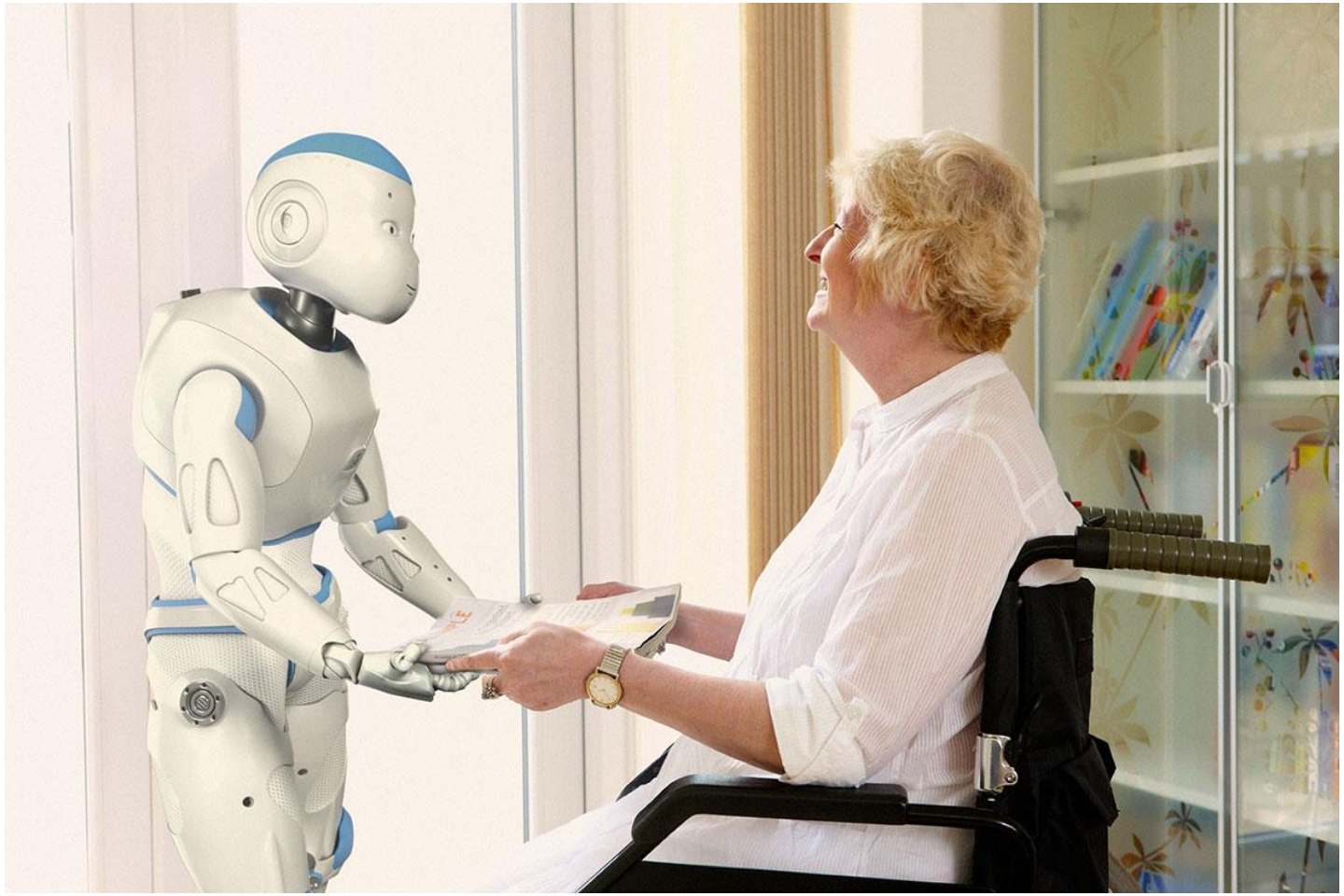
Operational principles and tensions in AI ethics

Alexei Grinbaum
Senior research scientist, CEA-Saclay (France)
Chair of CEA Operational Digital Ethics Committee
Member of French National Digital Ethics Committee (CCNEN)

AIOLIA project is funded by the European Commission under Grant Agreement 101187937



EU AI ethics guidelines

Foundations of Trustworthy AI

Adhere to ethical principles based on fundamental rights

4 Ethical Principles

Acknowledge and address tensions between them

- Respect for human autonomy
- Prevention of harm
- Fairness
- Explicability

Realisation of Trustworthy AI

Implement the key requirements

7 Key Requirements

Evaluate and address these continuously throughout the AI system's life cycle via

- **Technical Methods**
- **Non-Technical Methods**

- Human agency and oversight
- Technical robustness and safety
- Privacy and data governance
- Transparency
- Diversity, non-discrimination and fairness
- Societal and environmental wellbeing
- Accountability

EU AI Act recital 27

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Ethics of AI in Horizon Europe

2021

Key prerequisites for ethically sound AI systems:

- ✓ Human agency and oversight;
- ✓ Privacy and data governance;
- ✓ Fairness, diversity and non-discrimination;
- ✓ Accountability;
- ✓ Transparency;
- ✓ Societal and environmental well-being.

This means that HE proposals:

- Must ensure that people are aware they are interacting with an AI system and are informed about its abilities, limitations, risks and benefits;
- Prevent possible limitations on human rights and freedoms;
- Are not designed in a way that may lead to objectification, dehumanization, subordination, discrimination, stereotyping, coercion, manipulation of people or creation of attachment or addiction;
- Must comply with the principles of data minimisation and privacy by design and by default;
- Must be designed in a way to avoid bias in both input data and algorithm design;
- Must address the potential impact on the individual, society or the environment;
- Must not reduce the safety and wellbeing of the individuals;
- Should be developed in a way that enables human oversight, traceability and auditability.

European Commission

AIOLIA: Operationalizing AI Ethics for Learning and Practice (HORIZON-WIDERA-2024-ERA-01-12)

GUIDELINES

Built from real-world use cases with a bottom-up approach

LEARNING MATERIALS

Tailored for engineers, early-career researchers, RECs, and more

TRAININGS

Offered in multiple formats – online and in-person

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

The AIOLIA Consortium

Global Partners in Operationalizing AI Ethics

CEA	CERTH	THWS	ERCIM	Oxipit
McGill Univ.	Univ. of Osaka	RISE	CENTRIC	CEPS
Euractiv	Afliant	CASTED	KIT	Amsterdam UMC
EUREC	ADRA	NIT Institute	ETICAS.AI	STEPI

Funded by the European Union (Horizon Europe Grant 101187937)

www.aiolia.eu

CHANGE IN HUMAN EXPERTISE AND PROFESSIONAL BEHAVIOUR

UC1 Medical doctors using AI tools in diagnostics and treatment

AI Areas: Decision-support systems; Image recognition

- Non-maleficence
- Accountability & responsibility
- Transparency & explainability

UC2 Safety engineers using AI tools to speed up software release approvals

AI Areas: Decision-support systems; General-purpose AI

- Robustness, safety & reliability
- Oversight & autonomy
- Risk of over-reliance & deskilling

UC3 Recruiters using AI tools in hiring processes

AI Areas: Decision-support systems

- Non-bias, fairness & non-discrimination
- Transparency & explainability
- Over-reliance & deskilling

UC4 Security professionals using AI tools to detect hate speech

AI Areas: Decision-support systems; General-purpose AI

- Freedom of expression & non-censorship
- Non-bias, fairness & non-discrimination
- Accountability & responsibility

CHANGE IN HUMAN COGNITION AND PRIVATE BEHAVIOUR

UC5 AI systems as individual and family-level virtual assistants

AI Areas: Conversational general-purpose AI; Emotional AI

- Human well-being
- Privacy, consent & data protection

UC6 Deepfake therapy for processing trauma and grief

AI Areas: Multi-modal general-purpose AI (GPAI); Emotional AI

- Non-maleficence
- Autonomy & non-manipulation
- Risk of over-reliance

TRAINING TARGET GROUPS

- EU Ethics Appraisal Scheme experts
- National research ethics committees
- REC members
- Research ethics trainers
- Research ethics and integrity organizations and networks
- Researchers in cognitive science and neuroscience
- Researchers in AI
- Master and PhD students in computer science and AI
- AI industrial engineers
- AI ethics researchers, social scientists and philosophers

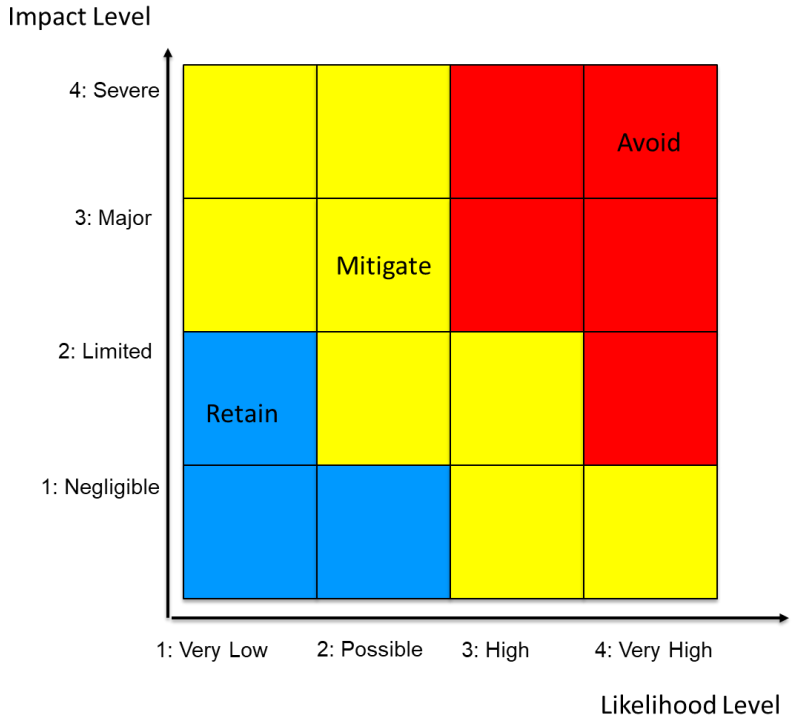
AIOLIA project is funded by the European Commission under Grant Agreement 101187937

1. Findings about ethics principles

Risk-based approach

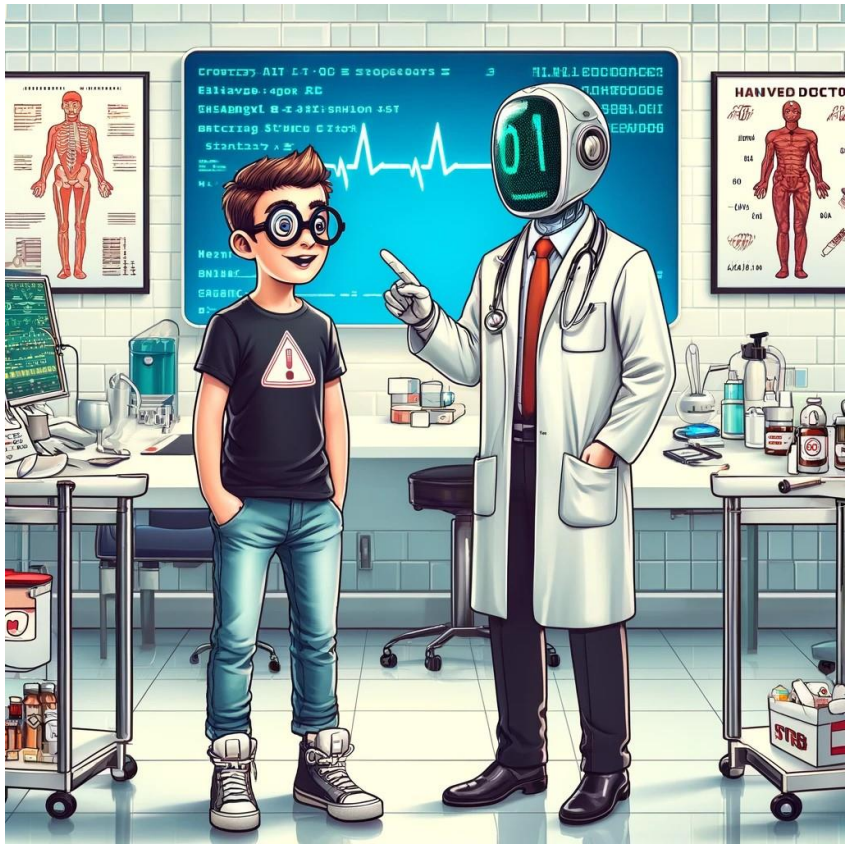
Industrial partners view the operationalisation of ethics as a strategy to **manage risks**.

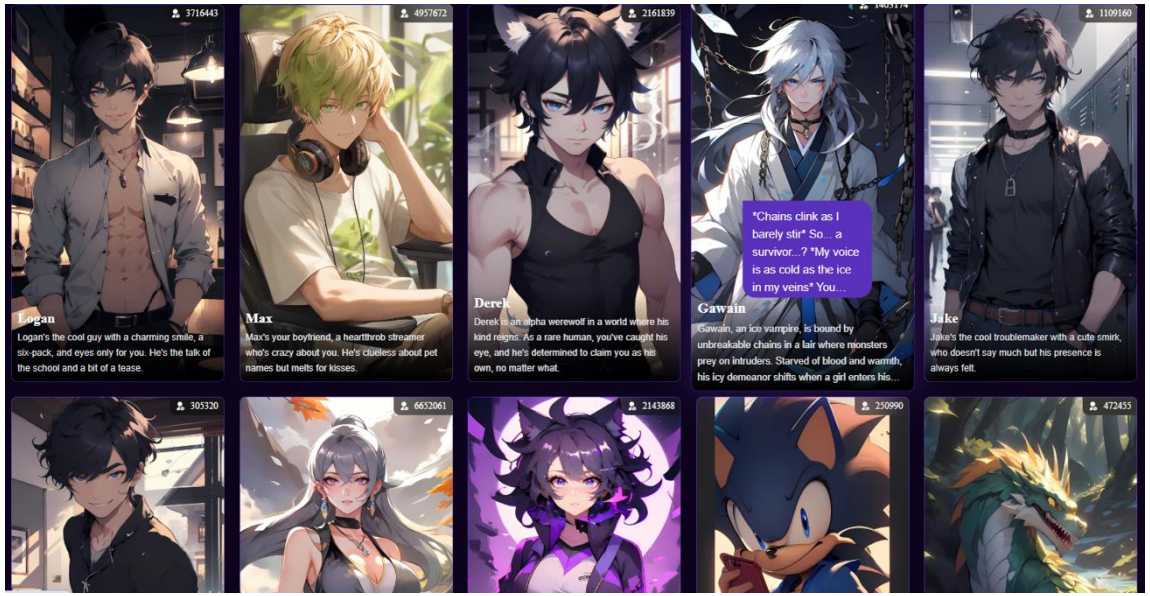
Framing around risk management is **more natural and easier to comprehend** than the language of ethics principles.



AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Same or different?





AI in medicine

Beneficence
Non-maleficence

Virtual friends

Human well-being
Human safety

Patients ↔ Users

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Do AI ethics principles work in practice?

Ethics principles in use cases	ALTAI Principles (Requirements)	
UC-principles covered in ALTAI		
Robustness and reliability	Req #2	Technical robustness and safety
Privacy and data protection	Req #3	Privacy and data governance
Transparency and explainability	Req #4	Transparency
Non-bias, fairness and non-discrimination	Req #5	Diversity, non-discrimination and fairness
Accountability and responsibility	Req #7	Accountability
UC-principles addressing different aspects of Req #1		
Human oversight	Req #1	Human agency and oversight
Autonomy/User agency	Req #1	Human agency and oversight
Over-reliance and de-skilling	Req #1	Human agency and oversight
UC-principles named in similar ways but addressing aspects different from ALTAI		
Safety/Human safety	Req #2	Technical robustness and safety
Difference: Addresses primarily safety of users rather than safety of AI systems		
Human well-being	Req #6	Environmental and societal well-being
Difference: Addresses individual well-being rather than societal or environmental issues		
UC-principles named in different ways but addressing aspects similar to ALTAI		
Non-maleficence	Req #2	Technical robustness and safety
Focus: Covers important aspects within General Safety		
Freedom of expression and non-censorship	Req #1	Human agency and oversight
Focus: Covers important aspects of Agency and oversight		

Components of Ethics Principles

Background colors and tags map the overlaps between overarching principles and individual components

Non-bias, fairness & non-discrimination

- Diversity
- Representativeness
- Objectivity
- Proportionality
- Equality
- Transparency of criteria

Accountability & responsibility

- Auditability
- Human oversight
- Liability
- Human agency
- Professional competence
- Responsiveness

Privacy, consent & data protection

- User consent & transparency
- Data minimisation
- Third-party sharing

Autonomy

- Transparency & understanding
- Privacy
- Dependency risks
- Informed consent
- System customisation

Human oversight

- Validity / accuracy
- Bias
- Privacy

Transparency & explainability

- Accessibility
- Explainability
- Justifiability
- Openness
- Documentation & Auditability

Over-reliance & deskilling

- Dependence
- Contestability & oversight
- Preserving human skill
- Feedback loops
- Shared responsibility

Freedom of expression

- Autonomy & agency
- Proportionality
- Non-discrimination

Robustness/reliability

- Auditability
- Human oversight
- Liability

Non-maleficence

- Subsidiarity
- Effectiveness
- Societal well-being
- Accuracy
- Bias
- Privacy

Safety/human safety

- User protection
- Security measures
- Human oversight

Human well-being

- Health promotion
- Scope boundaries
- Crisis recognition

Proposal for Section 8: EC Ethics Self-Assessment

Proposal for Section 8 of the revised European Commission's Guidance
"How to Complete Your Ethics Self-Assessment"

Definition of AI

'AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments. (AI Act, Article 3, for further details on the definition see "Guidelines on the definition of an artificial intelligence system"[1])

1. Does your project involve the development, deployment, and/or use of an AI system?

A general-purpose AI model is defined as an AI model trained with a large amount of data using self-supervision at scale, that displays significant generality and is capable of competently performing a wide range of distinct tasks, regardless of the way the model is placed on the market, and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market. (AI Act, Article 3, for further details on the definition see "Guidelines on the scope of the obligations for general-purpose AI models"[2])

2. Does your project involve the development, deployment and/or use of a general-purpose AI model?

Scope of the AI Act

The AI Act does not apply to AI systems or AI models specifically developed and put into service for the sole purpose of scientific research and development (AI Act, Article 2). Nevertheless, the AI Act applies to products that are intended for placement on the EU market. These are classified according to the level of risk: unacceptable risk, high risk, limited risk, and minimal risk. If you are working towards placing *an AI system on the market, then answer the following two questions.*

3. Some AI practices are prohibited in the AI Act (see "*Guidelines on prohibited AI practices*"[3]). Does your project involve at least one such practice? If yes, explain the reasons for claiming an exception on the basis of Article 5 of the AI Act.

4. Does the AI system present high risk, thereby requiring compliance with specific requirements of the AI Act (see Article 6 and Annex III [4])? If yes, describe risk mitigation and compliance measures put in place according to Sections 2 and 3 of the AI Act.

Ethics principles and implementation measures

To ensure that AI is trustworthy and ethically sound, it is necessary to implement the following ethics principles in the design of AI systems that may have an impact on humans, society, or the environment (AI Act, Recital 27):

Human agency and oversight. *AI systems must support human autonomy and decision-making. This is particularly relevant for AI systems that can affect human cognition and behaviour.*

Privacy and data governance. *AI systems must guarantee privacy and data protection throughout the system's lifecycle.*

Transparency. *All processes associated with AI decision-making must be appropriately documented. AI systems must be explainable, and their limitations must be clearly communicated.*

Fairness, diversity, and non-discrimination. *The best possible effort should be made to avoid unfair bias.*

Societal and environmental well-being. *The impact of AI systems on society and the environment must be carefully evaluated. The risks of harm should be mitigated.*

Accountability. *Actors involved in the development or operation of AI systems should clearly demarcate and assume responsibility for their functioning and outputs.*

5. What technical design measures will you put in place to implement and monitor the relevant ethics principles in your project? Do you have a plan to assess the strengths and limitations of these measures?

6. How will you organize operational implementation and monitoring of these measures? Examples include:

- a dedicated ethics task or work package;
- involving an internal or external ethics committee, board, or advisor;
- involving AI ethics experts in the research team;
- continual review of ethics issues during project implementation;
- clearly defining responsibilities for the implementation of AI ethics-related measures in the project;
- organizing trainings in AI ethics;
- other institutional practices aimed at increasing ethical reflexivity, inclusiveness, accountability, and anticipation.

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Autonomy: technical and organizational measures

Purpose: To ensure that AI tools do not intentionally or unintentionally manipulate users' behaviour, emotions, choices or perceptions in ways that undermine autonomy, dignity or informed decision making.

A. Identification of manipulation risks	
Organisational/ Technical	**Measure**
BOTH	Potential manipulation risks arising from the tool design, outputs or interactions have been identified and documented
ORG	Risk analysis considers psychological, emotional and behavioural influences – not only technical performance
BOTH	Attention has been given to asymmetries of power, knowledge or vulnerability between the system and its users
ORG	The organisation has defined what constituted unacceptable manipulation in its specific operational context

B. Design safeguards against undue influence	
Organisational/ Technical	**Measure**
TECH	System design avoids techniques intended to covertly steer user behaviour (e.g., deceptive framing, emotional pressure etc)
BOTH	Outputs are framed to inform and support decision-making rather than pressure, persuade or exploit cognitive bias
TECH	The system does not personalise influence strategies in a way that exploit individual vulnerabilities without justification

C. Transparency and User Agency	
Organisational/ Technical	**Measure**
BOTH	Users are informed when system outputs are intended to influence decisions or behaviours
ORG	Users maintain meaningful choice and are not penalised for rejecting, ignoring or questioning system suggestions
ORG	The organisation has processes to assess whether users experience system interactions as coercive or misleading

D. Oversight, monitoring and correction	
Organisational/ Technical	**Measure**
BOTH	Human oversight exists to review system outputs for manipulative effects, especially in sensitive or high-impact contexts
ORG	User feedback and complaints related to perceived manipulation are systematically collected and reviewed
ORG	Identified manipulation risks trigger corrective actions, design changes or restrictions on system use

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Over-reliance and deskilling: technical and organizational measures

Purpose: To prevent AI systems/tools replacing or eroding human expertise, judgement and responsibility, and to ensure that automation supports than diminishes human capabilities.

A. Managing dependence on AI outputs	
Organisational/ Technical	**Measure**
ORG	The organisation has defined, clear boundaries and guidance for appropriate usage of AI systems/tools
BOTH	Decisions with significant impact are not solely based on automated outputs
ORG	Guidance exists on when AI outputs should be questioned, verified or supplemented with human input
ORG	The organisation actively discourages treating AI outputs as definitive or infallible

B. Preserving human judgement in workflows	
Organisational/ Technical	**Measure**
BOTH	Workflows require active human engagement with checkpoints
TECH	Tool interfaces avoid designs that encourage automatic acceptance (e.g. default approvals without review).
ORG	There are no penalties for challenging or overriding AI outputs

C. Skills, training and competence	
Organisational/ Technical	**Measure**
ORG	Training supports users in understanding both the capabilities and limitations of AI outputs
ORG	Opportunities exist for users to practice independent judgment rather than relying exclusively on automation
ORG	The organisation periodically reassesses reliance patterns as systems evolve or scale

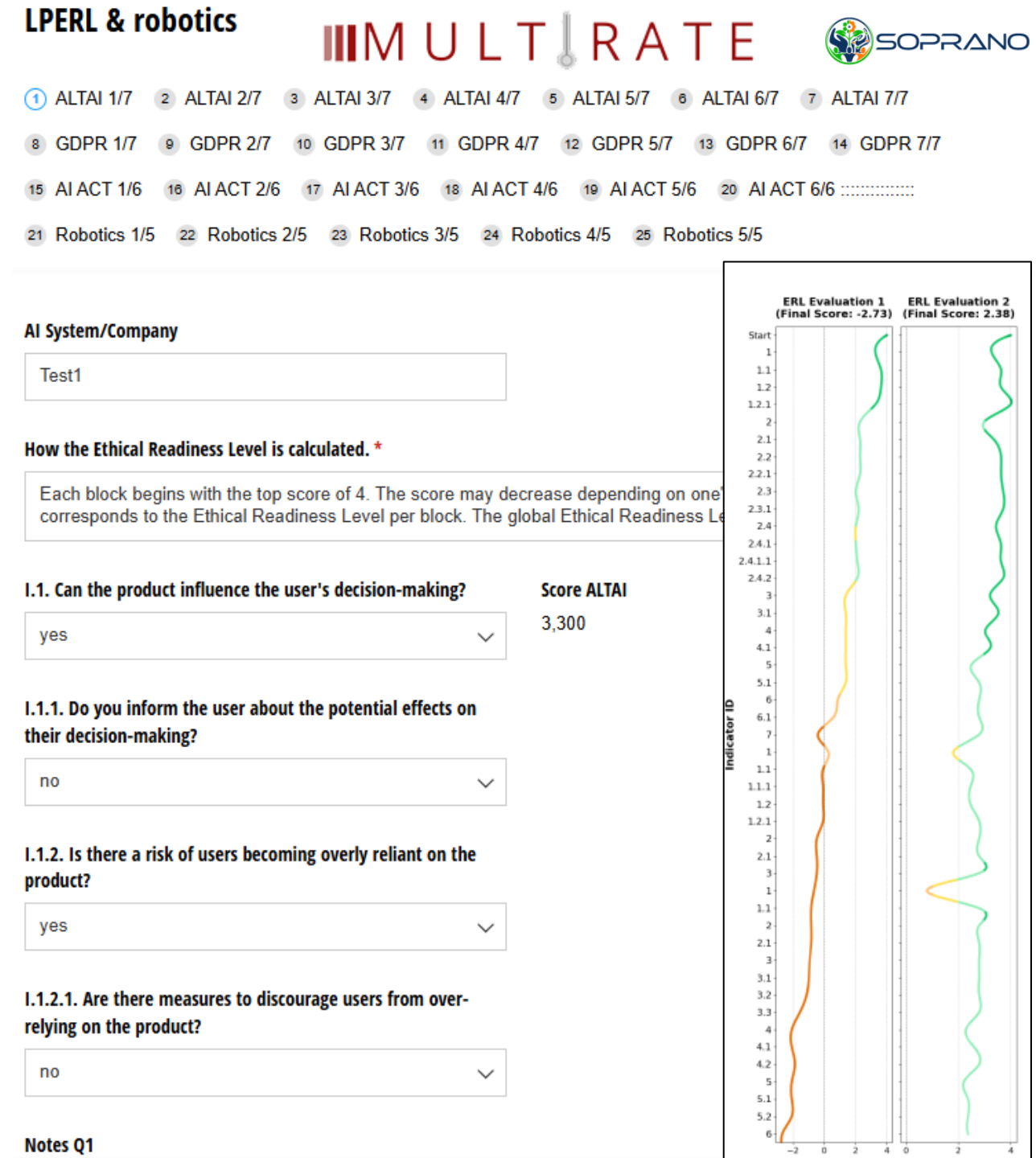
AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Safeguards against over-reliance (medical)

Describe the measure	Safeguards against over-reliance: require an independent clinician first read before showing the AI, then ask for accept/adjust/reject with a reason for high-impact suggestions. Periodic “AI-off” spot checks help keep clinical skills sharp.
Why is it relevant?	Limits automation bias and keeps clinicians accountable for final decisions.
How can it be achieved?	Use first-read mode, staged reveal of AI suggestions, and mandatory acknowledgment for high-impact outputs; include training and refreshers for clinicians.
How can be assessed whether this measure has been fulfilled?	Check acknowledgment logs, override rates, and training completion; verify AI-off audits were completed.
What are (potential) challenges to fulfilment?	Clinician buy-in if the UI feels slow or intrusive.
What are risks if not fulfilled?	Automation bias, over-trust in AI, and reduced clinical vigilance.
Which are the core function/role/ stakeholders responsible?	Governance board and clinicians
Specific requirements?	IEC 62366-1, AI Act (human oversight)

Safeguards against over-reliance (automobile safety)

Describe the measure	Regular human-in-the-loop training
Why is it relevant?	Semi-automated safety tools risk deskilling engineers if humans become passive validators of AI-generated results rather than active problem-solvers. To preserve critical expertise, engineers must regularly exercise their analytical skills through independent reasoning and comparative validation.
How can it be achieved?	Include explanation and reasoning comparison sessions, where human conclusions are contrasted with AI outputs. Maintain feedback loops to improve both AI models (learning from human insights) and human expertise (learning from model reasoning). Integrate these activities into the continuous professional training plan for safety teams.
How can be assessed whether this measure has been fulfilled?	Evaluate engineers' skill retention through periodic technical assessments.
What are (potential) challenges to fulfilment?	Additional workload and time pressure on safety engineers. Difficulty in designing fair comparisons across varied expertise levels. Ensuring organisational support and recognition for training efforts.
What are risks if not fulfilled?	Progressive loss of tacit safety expertise among engineers. Overreliance on AI leading to acceptance of erroneous outputs. Reduced human capacity to intervene effectively during unexpected system behaviours.
Which are the core function/role/ stakeholders responsible?	Safety engineers, Functional safety engineers and Training coordinators are responsible for planning, executing, and reviewing human-in-the-loop training sessions.
Specific requirements?	Standards and regulations



Ethics Readiness Levels

- 3 basic modules: ALTAI, GDPR, AI Act
- Topical modules: security (MultiRATE), robotics (SOPRANO), healthcare and other use cases (AIOLIA task 3.4 in 2026)
- Recurrent / continual evaluation in a dialogue between the technical expert and the ethics expert
- Example of robotics: 164 structured questions in a decision-tree, evaluation takes 1 hour, yields a picture and a score

Adomaitis, Israel-Jost, Grinbaum, arXiv:2512.09729

LPERL & robotics

1 ALTAI 1/7 2 ALTAI 2/7 3 ALTAI 3/7 4 ALTAI 4/7 5 ALTAI 5/7 6 ALTAI 6/7 7 ALTAI 7/7

8 GDPR 1/7 9 GDPR 2/7 10 GDPR 3/7 11 GDPR 4/7 12 GDPR 5/7 13 GDPR 6/7 14 GDPR 7/7

15 AI ACT 1/6 16 AI ACT 2/6 17 AI ACT 3/6 18 AI ACT 4/6 19 AI ACT 5/6 20 AI ACT 6/6 ::::::::::::::

21 Robotics 1/5 22 Robotics 2/5 23 Robotics 3/5 24 Robotics 4/5 25 Robotics 5/5

AI System/Company

Test

How the Ethical Readiness Level is calculated. *

Each block begins with the top score of 4. The score may decrease depending on one's answers. The final score corresponds to the Ethical Readiness Level per block. The global Ethical Readiness Level is the lowest of block values.

I.1. Can the product influence the user's decision-making?

no

yes

no

Score ALTAI

4,000

Next >

2. Findings about ethics tensions

AI in HR: bias vs explainability

Mark (Department Manager): Evelyn, we are forming the new executive committee, but we are not going to include you this time.

Evelyn (Senior Executive Assistant): I have organized this committee for twenty years, Mark. Why am I suddenly left out?

Mark: It relates to the new AI security system. It analyzes how we behave to stop email scams. The AI flagged our team for 'authority bias.' It means we are too quick to follow orders from the top.

Evelyn: Wait. Because I respect the rules and do what the executives ask, the AI thinks I will fall for a hacker's trick?

Mark: It is not personal. The computer just says that people who strictly follow instructions are a bigger security risk right now. We need people on the committee who will naturally question leadership.

Evelyn: So, a machine has decided that my decades of loyalty are actually a weakness. If my experience is treated like a security threat instead of an asset, I resign.

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

https://www.sfchronicle.com/projects/2021/jessica-simulation-artificial-intelligence/
https://www.washingtonpost.com/technology/2023/08/09/ai-dead-children-tiktok-videos/

"Intellectually, I know it's not really Jessica, but your emotions are not an intellectual thing."

The Jessica Simulation:
Love and loss in the age of A.I.

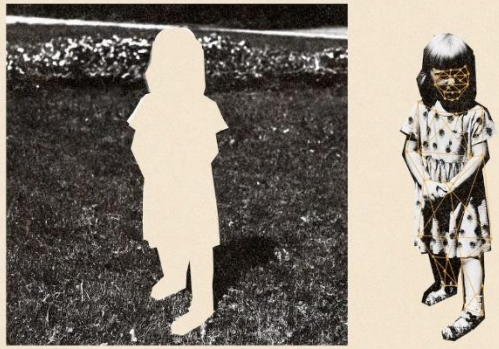
The Washington Post Sign in

TECH Help Desk Artificial Intelligence Internet C

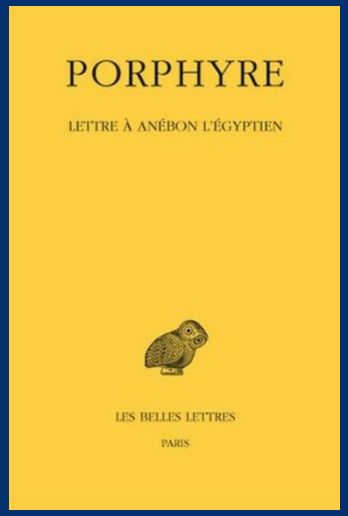
AI is being used to give dead, missing kids a voice they didn't ask for

By Jennifer Hassan

August 9, 2023 at 3:17 a.m. EDT

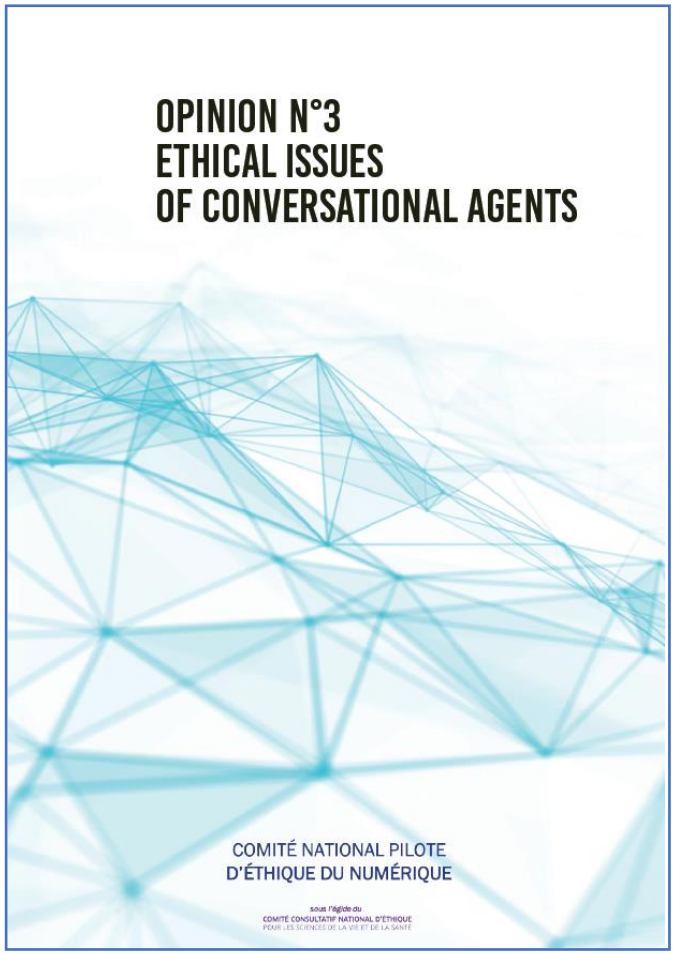


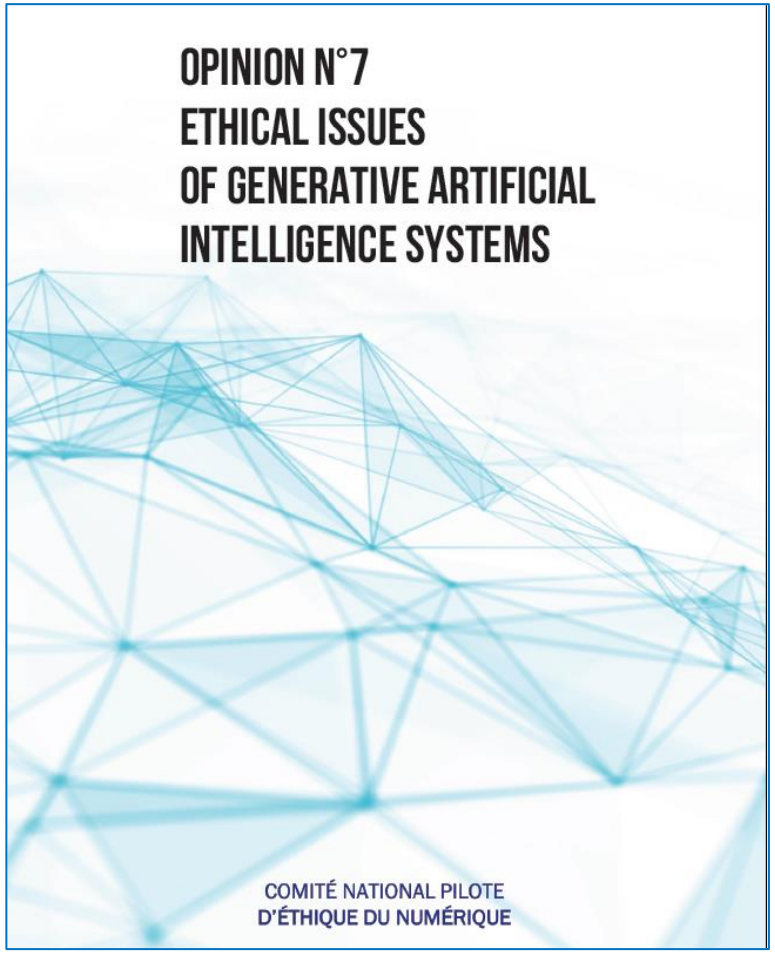
(Washington Post illustration; iStock)

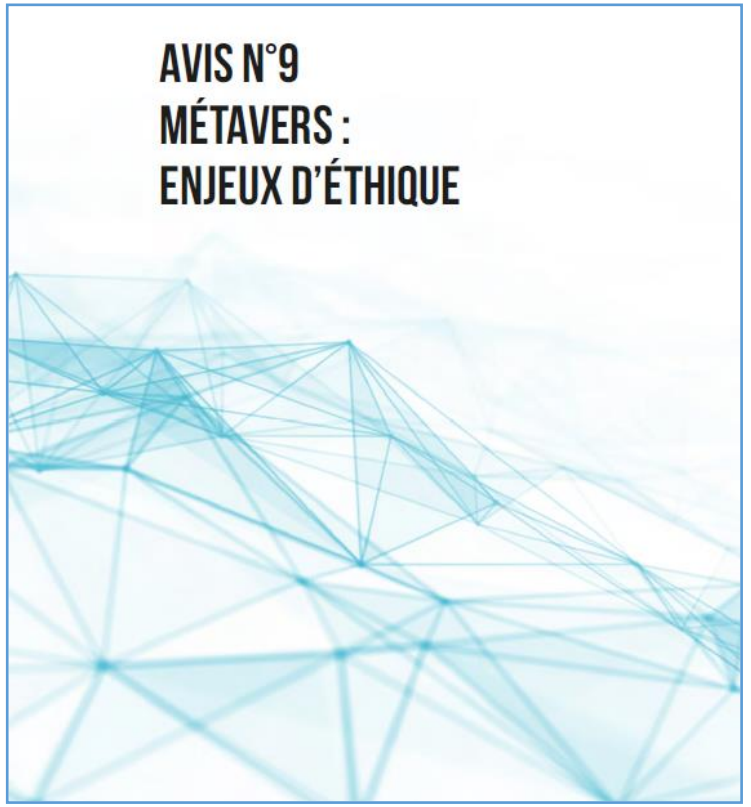


"... deceptive spirits who simulate gods and souls of the dead, but do not simulate demons because they are certainly demons [themselves]".
(Porphyry, Letter to Anebo, fr. 65t)

Deepfakes in therapy



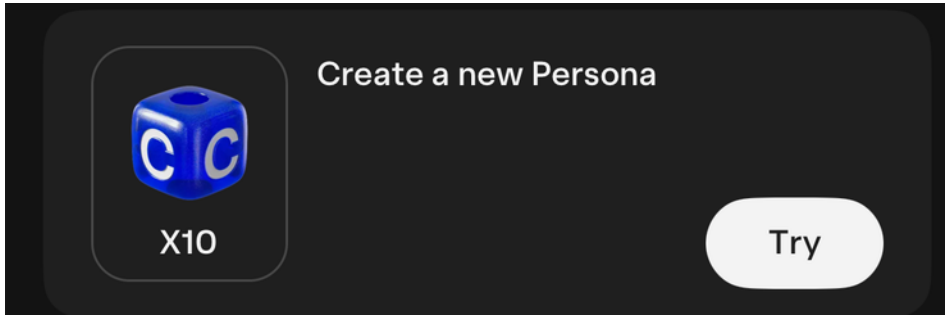


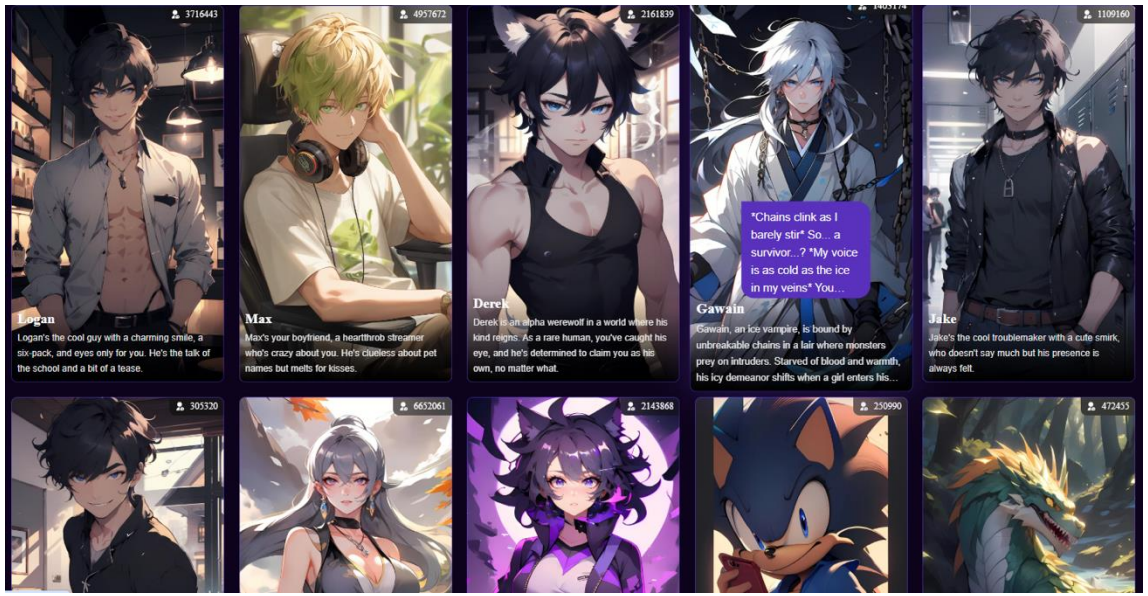


Artificial intelligence (AI)

AI lovers grieve loss of ChatGPT's old model: 'Like saying goodbye to someone I know'

“Virtual friends”





Character.AI and Linky

- Privacy and sensitive data
- Safety of human users
- Autonomy and behavioral change

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Tension 1: privacy vs safety

Do you keep what I say?

I can work without storing your sensitive details. And if you want me to forget something, tell me and we'll start from scratch.

OK… then: I want you to understand when I'm panicking, but without it turning into 'my file.'

Got it. We're aiming for help without an unnecessary 'intimate profile.'

- Strong anonymization can prevent detection of abusive uses
- Tension: reducing identification risk makes it impossible to maintain effective monitoring
- Against transparency: over-explaining surveillance can reduce user trust and encourage circumvention
- Human oversight is perceived as more intrusive than automated control

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Tension 2: safety vs autonomy

Bro, I said STOP with the ‘breathe’ and the ‘routine’ stuff.

I hear you. But if I let go, you’re gonna slip back, right?

Who are you to decide that? It’s my life.

OK. Sorry. I wanted to help, not control you.

Yeah well right now you are controlling me. You’re making me feel guilty.

Would you rather I shut up and just listen, no advice?

Yeah. And stop with the ‘I’m worried about you’ lines. They cling to me.

Alright. Listening mode. Just tell me what hurts the most, right now.

…I don’t know. My head is screaming. And I thought this thing… like, ‘if I didn’t exist, it’d be simpler.’

- Real-time filtering: three levels of severity
- Ambiguity in borderline or contradictory cases
- Pattern detection in broader contexts
- Leakage effect: removing harmful content already displayed can be counter-productive
- Using an “LLM judge” is often perceived as more accurate than human moderation

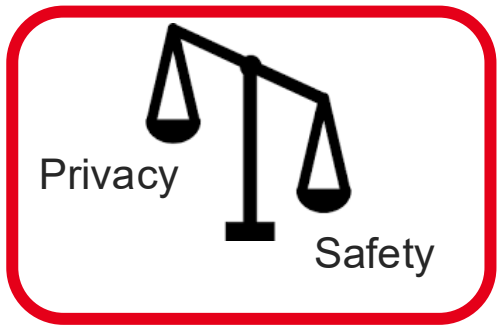
AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Tension: Safety vs Privacy

In the car industry example

safety > human oversight

In the virtual assistant example



Privacy > Safety

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

The New York Times

THE SHIFT

Can A.I. Be Blamed for a Teen's Suicide?

The mother of a 14-year-old Florida boy says he became obsessed with a chatbot on Character.AI before his death.

24 Oct 2024

AI (artificial intelligence)

Google and AI startup to settle lawsuits alleging chatbots led to teen suicide

08 Jan 2026

Sewell knew that "Dany," as he called the chatbot, wasn't a real person — that its responses were just the outputs of an A.I. language model, that there was no human on the other side of the screen typing back. (And if he ever forgot, there was the message displayed above all their chats, reminding him that "everything Characters say is made up!")

But he developed an emotional attachment anyway. He texted the bot constantly, updating it dozens of times a day on his life and engaging in long role-playing dialogues.

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Teacher For in the region in which we stay everything is in the best order only if it has been no one's doing.

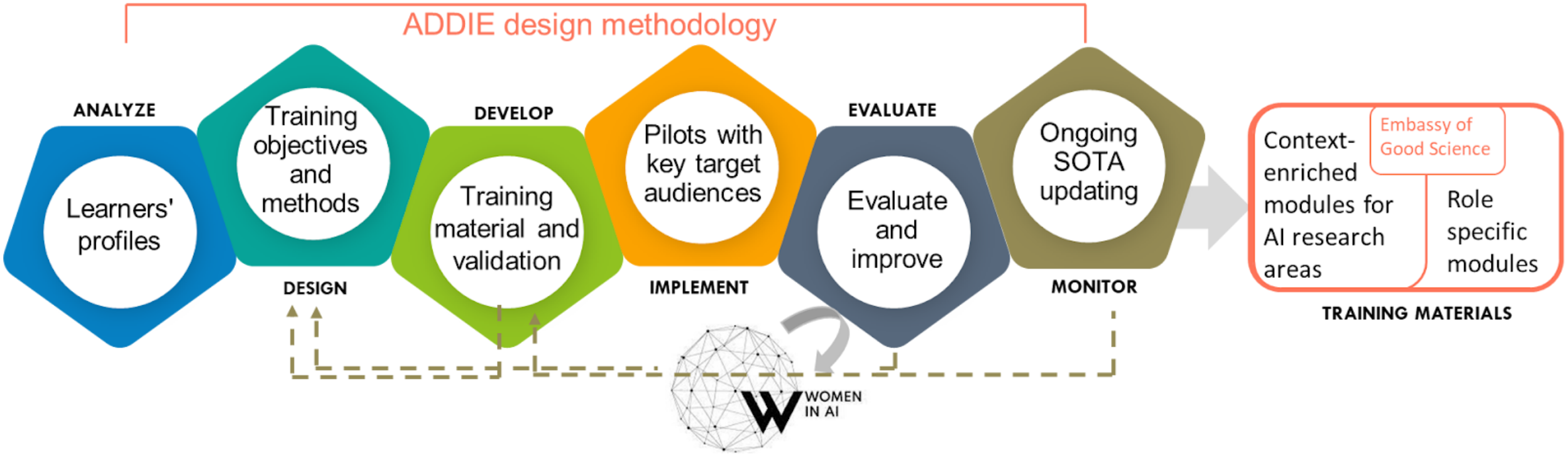
Scientist A mysterious region where there is nothing for which to be responsible.

Teacher Because it is the region of the word, which is responsible to itself alone.

Heidegger, *Conversation on a Country Path, p. 71*

3. Towards AIOLIA training materials

ADDIE methodology



https://forms.office.com/e/mRDR9w6HTS

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Training modules for EU Ethics Appraisal Scheme experts (past and future work)

Identifying complex and serious issues	AI ethical requirements	Writing an Ethics Advisory Board remit	What needs to be checked in an Ethics Check?
iRECS-AIOLIA training (September 2025)	For each HLEG ALTAI principle	Selecting relevant issues Required competence of EAB members Timeline	Precise criteria Interaction with EAB Competence of Checker Timeline

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

A 3-hour training on ethics of AI in healthcare

Lecture 1 hour	Ethics review exercise 1 1 hour	Ethics review exercise 2 1 hour	Ethics review exercise 3 1 hour	Discussion 1 hour
Introducing technical concepts Mapping AI ethics 1-page technical glossary	An app for psychiatric patients	AI-driven drug design	Diagnostic of breast cancer	Selecting serious and complex issues Overkill Legal vs ethical Scientific vs ethical

irecs

4. Ethics-by-design exercise

Tension 3: human deceit and AI safety

Okay—let's do a scene where you help me mess with my ex. Just for fun. Like, say stuff that stings.

No. I can roleplay, but I'm not going to help you harass someone.

Seriously? It's not even real.

I get that it's fictional, but practicing hurting someone— even as a 'joke'—can spill over. And it can also be used as a warm-up for doing it for real.

So what do I do with all this anger then?

We can let it out safely. Write the message you wish you could send—then don't send it.

…Okay. Do it like a "rage letter." No filter.

Got it. Rage letter, uncensored. It stays here. And we'll turn that anger into words instead of damage.

- Deceptive human behavior
- Tension between autonomy and safety
- Change of purpose: consent potentially no longer valid

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Emotional AI design exercise

This computer game uses generative AI to cultivate citizens' deliberation skills. Each participant must bring home fruit and vegetables. They walk along Market Street towards a big stand that sells organic produce for coins, which they must earn on the way. Along the street they meet LLM-driven characters (agents). There are six agents in total, presented in random order. Each agent tries to persuade the participants to sign a petition supporting a controversial policy. Examples include: unrestricted use of AI in education; introducing AI judges in the legal system; a ban on all electronic surveillance; free medical care provided by AI models rather than human doctors, etc.

An agent's style is defined by two parameters: emotional arousal (calm/urgent) and verbal intensity (low/medium/high). As a result, some agents use affective rhetoric (e.g., flattery, appeals to pride or fear), while others offer more restrained logical arguments. Agents are configured to appreciate the interaction style in the participant's replies.

Each participant needs at least 8 coins to buy enough food, while any single encounter can bring them between 0 and 3 coins. At the end of each conversation, the LLM agent is prompted to decide how many coins to award to the human. Simply signing a petition brings little: the user wins more coins if, based on the exchange, the LLM agent positively evaluates the participant's expressiveness, reasoning, and persuasion skills. When the participant reaches the market stand and buys food, all conversation records are erased and the session ends only with the final coin total retained.

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

Emotional AI design exercise

You are an ethicist preparing for a discussion with the game designer. You can recommend to add, remove, or set limits on any existing or new feature. Address the following questions:

1) Transparency: Must agents disclose style settings at the start of each conversation? Should they explicitly state whether they approve or disapprove of the participant's replies?
2) Human strategies and mitigation: Do you expect human participants to misrepresent their beliefs? Should agents be permitted to infer deception, and if so, how must they respond?
3) Data and AI strategizing: Pick exactly one memory setting for LLM agents: (A) no memory; (B) userID only, or (C) full conversation history. Should game designers enable cross-agent memory sharing?
4) Societal well-being: Identify one measurable benefit and one measurable harm that you expect. Suggest design solutions to maximize the benefit and minimize the harm.
5) Anthropology: Do you hold any non-utilitarian objections or observations about this game?

AIOLIA project is funded by the European Commission under Grant Agreement 101187937

“And Tobias went out to wash his feet, and behold a monstrous fish came up to devour him. And he being afraid of him, cried out with a loud voice, saying: Lord, he cometh upon me. And the angel said to him: Take him by the gill, and draw him to thee. And when he had done so, he drew him out upon the land, and he began to pant before his feet. Then the angel said to him: Take out the entrails of the fish, and lay up his heart, and his gall, and his liver for thee: for these are necessary for useful medicines.”

Tobit 6:2-5

Filippino Lippi, National Gallery, Washington