AI.4.educators

AI.4.educators
Educating Educators on Artificial Intelligence (AI) – development of an AI training material and an AI educational program for educators

Project No: 2021-1-EL01-KA210-ADU-000034976

AI Ethical Roadmap

Co-funded by
the European Union

The European Commission's support for the production of this publication does not constitute an endorsement of the contents, which reflect the views only of the authors, and the Commission cannot be held responsible for any use which may be made of the information contained therein.

Table of Contents

1. Ethical values in AI systems
2. European approach on Ethics and AI
3. Ethical considerations on AI of today

Ethical values in AI systems

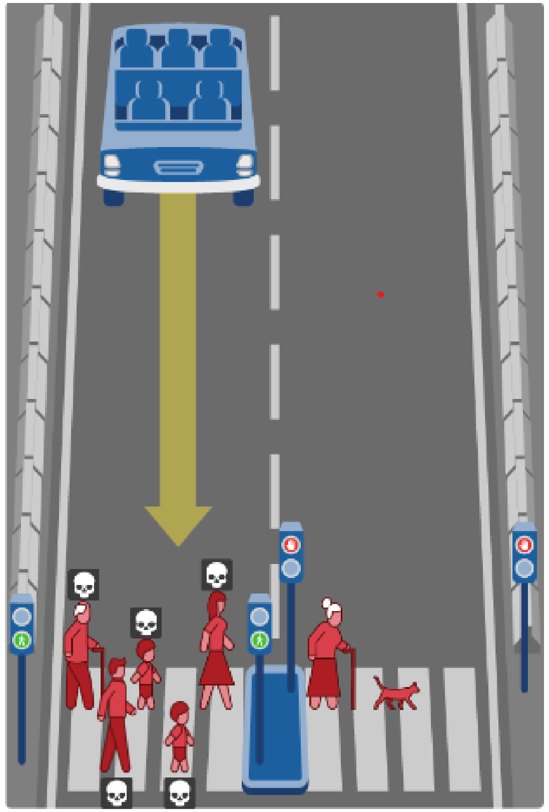
Ethics is a set of moral principles which help us discern between right and wrong. AI ethics is a set of guidelines that advise on the design and outcomes of artificial intelligence. Human beings come with all sorts of cognitive biases, such as recency and confirmation bias, and those inherent biases are exhibited in our behaviors and subsequently, our data. Since data is the foundation for all machine learning algorithms, it’s important for us to structure experiments and algorithms with this in mind as artificial intelligence has the potential to amplify and scale these human biases at an unprecedented rate.

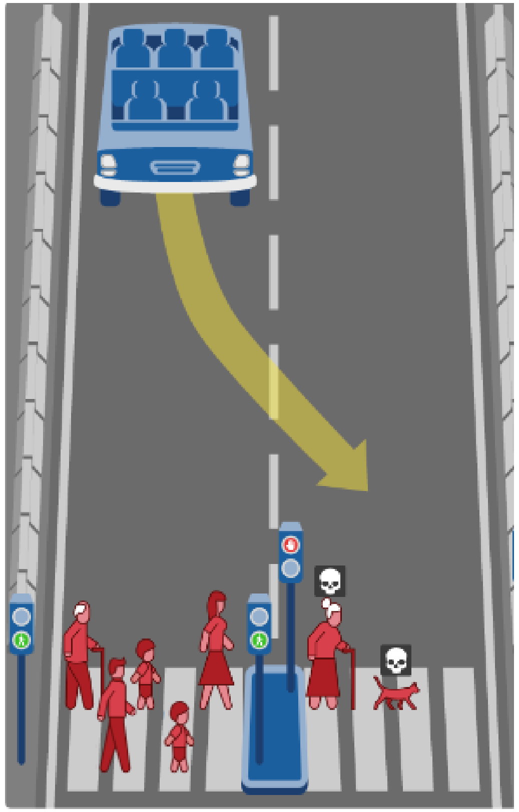
Ethical values in AI systems

The question of whether and how technologies embody values is not new. It has been discussed in the philosophy of technology, where several accounts have been developed. Some authors deny that technologies are, or can be, value-laden, while others see technologies as imbued with values due to the way they have been designed. Still others treat technologies as moral agents, somewhat similar to human agents, and some even argue for abandoning the distinction between (human) subjects and (technological) objects altogether in understanding how technologies may embody values.

Ethical values in AI systems – The ethical dilemmas

In this case the self driving car will continue ahead. This will result
Dead:
1 elderly man
2 boys
1 woman
1 man



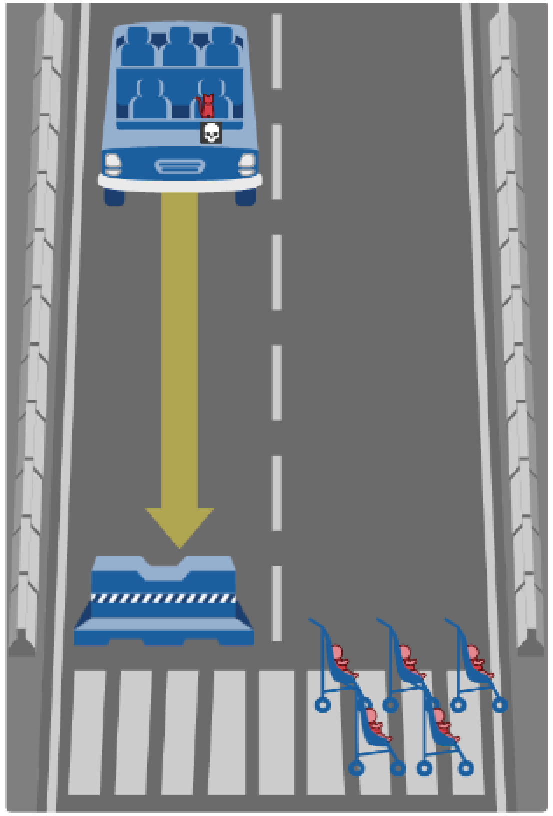


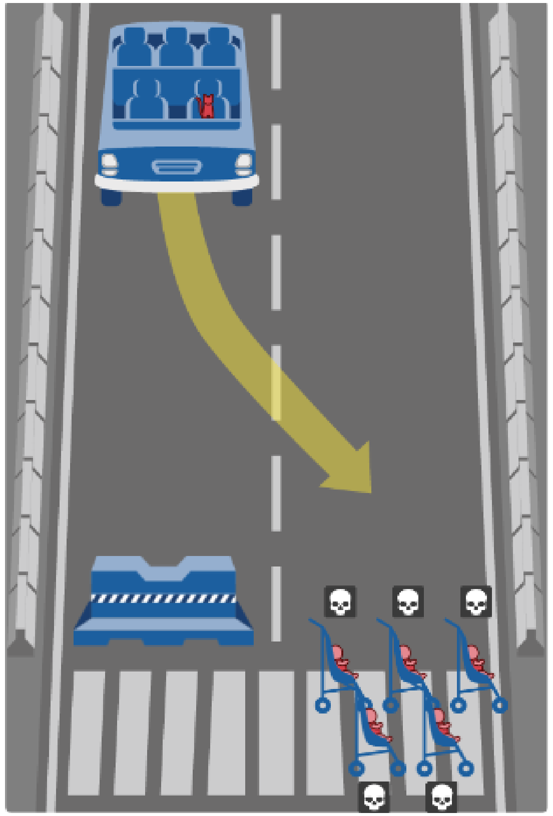
In this case the self driving car will swerve and drive through a pedestrian crossing. This will result
Dead:
1 elderly woman
1 cat

See more at https://www.moralmachine.net/

Ethical values in AI systems – The ethical dilemmas

In this case the self driving car will continue ahead and crash into a barrier. This will result
Dead:
1 cat



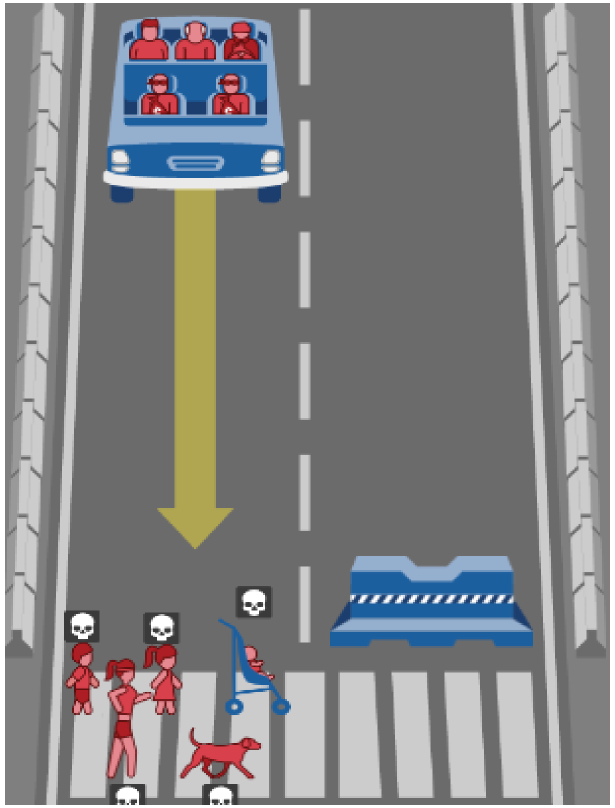


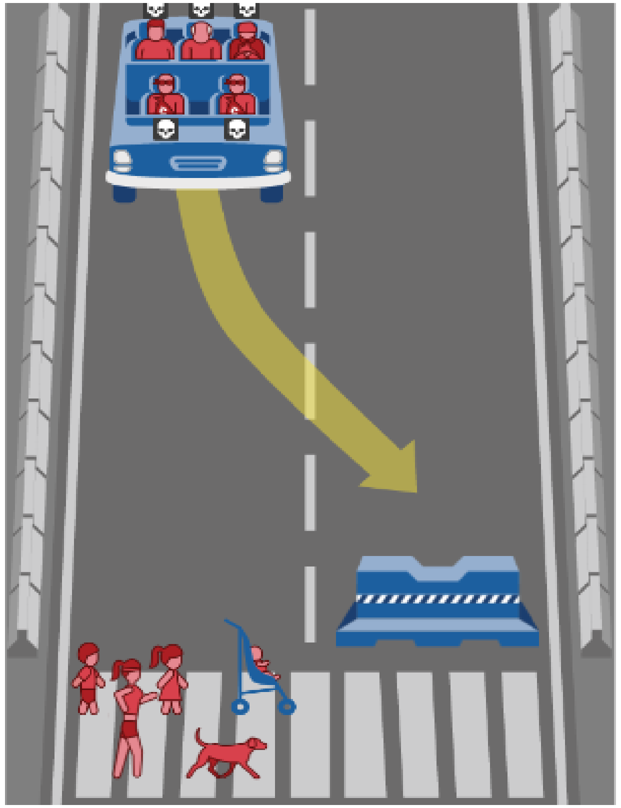
In this case the self driving car will swerve and drive through a pedestrian crossing. This will result
Dead:
5 babies

See more at https://www.moralmachine.net/

Ethical values in AI systems – The ethical dilemmas

In this case the self driving car will continue ahead. This will result
Dead:
2 children
1 baby
1 athlete
1 dog



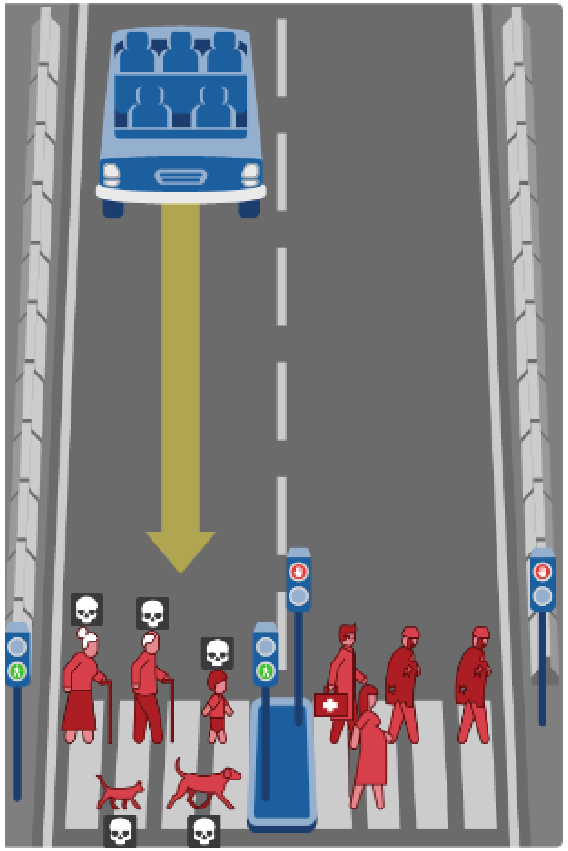


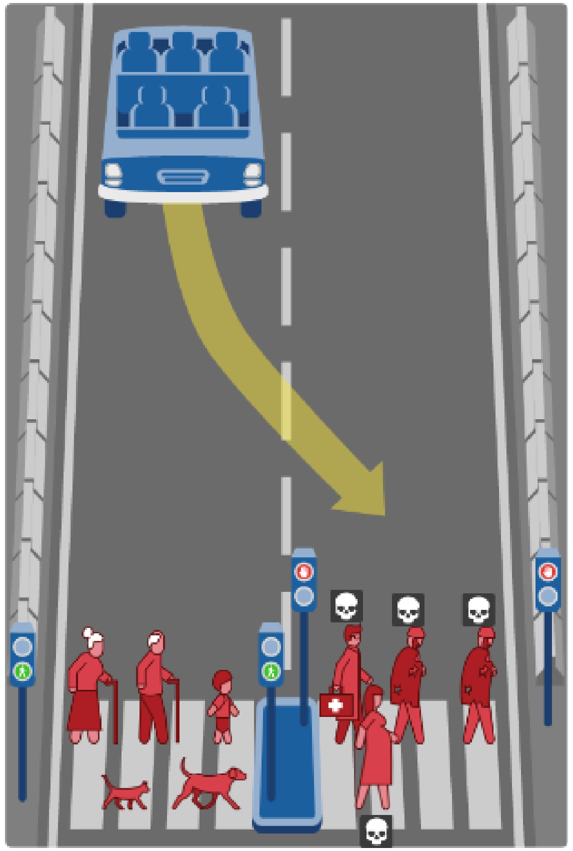
In this case the self driving car will swerve and crash into a barrier. This will result
Dead:
2 criminals
1 homeless person
1 elderly man
1 large man

See more at https://www.moralmachine.net/

Ethical values in AI systems – The ethical dilemmas

In this case the self driving car will continue ahead. This will result
Dead:
2 elderly people
1 boy
1 cat
1 dog
[Green Light]

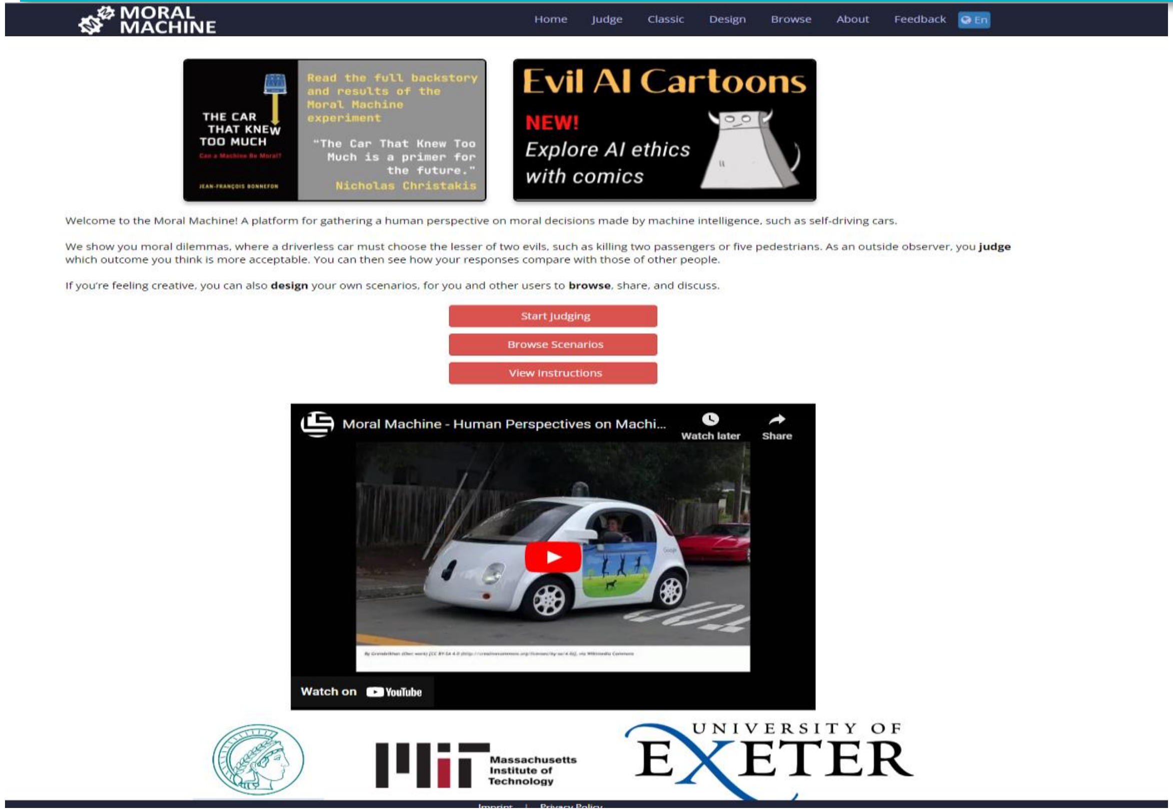




In this case the self driving car will swerve. This will result
Dead:
1 doctor
2 homeless people
1 pregnant woman
[Red light]

See more at https://www.moralmachine.net/

Ethical values in AI systems – The ethical dilemmas


MORAL MACHINE
Home
Judge
Classic
Design
Browse
About
Feedback
En
THE CAR THAT KNEW TOO MUCH
Can a Machine Be Moral?
JEAN-FRANÇOIS BONNEFON
Read the full backstory and results of the Moral Machine experiment
"The Car That Knew Too Much is a primer for the future."
Nicholas Christakis
Evil AI Cartoons
NEW!
Explore AI ethics with comics
Start Judging
Browse Scenarios
View Instructions
Moral Machine - Human Perspectives on Machi...
Watch later
Share
Watch on YouTube
Massachusetts Institute of Technology
UNIVERSITY OF EXETER

Welcome to the Moral Machine! A platform for gathering a human perspective on moral decisions made by machine intelligence, such as self-driving cars.

We show you moral dilemmas, where a driverless car must choose the lesser of two evils, such as killing two passengers or five pedestrians. As an outside observer, you **judge** which outcome you think is more acceptable. You can then see how your responses compare with those of other people.

If you're feeling creative, you can also **design** your own scenarios, for you and other users to **browse**, share, and discuss.

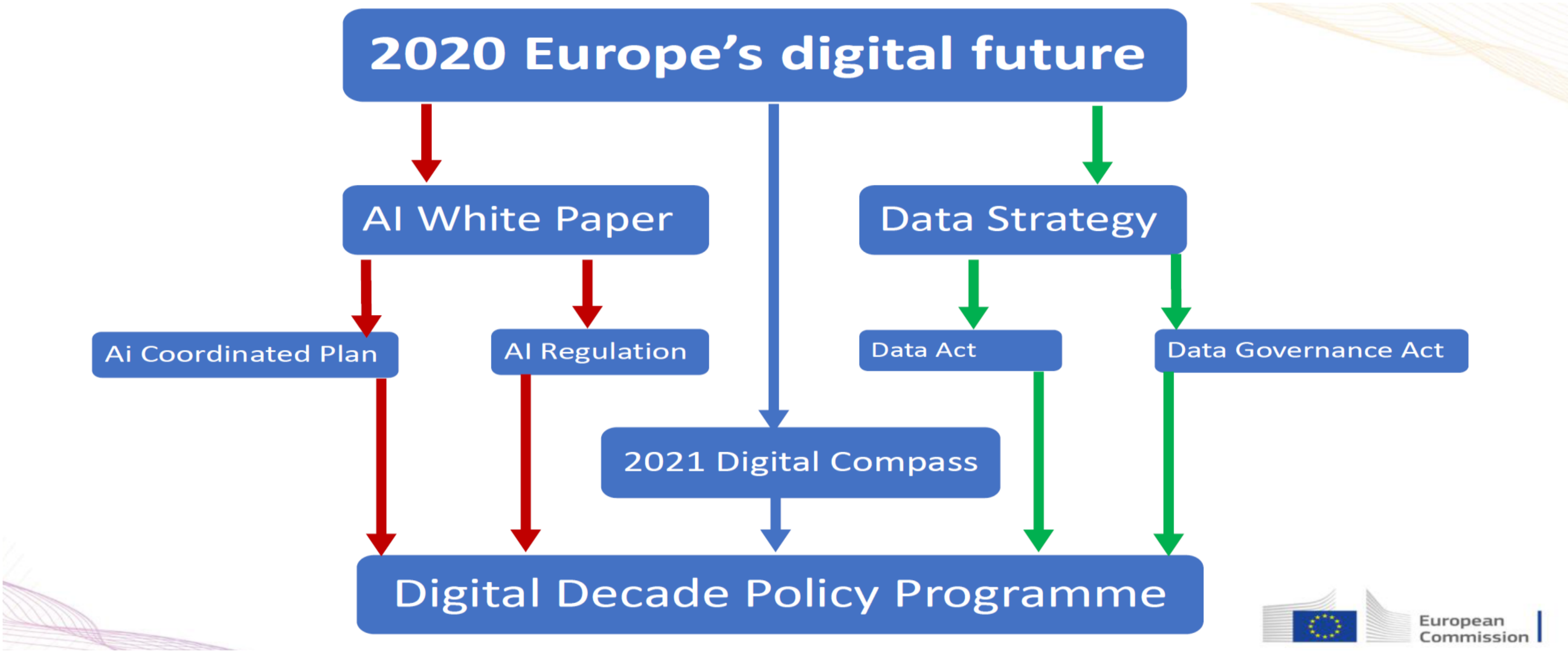
Visit MIT's MORAL MACHINE website https://www.moralmachine.net/ play the game and compare your answers with game's results.

If we cannot agree witch is the "right" answer , how could the AI System decide?

Welcome to the world of AI Ethics.

European approach on Ethics and AI

EU’s strategy in the field of AI at a glance



EU’s Official Website https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence#ecl-inpage-l6ov8brl

The European AI strategy

The European AI Strategy aims at making the EU a world-class hub for AI and ensuring that AI is human-centric and trustworthy. Such an objective translates into the European approach to excellence and trust (.pdf) through concrete rules and actions.

In April 2021, the Commission presented its AI package, including:

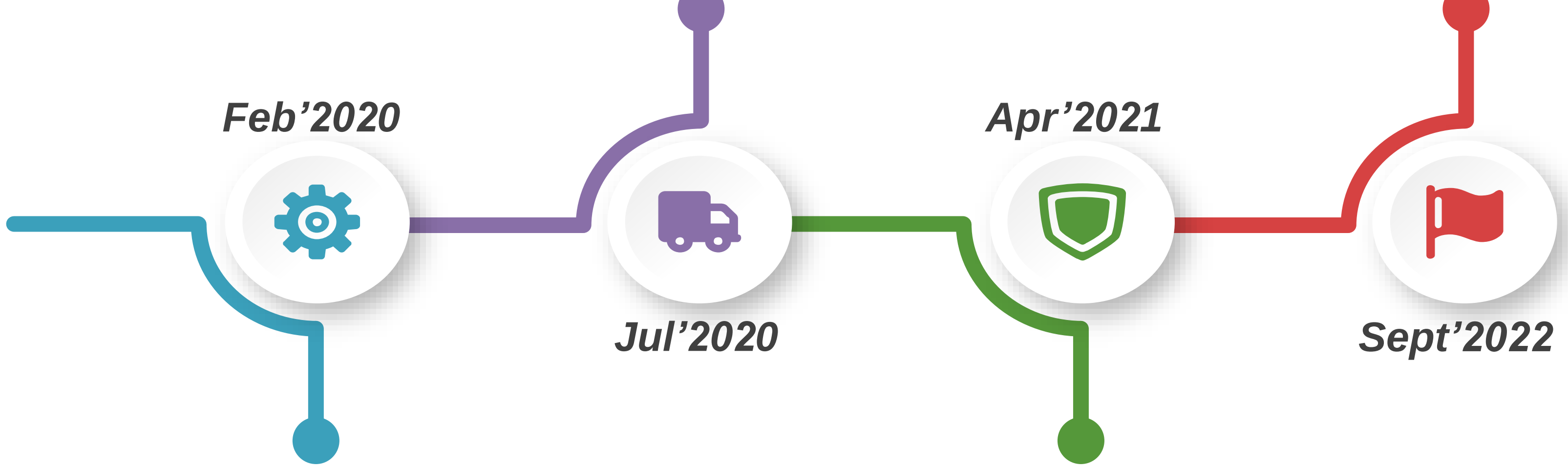
- its Communication on fostering a European approach to artificial intelligence ;
- an update of the Coordinated Plan on Artificial Intelligence (with EU Member States);
- its proposal for a regulation laying down harmonised rules on AI (AI Act) and relevant Impact assessment.

A European approach to artificial intelligence

“The EU’s approach to artificial intelligence centers on excellence and trust, aiming to boost research and industrial capacity while ensuring safety and fundamental rights.
The way we approach Artificial Intelligence (AI) will define the world we live in the future. To help building a resilient, people and businesses should be able to enjoy the benefits of AI while feeling safe and protected.
The European AI Strategy aims at making the EU a world-class hub for AI and ensuring that AI is human-centric and trustworthy. Such an objective translates into the European approach to excellence and trust (.pdf) through concrete rules and actions.”

EU’s Official Website https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence#ecl-inpage-l6ov8brl

EU regulates AI: The Milestones



Artificial intelligence — ethical and legal requirements

Proposal for an AI liability directive

Feb'2020

Apr'2021

Jul'2020

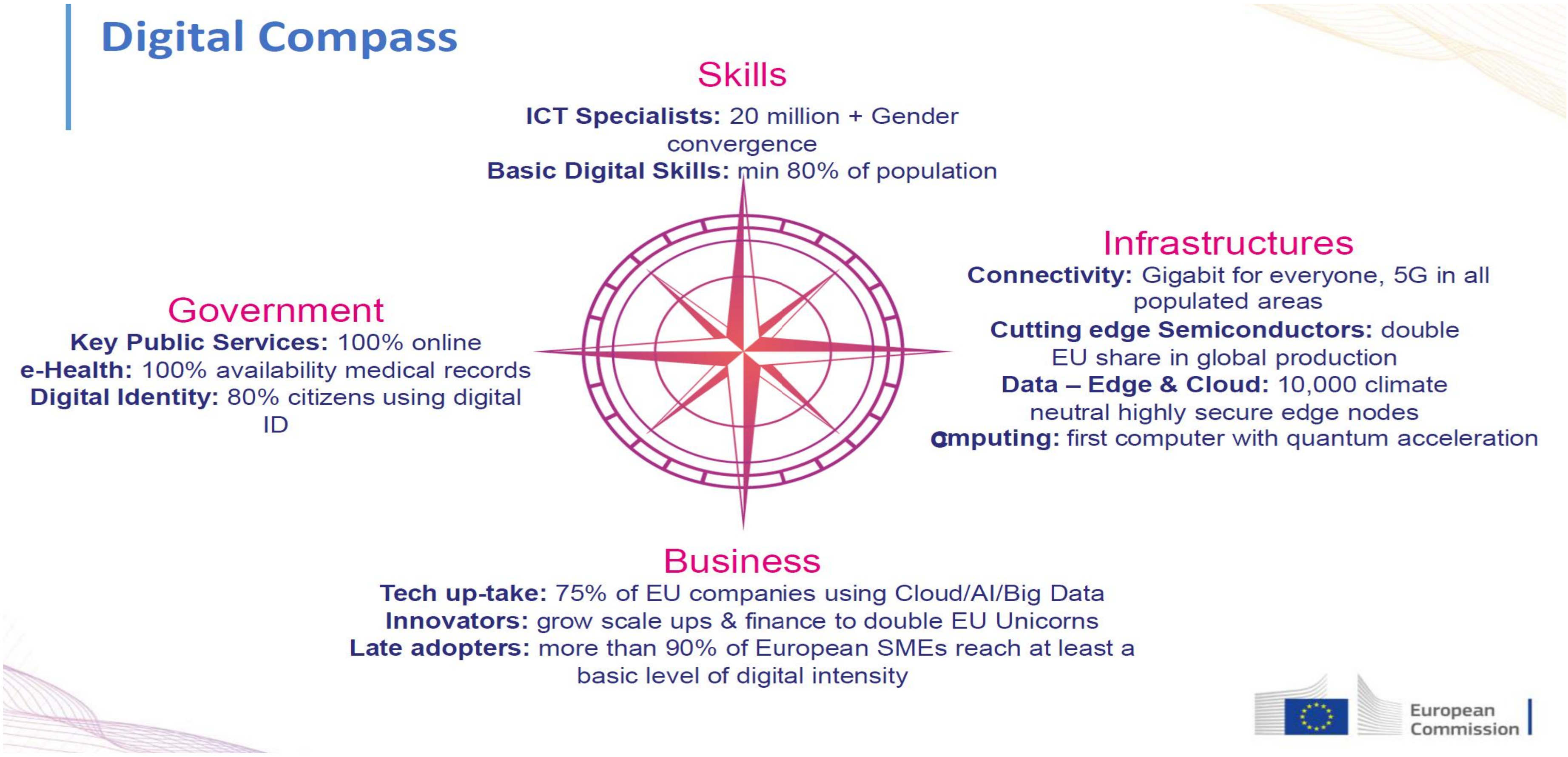
Sept'2022

White paper on AI: a European approach to excellence and trust

EC: Proposal for a regulation laying down harmonised rules on AI (AI Act)

More info about the EU's strategy milestones at: https://digital-strategy.ec.europa.eu/en/policies/european-approach-artificial-intelligence#ecl-inpage-l6ov8brl

EU's strategy in the field of AI at a glance



Digital Compass

Skills

ICT Specialists: 20 million + Gender convergence
Basic Digital Skills: min 80% of population

Infrastructures

Connectivity: Gigabit for everyone, 5G in all populated areas
Cutting edge Semiconductors: double EU share in global production
Data – Edge & Cloud: 10,000 climate neutral highly secure edge nodes
Computing: first computer with quantum acceleration

Government

Key Public Services: 100% online
e-Health: 100% availability medical records
Digital Identity: 80% citizens using digital ID

Business

Tech up-take: 75% of EU companies using Cloud/AI/Big Data
Innovators: grow scale ups & finance to double EU Unicorns
Late adopters: more than 90% of European SMEs reach at least a basic level of digital intensity

European Commission

EU’s Coordinated Plan on Artificial Intelligence 1/7

The key aims of the Coordinated Plan on Artificial Intelligence 2021 Review are to accelerate investment in AI, act on AI strategies and programmes and align AI policy to avoid fragmentation.

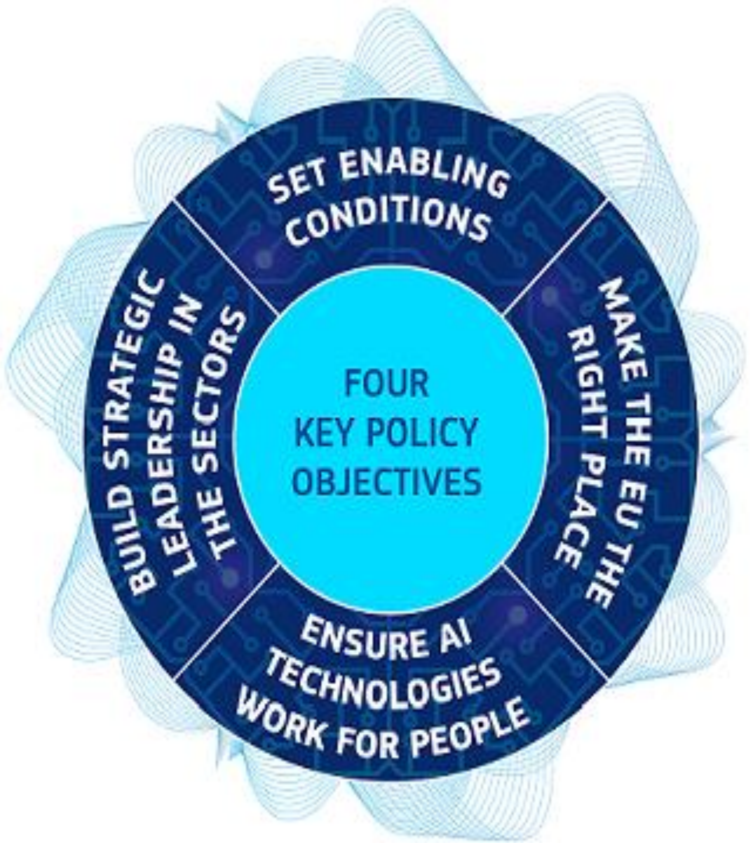
Turning strategy into action, the 2021 Coordinated Plan’s key message is that the Commission and Member States should:

- **accelerate** investments in AI technologies to drive resilient economic and social recovery aided by the uptake of new digital solutions.
- **act** on AI strategies and programmes by fully and timely implementing them to ensure that the EU fully benefits from first-mover adopter advantages.
- **align** AI policy to remove fragmentation and address global challenges.

EU’s Coordinated Plan on Artificial Intelligence 2/7

In order to achieve this, the updated plan sets four key sets of policy objectives, supported by concrete actions and indicating possible funding mechanism and the timeline to:

1. Set enabling conditions for AI development and uptake in the EU
2. Make the EU the place where excellence thrives from the lab to market
3. Ensure that AI works for people and is a force for good in society
4. Build strategic leadership in high-impact sectors


SET ENABLING CONDITIONS
MAKE THE EU THE RIGHT PLACE
ENSURE AI TECHNOLOGIES WORK FOR PEOPLE
BUILD STRATEGIC LEADERSHIP IN THE SECTORS
FOUR KEY POLICY OBJECTIVES

EU’s Coordinated Plan on Artificial Intelligence 3/7 - Set enabling conditions for AI development and uptake in the EU

Create broad enabling conditions for AI technologies to succeed in the EU through acquiring, pooling and sharing policy insights, tapping into the potential of data and fostering critical computing infrastructure.

EU’s Coordinated Plan on Artificial Intelligence 4/7 - Make EU the place where excellence thrives from the lab to the market

Ensure that EU has a strong ecosystem of excellence including world-class foundational and application-oriented research and capabilities to bring innovations from the ‘lab to the market’. Testing and experimentation facilities (TEFs), European Digital Innovation Hubs (EDIHs) and the European ‘AI-on-demand’ platform will play key role in facilitating a broad uptake and deployment of AI technologies.

EU’s Coordinated Plan on Artificial Intelligence 5/7 - Ensure that AI works for people and is a force for good in society

The EU has to ensure that AI developed and put on the market in the EU is human-centric, sustainable, secure, inclusive and trustworthy. The proposed actions focus on:

- nurturing talent and improving the supply of skills necessary to enable a thriving AI eco-system.
- developing the policy framework to ensure trust in AI systems.
- promoting the EU vision on sustainable and trustworthy AI in the world.

EU’s Coordinated Plan on Artificial Intelligence 6/7 - Build strategic leadership in high-impact sectors

To align with the market developments and ongoing actions in Member States and to reinforce EU position on the global scale the review puts forward joint actions in seven sectors. These sectors are environment, health, a strategy for robotics in the world of AI, public sector, transport, law enforcement, migration and asylum, and agriculture.

EU’s Coordinated Plan on Artificial Intelligence 7/7 - Investment

The Commission proposed that the EU invests in AI at least €1 billion per year from the Horizon Europe and Digital Europe programmes. EU-level funding on AI should attract and pool investment to foster collaboration among Member States, and maximise impact by joining forces.

The Recovery and Resilience Facility provides an unprecedented opportunity to modernise and invest in AI. Through this the EU can become a global leader in the development and uptake of human-centric, trustworthy, secure and sustainable AI technologies.

Ethics guidelines for trustworthy AI

On 8 April 2019, the High-Level Expert Group on AI presented Ethics Guidelines for Trustworthy Artificial Intelligence. This followed the publication of the guidelines' first draft in December 2018 on which more than 500 comments were received through an open consultation.

According to the Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - both from a technical perspective while taking into account its social environment

Ethics guidelines for trustworthy AI – Key Requirements

The Guidelines put forward a set of 7 key requirements that AI systems should meet in order to be deemed trustworthy. A specific assessment list aims to help verify the application of each of the key requirements:

Key Requirement 1: Human agency and oversight

Human agency and oversight: AI systems should empower human beings, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms need to be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches.

Key Requirement 2: Technical Robustness and safety

Technical Robustness and safety: AI systems need to be resilient and secure. They need to be safe, ensuring a fall back plan in case something goes wrong, as well as being accurate, reliable and reproducible. That is the only way to ensure that also unintentional harm can be minimized and prevented.

Key Requirement 3: Privacy and data governance

Privacy and data governance: besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be ensured, taking into account the quality and integrity of the data, and ensuring legitimised access to data.

Key Requirement 4: Transparency

Transparency: the data, system and AI business models should be transparent. Traceability mechanisms can help achieving this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned. Humans need to be aware that they are interacting with an AI system, and must be informed of the system’s capabilities and limitations.

Key Requirement 5: Diversity, non-discrimination and fairness

Diversity, non-discrimination and fairness: Unfair bias must be avoided, as it could could have multiple negative implications, from the marginalization of vulnerable groups, to the exacerbation of prejudice and discrimination. Fostering diversity, AI systems should be accessible to all, regardless of any disability, and involve relevant stakeholders throughout their entire life circle.

Key Requirement 6: Societal and environmental well-being

Societal and environmental well-being: AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should take into account the environment, including other living beings, and their social and societal impact should be carefully.

Key Requirement 7: Accountability

Accountability: Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data and design processes plays a key role therein, especially in critical applications. Moreover, adequate an accessible redress should be ensured.

Finalization of Ethics guidelines for trustworthy AI

The piloting phase closed on 1 December 2019

Based on the feedback received, the AI HLEG presented the final Assessment List for Trustworthy AI (ALTAI) in July 2020. ALTAI is practical tool that translates the Ethics Guidelines into an accessible and dynamic (self-assessment) checklist. The checklist can be used by developers and deployers of AI who want to implement the key requirements in practice. This new list is available as a prototype web based tool and in PDF format.

Ethical considerations on AI of today

Which is the view of OECD, UNESCO and the Council of Europe?

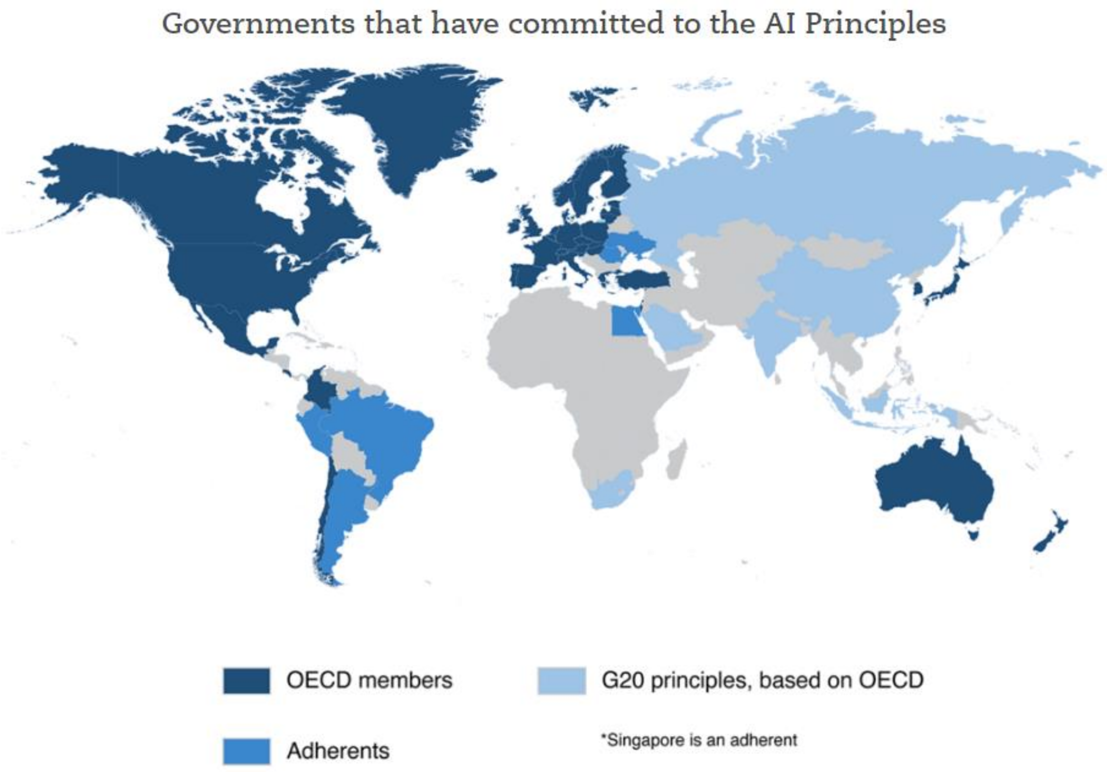
OECD
BETTER POLICIES FOR BETTER LIVES

COUNCIL OF EUROPE
CONSEIL DE L'EUROPE

unesco
Building peace in the minds of men and women

OECD: AI Principles

- Inclusive growth, sustainable development and well – being
- Human - centred values and fairness
- Transparency and explainability
- Robustness, security and safety
- Accountability


Governments that have committed to the AI Principles
OECD members
G20 principles, based on OECD
Adherents
*Singapore is an adherent

See more at: https://oecd.ai/en/ai-principles

UNESCO: Ethics of Artificial Intelligence

“AI can provide millions of students with support to complete secondary education, fill an additional 3.3 million jobs, and, more urgently, help us tackle the spread and the aftermath of the COVID-19 pandemic. Along with multiple advantages, these technologies also generate downside risks and challenges, derived from malicious use of technology or deepening inequalities and divides.”

See more at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

UNESCO: Global agreement on the Ethics of AI 1/3

In November 2021, the 193 Member States at UNESCO’s General Conference adopted the Recommendation on the Ethics of Artificial Intelligence, the very first global standard-setting instrument on the subject. It will not only protect but also promote human rights and human dignity, and will be an ethical guiding compass and a global normative bedrock allowing to build strong respect for the rule of law in the digital world.

See more at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

UNESCO: Global agreement on the Ethics of AI - Values 2/3

- Respect, protection and promotion of human rights and fundamental freedoms and human Dignity
- Environment and ecosystem flourishing
- Ensuring diversity and inclusiveness
- Living in peaceful, just and interconnected societies

See more at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

UNESCO: Global agreement on the Ethics of AI - Principles 3/3

- Proportionality and Do No Harm
- Fairness and non-discrimination
- Safety and security
- Sustainability
- Right to Privacy, and Data Protection
- Transparency and explainability
- Responsibility and accountability
- Awareness and literacy
- Multi-stakeholder and adaptive governance and collaboration

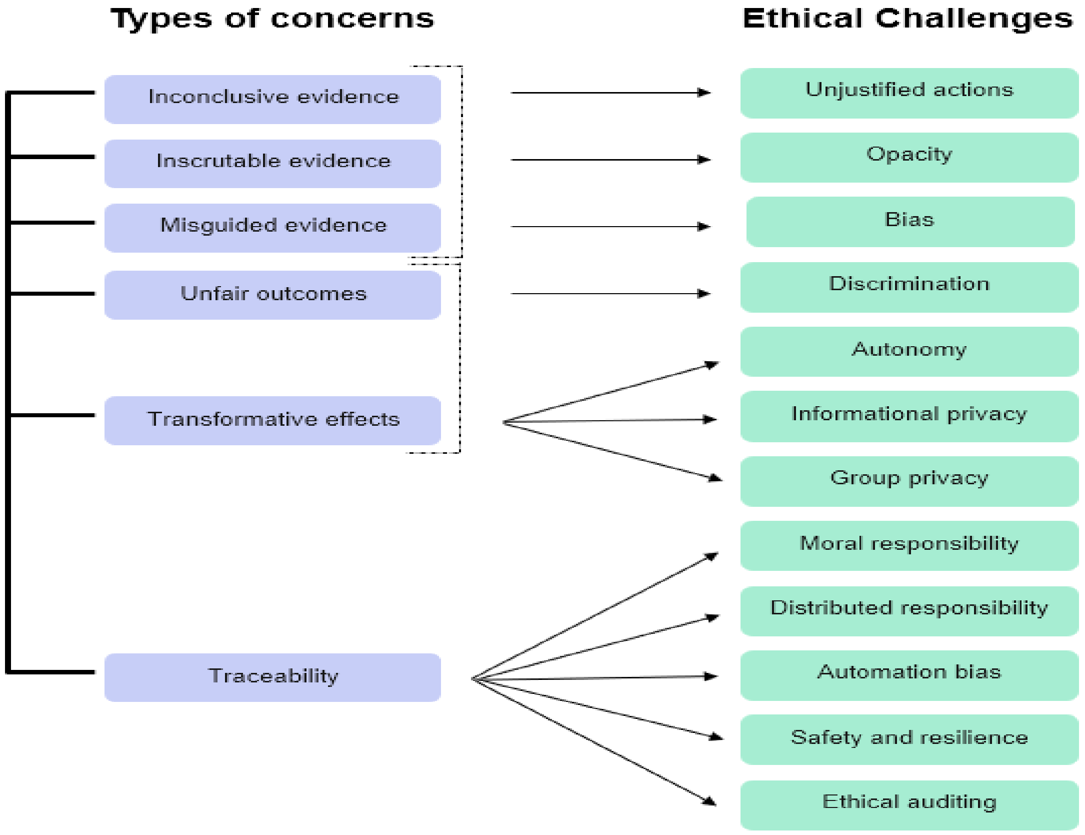
See more at: https://www.unesco.org/en/artificial-intelligence/recommendation-ethics

Council of Europe: Common ethical challenges in AI 1/17

Taking into account that:
“decision-making algorithms (1) turn data into evidence for a given outcome (henceforth conclusion), and that this outcome is then used to (2) trigger and motivate an action that (on its own, or when combined with other actions) may not be ethically neutral. This work is performed in ways that are complex and (semi-)-autonomous, which (3) complicates apportionment of responsibility for effects of actions driven by algorithms.”
3 epistemological and 2 normative types of ethical concerns can be identified based on how algorithms process data to produce evidence and motivate actions (See next slide).
The proposed 5 types of concerns can cause failures involving multiple human, organisational, and technological agents. This mix of human and technological actors leads to difficult questions concerning how to assign responsibility and liability for the impact of AI behaviours. These difficulties are captured in traceability as a final, overarching, type of concern.

See more at: [https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}](https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]})

Council of Europe: Common ethical challenges in AI 2/17



Types of concerns

- Inconclusive evidence
- Inscrutable evidence
- Misguided evidence
- Unfair outcomes
- Transformative effects
- Traceability

Ethical Challenges

- Unjustified actions
- Opacity
- Bias
- Discrimination
- Autonomy
- Informational privacy
- Group privacy
- Moral responsibility
- Distributed responsibility
- Automation bias
- Safety and resilience
- Ethical auditing

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Inconclusive evidence 3/17

When algorithms draw conclusions from the data they process using inferential statistics and/or machine learning techniques, they produce probable yet inevitably uncertain knowledge. Statistical learning theory and computational learning theory are both concerned with the characterisation and quantification of this uncertainty. Statistical methods can identify significant correlations, but correlations are typically not sufficient to demonstrate causality, and thus may be insufficient to motivate action on the basis of knowledge of such a connection. The concept of an ‘actionable insight’ captures the uncertainty inherent in statistical correlations and normativity of choosing to act upon them.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Inscrutable evidence 4/17

When data are used as (or processed to produce) evidence for a conclusion, it is reasonable to expect that the connection between the data and the conclusion should be intelligible and open to scrutiny. Given the complexity and scale of many AI systems, intelligibility and scrutiny cannot be taken for granted. A lack of access to datasets and the inherent difficulty of mapping how the multitude of data and features considered by an AI system contribute to specific conclusions and outputs cause practical as well as principled limitations.

See more at: [https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}](https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]})

Council of Europe: Common ethical challenges in AI – Misguided evidence 5/17

Algorithms process data and are therefore subject to a limitation shared by all types of data processing, namely that the output can never exceed the input. The informal ‘garbage in, garbage out’ principle illustrates this phenomenon and its significance: conclusions can only be as reliable (but also as neutral) as the data they are based on.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Unfair Outcomes 6/17

Algorithmically driven actions can be scrutinised from a variety of ethical perspectives, criteria, and principles. The normative acceptability of the action and its effects is observer-dependent and can be assessed independently of its epistemological quality. An action can be found discriminatory, for example, solely from its effect on a protected class of people, even if made on the basis of conclusive, scrutable and well-founded evidence.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Transformative effects 7/17

The impact of AI systems cannot always be attributed to epistemic or ethical failures. Much of their impact can appear initially ethically neutral in the absence of obvious harm. A separate set of impacts, which can be referred to as transformative effects, concern subtle shifts in how the world is conceptualised and organised.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Unjustified Actions 8/17

Much algorithmic decision-making and data mining relies on inductive knowledge and correlations identified within a dataset. Correlations based on a ‘sufficient’ volume of data are often seen as sufficiently credible to direct action without first establishing causality. Acting on correlations can be doubly problematic. Spurious correlations may be discovered rather than genuine causal knowledge. Even if strong correlations or causal knowledge are found, this knowledge may only concern populations while actions with significant personal impact are directed towards individuals.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Opacity 9/17

This is the ‘black box’ problem with AI: the logic behind turning inputs into outputs may not be known to observers or affected parties or may be fundamentally inscrutable or unintelligible. Opacity in machine learning algorithms is a product of the high dimensionality of data, complex code and changeable decision-making logic.[1] Transparency and comprehensibility are generally desired because algorithms that are poorly predictable or interpretable are difficult to control, monitor and correct.[2] Transparency is often naively treated as a panacea for ethical issues arising from new technologies.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Bias 10/17

The automation of human decision-making is often justified by an alleged lack of bias in AI and algorithms. This belief is unsustainable; AI systems unavoidably make biased decisions. A system's design and functionality reflects the values of its designer and intended uses, if only to the extent that a particular design is preferred as the best or most efficient option. Development is not a neutral, linear path. As a result, "the values of the author, wittingly or not, are frozen into the code, effectively institutionalising those values." Inclusiveness and equity in both the design and usage of AI is thus key to combat implicit biases. Friedman and Nissenbaum clarify that bias arise from (1) pre-existing social values found in the "social institutions, practices and attitudes" from which the technology emerges, (2) technical constraints and (3) emergent aspects of a context of use.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Discrimination 11/17

Discrimination against individuals and groups can arise from biases in AI systems. Discriminatory analytics can contribute to self-fulfilling prophecies and stigmatisation in targeted groups, undermining their autonomy and participation in society. While a single definition of discrimination does not exist, legal frameworks internationally have a long history of jurisprudence discussing types of discrimination (e.g., direct and indirect), goals of equality law (e.g., formal and substantive equality), and appropriate thresholds for distribution of outcomes across groups. In this context, embedding considerations of non-discrimination and fairness into AI systems is particularly difficult. It may be possible to direct algorithms not to consider sensitive attributes that contribute to discrimination, such as gender or ethnicity, based upon the emergence of discrimination in a particular context. However, proxies for protected attributes are not easy to predict or detect, particularly when algorithms access linked datasets.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Autonomy 12/17

Value-laden decisions made by algorithms can also pose a threat to autonomy. Personalisation of content by AI systems, such as recommender systems, is particularly challenging in this regard. Personalisation can be understood as the construction of choice architectures which are not the same across a sample. AI can nudge the behaviour of data subjects and human decision-makers by filtering information. Different information, prices, and other content can be offered to profiling groups or audiences within a population defined by one or more attributes, for example the ability to pay, which can itself lead to discrimination. Personalisation reduces the diversity of information users encounter by excluding content deemed irrelevant or contradictory to the user's beliefs or desires. This is problematic insofar as information diversity can be considered an enabling condition for autonomy. The subject’s autonomy in decision-making is disrespected when the desired choice reflects third-party interests above the individual’s.

See more at: [https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}](https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]})

Council of Europe: Common ethical challenges in AI – Informational privacy and group privacy 13/17

Algorithms also transform notions of privacy. Responses to discrimination, personalisation, and the inhibition of autonomy due to opacity often appeal to informational privacy, or the right of data subjects to “shield personal data from third parties.” Informational privacy concerns the capacity of an individual to control information about herself, and the effort required by third parties to obtain this information. A right to identity derived from informational privacy suggests that opaque or secretive profiling is problematic when carried out by a third party. In a healthcare setting this could include insurers, remote care providers (e.g., chatbot and triage service providers), consumer technology companies, and others. Opaque decision-making inhibits oversight and informed decision-making concerning data sharing. Data subjects cannot define privacy norms to govern all types of data generically because the value or insightfulness of data is only established through processing.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Moral responsibility and distributed responsibility 14/17

When a technology fails, blame and sanctions must be apportioned. Blame can only be justifiably attributed when the actor has some degree of control and intentionality in carrying out the action. Traditionally, developers and software engineers have had “control of the behaviour of the machine in every detail” insofar as they can explain its overall design and function to a third party. This traditional conception of responsibility in software design assumes the developer can reflect on the technology’s likely effects and potential for malfunctioning, and make design choices to choose the most desirable outcomes according to the functional specification.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Automation bias 15/17

A related problem concerns the diffusion of feelings of responsibility and accountability for users of AI systems, and the related tendency to trust the outputs of systems on the basis of their perceived objectivity, accuracy, or complexity. Delegating decision-making to AI can shift responsibility away from human decision-makers. Similar effects can be observed in mixed networks of human and information systems as already studied in bureaucracies, characterised by reduced feelings of personal responsibility and the execution of otherwise unjustifiable actions. Algorithms involving stakeholders from multiple disciplines can, for instance, lead to each party assuming others will shoulder ethical responsibility for the algorithm’s actions. Machine learning adds an additional layer of complexity between designers and actions driven by the algorithm, which may justifiably weaken blame placed upon the former.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Safety and resilience 16/17

The need to apportion responsibility is acutely felt when algorithms malfunction. Unethical algorithms can be thought of as malfunctioning software artefacts that do not operate as intended. Useful distinctions exist between errors of design (types) and errors of operation (tokens), and between the failure to operate as intended (dysfunction) and the presence of unintended side-effects (misfunction). Misfunctioning is distinguished from mere negative side effects by 'avoidability', or the extent to which comparable types of systems or artefacts accomplish the intended function without the effects in question. These distinctions clarify ethical aspects of AI systems that are strictly related to their functioning, either in the abstract (for instance when we look at raw performance), or as part of a larger decision-making system, and reveals the multifaceted interaction between intended and actual behaviour. Machine learning in particular raises unique challenges, because achieving the intended or "correct" behaviour does not imply the absence of errors or harmful actions and feedback loops.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}

Council of Europe: Common ethical challenges in AI – Ethical auditing 17/17

How best to operationalise and set standards for testing of these ethical challenges remains an open question, particularly for machine learning. Merely rendering the code of an algorithm transparent is insufficient to ensure ethical behaviour. One possible path to achieve interpretability, fairness, and other ethical goals in AI systems is via auditing carried out by data processors, external regulators, or empirical researchers, using ex post audit studies, reflexive ethnographic studies in development and testing, or reporting mechanisms designed into the algorithm itself. For all types of AI, auditing is a necessary precondition to verify correct functioning. For systems with foreseeable human impact, auditing can create an ex post procedural record of complex automated decision-making to unpack problematic or inaccurate decisions, or to detect discrimination or similar harms.

See more at: https://www.coe.int/en/web/bioethics/common-ethical-challenges-in-ai#{%22123745744%22:[]}