8. Artificial intelligence

This section concerns projects with activities involving the development or/and use of artificial intelligence (AI)-based systems or techniques.

⚠ The manner in which an AI solution is deployed or used may change the ethical characteristics of the system. It is therefore important to ensure ethics compliance even in cases where your project does not develop itself an AI based system/technique.

All AI-based systems/techniques must be ethical and comply with the principles and values as enshrined in the EU Charter of Fundamental rights and the EU Treaties. This requires specific ethically-focused approach during the design, development, deployment and/or use of AI based solutions.

Such approach must be built upon the following key prerequisites for ethically sound AI systems[8]:

Human agency and oversight — AI systems must support human autonomy and decision-making, enabling users to make informed autonomous decisions regarding the AI systems.

This is particularly relevant for AI systems that can affect human behaviour by guiding, influencing or supporting humans in decision-making processes *(e.g. recommendation systems, predictive algorithms, disease diagnosing tools)*. The right to human agency should be safeguarded by setting up appropriate oversight mechanisms to prevent possible adverse effects and uphold human autonomy.

⚠ AI systems must not subordinate, coerce, deceive or manipulate people, and should not create attachment or stimulate addiction.

Privacy and data governance — AI systems must guarantee privacy and data protection throughout the system's lifecycle. The principles of privacy by design and by default must be taken into account in the process of designing, developing, selecting and using AI. The quality, integrity and security of data should be rigorously checked and adequately managed.

Fairness, diversity and non-discrimination — Best possible efforts should be made to avoid unfair bias *(e.g. stemming from the used data sets or the ways the AI is developed)*.

Use diverse design teams and ensure participation of affected stakeholders to ensure objectivity and inclusiveness of the developed systems/approaches.

Accountability — Appropriate mechanisms should be set in place to ensure auditability and accountability of the AI solutions and their outcomes, both before and after their development, deployment and use. Potential negative impacts should be identified and addressed at early stages.

Transparency — All data sets and processes associated with AI decisions must be well communicated and appropriately documented. The principle of transparency is closely linked to the principles of tractability and explicability and facilitates the implementation of human agency, data governance and human

8 *As identified by the Independent High Level Expert Group on AI set up by the European Commission in the 'Ethics Guidelines for Trustworthy AI'.*

oversight. It includes all elements relevant to an AI system *(e.g. the data, the system and the processes by which it is designed, deployed and operated)*.

Societal and environmental well-being — The impact of the developed and/or used AI system/technique on the individual, society and environment must be carefully evaluated and any possible risk of harm must be avoided. Increased vigilance is needed for solutions that may potentially have significant negative social or environmental impact.

> ***Examples of social impact:*** *negative impact on human rights, democratic processes, functioning of media and mass communication, labour and labour market; educational choices; consumer interests and consumer protection, social cohesion and social exclusion, cultural diversity and cultural heritage, international co-operation, mass surveillance.*

⚠ The involvement of an **ethics advisor/mentor** with appropriate expertise in ethics of new and emerging technologies is highly recommended for projects which may raise significant ethics risks. This is particularly relevant for systems that have the potential to lead to significant negative individual, social and environmental impacts; stigmatisation or discrimination of people; interaction, replacement or influence on human decision-making processes.

At the development stage, the implementation of the key prerequisites for ethically sound AI systems can be achieved by adopting the 'ethics by design' approach. The latter is aimed at preventing ethics issues from occurring by integrating ethics values-based requirements into the design of the developed/used AI solution. The ethics by design approach will greatly facilitate your ethics compliance. For more information, please consult Guidelines on ethics by design for AI — *to be published soon*.

⚠ Some types of objectives, research methodologies, system architecture or design may be inherently problematic and could not be funded (due to serious ethical non-compliance). This is the case for instance for AI systems that risk to:

- limit human rights, subordinate, deceive or manipulate people, violate bodily or mental integrity, create attachment or addiction, or hide the fact people are interacting with an AI system
- cause people to be disadvantaged socially or politically, reduce the power that they have over their lives, or result in discrimination, either by the system, or by the way it will be used
- cause people to suffer physical, psychological or financial harm, cause environmental damage, or significantly damage social processes and institutions *(for example, by contributing to misinformation of the public)*.

8.1 Ethics issues checklist

Section 8: ARTIFICIAL INTELLIGENCE	YES/NO		Information to be provided	Documents to be provided/kept on file
Does this activity involve the development, deployment and/or use of Artificial Intelligence?	☐	☐	1) Explanation as to how the participants and/or end-users will be informed about: - their interaction with an AI	1) Detailed risk assessment accompanied by a risk mitigation plan (if relevant). These must cover the development, deployment and post-deployment

				system/technology (if relevant); - the abilities, limitations, risks and benefits of the proposed AI system/technique; - the manner in which decisions are taken and the logic behind them (if relevant). 2) Details on the measures taken to avoid bias in input data and algorithm design; 3) Explanation as to how the respect to fundamental human rights and freedoms (e.g. human autonomy, privacy and data protection) will be ensured; 4) Detailed explanation on the potential ethics risks and the risk mitigation measures.	phases. 2) Copies of ethics approvals (if relevant).
If **YES**:	Could the AI based system/technique potentially stigmatise or discriminate against people *(e.g. based on sex, race, ethnic or social origin, age, genetic features, disability, sexual orientation, language, religion or belief, membership to a political group, or membership to a national minority)*?	☐	☐	1) Detailed explanation of the measures set in place to avoid potential bias, discrimination and stigmatisation.	
	Does the AI system/technique interact, replace or influence human decision-making processes *(e.g. issues affecting human life, health, well-being or human rights, or economic, social or political decisions)*?	☐	☐	1) Detailed explanation on how humans will maintain meaningful control over the most important aspects of the decision-making process; 2) Explanation on how the	1) Information sheets/Template Informed consent forms (if relevant)

				presence/role of the AI will be made clear and explicit to the affected individuals.	
	Does the AI system/technique have the potential to lead to negative social *(e.g. on democracy, media, labour market, freedoms, educational choices, mass surveillance)* and/or environmental impacts either through intended applications or plausible alternative uses?	☐	☐	1) Justification of the need for developing/using this particular technology; 2) Assessment of the ethics risks and detailed description of the measures set in place to mitigate the potential negative impacts during the research, development, deployment and post-deployment phase.	For serious and/or complex cases: Algorithmic impact assessment/ human right assessment. These must cover the development, deployment and post-deployment phases.
	Does the AI to be developed/used in the project raise any other ethical issues not covered by the questions above *(e.g., subliminal, covert or deceptive AI, AI that is used to stimulate addictive behaviours, life-like humanoid robots, etc.)*?	☐	☐	1) Detailed explanation on how the potential ethics issues will be addressed and the measures set in place to mitigate ethics risks.	1) Detailed risk assessment accompanied by a risk mitigation plan. These must cover the development, deployment and post-deployment phases.

8.2 How to deal with the issues?

Your activities must comply with the ethics provisions set out in the Grant Agreement, and notably:

- highest ethical standards
- applicable international, EU and national law.

This implies that, amongst others, the developed/used AI solutions must:

- ensure that people are aware they are interacting with an AI system and are informed (in a language and terms understandable by all) about its abilities, limitations, risks and benefits. The manner in which this is done must be described in the research proposal

  ⚠ The manner in which information is provided should not depend on particular educational backgrounds, technical knowledge, or other skills which cannot be assumed of all people.

- prevent possible limitations on human rights and freedoms *(e.g. freedom of expression, access to information, freedom of movement etc.)*

- not be designed in a way that may lead to objectification, dehumanization, subordination, discrimination, stereotyping, coercion, manipulation of people or creation of attachment or addiction

- be able to demonstrate compliance with the principles of data minimisation and privacy by design and by default when processing personal data. The principles of lawfulness, transparency and fairness of the data processing must be respected at all times. For more information, please consult the Guidance on ethics and data protection in research projects

- must be designed in a way to avoid bias in both input data and algorithm design. The systems must be able to prevent potential discrimination, stigmatisation or any other adverse effects on the individual related to the use of the developed/deployed AI system/technique. The manner in which this is done must be described in your project proposal

- must address the potential impact on the individual, society or the environment. An evaluation of the potential negative individual, societal and/or environmental impacts must be carried out and be included in the project proposal along with the measures to be set in place to mitigate any potential adverse effect

  ⚠ The ethics risk assessment and risk mitigation measures must cover the development, deployment and post-deployment phases.

- must not reduce the safety and wellbeing of the individuals. Whenever relevant, the safety of the developed/used systems must be demonstrated in the project proposal

- should be developed in a way that enables human oversight (human-in-the-loop, human-on-the-loop, human-in-command), traceability and auditability. Whenever possible, explanation on how decisions are taken by the developed/used AI along with the logic behind it should be provided to the users.

For all issues related to the involvement of humans, data protection, safety and environmental impacts, please consult the relevant sections of this guidance.

8.3 What do you need to provide?

If your proposal raises one of the issues listed in the ethics issue checklist above, you must complete the **ethics self-assessment** in **Part A** of your proposal.

Your grant proposal must include the **information** referred to in the ethics issues checklist and any of the **documents** already available. Documents that are not submitted together with the proposal should be kept on file and may have to be provided later on, if requested by the granting authority.

Background documents & further reading

Ethics guidelines for trustworthy AI, Independent High Level Expert Group on AI

Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment, Independent High-Level Expert Group on AI

Guidelines on ethics by design/operational use for AI — *to be published soon*