European Commission

Ethics By Design and Ethics of Use Approaches for Artificial Intelligence

Version 1.0
25 November 2021

Acknowledgements: The preparation of this document was coordinated by Albena KUYUMDZHIEVA (DG R&I, currently EISMEA). The document has been written by Brandt DAINOW and Philip BREY and reviewed by Giovanni COMMANDE, Gemma GALDON CLAVELL, Iñigo DE MIGUEL BERIAIN; Bernd STAHL, Laurence BROOKS and the staff members of the Ethics and Research Integrity Sector, DG R&I: Lisa DIEPENDAELE, Francesco DURANTI (currently HADEA), Edyta SIKORSKA, Mihalis KRITIKOS and Yves DUMONT.

The Ethics and Research Integrity Sector, DG R&I would like to thank all contributors.

European Commission
DG Research & Innovation RTD.03.001-
Research Ethics and Integrity Sector
E-mail: RTD-ETHICS-REVIEW-HELPDESK@ec.europa.eu
ORBN B-1049 Brussels/Belgium

0. Introduction

This Guidance concerns all research activities involving the development or/and use of artificial intelligence (AI)-based systems or techniques, including robotics.[1] It builds on the work of the Independent High-Level Expert Group on AI and their 'Ethics Guidelines for Trustworthy AI' as well as on the results of the EU-funded SHERPA and SIENNA projects.[2]

This document offers guidance for adopting an ethically-focused approach while designing, developing, and deploying and/or using AI based solutions. It explains the ethical principles which AI systems must support and discusses the key characteristics that an AI-based system/ applications must have in order to preserve and promote:

- respect for human agency;
- privacy, personal data protection and data governance;
- fairness;
- individual, social, and environmental well-being;
- transparency;
- accountability and oversight.

Furthermore, it details specific tasks which must be undertaken in order to produce an AI which possess these characteristics.

For researchers who intend to use existing AI-based systems for their research, this document details ethical features which should be present in those systems to enable their use.

The central approach used in this guidance is known as "Ethics by Design." The aim of Ethics by Design is to incorporate ethical principles into the development process allowing that ethical issues are addressed as early as possible and followed up closely during research activities. It explicitly identifies concrete tasks which can be taken and can be applied to any development methodology (e.g. AGILE, V-Method or CRISP-DM). However, the advised approach should be tailored to the type of research being proposed keeping also in mind that ethics risks can be different during the research phase and the deployment or implementation phase. The Ethics by Design approach presented in this guideline offers an additional tool for addressing ethics-related concerns and for demonstrating ethics compliance. The adoption of the ethics by design approach, however, does not preclude additional measures to ensure adherence to all major AI ethics principles and compliance with the EU legal framework, in order to guarantee full ethical compliance and implementation of the ethical requirements.

The suggested approach is not mandatory for Horizon Europe applicants and beneficiaries, but aims to offer additional guidance for addressing ethics-related concerns and for demonstrating ethics compliance. This document is divided in three main parts:

- *Part 1: Principles and requirements*: This part defines the ethical principles that AI systems should adhere to and derives requirements for their development;

---

[1] For a comprehensive definition of AI-based systems and applications referred to here, please see: https://digital-strategy.ec.europa.eu/en/library/definition-artificial-intelligence-main-capabilities-and-scientific-disciplines

[2] The proposed approach is built on the principles elaborated by the High-Level Expert Group on AI in their *Ethics Guidelines for Trustworthy AI (https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai)* High-Level Expert Group on Artificial Intelligence, as well as on the results of the EU-funded SHERPA (https://www.project-sherpa.eu) and SIENNA projects (https://www.sienna-project.eu).

- *Part 2: Practical steps for applying Ethics by Design in AI development*: This section explains the Ethics by Design concept and relates it to a generic model for the development of AI systems. It defines the actions to be taken at different stages in the AI development in order to adhere to the ethics principles and requirements listed in Part 1;
- *Part 3: Ethical deployment and use presents guidelines for deploying or using AI in an ethically responsible manner*.

1. PART I – ETHICS BY DESIGN: PRINCIPLES & REQUIREMENTS

This part defines the ethical principles that AI systems should adhere to and derives the requirements which the AI system must comply with.

The ethical requirements embody the principles as characteristics of the system. While many of the ethical requirements are backed by legal requirements, ethical compliance cannot be achieved by adhering to legal obligations alone. Ethics is concerned with the protection of individual rights like freedom and privacy, equality and fairness, avoiding harm and promoting individual well-being, and building a better and more sustainable society often anticipating solutions that eventually becomes legal requirements to comply with.

Ethical Principles and Requirements

There are six general *ethical principles*[3] that any AI system must preserve and protect based on fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (EU Charter), and in relevant international human rights law:

1. **Respect for Human Agency:** human beings must be respected to make their own decisions and carry out their own actions. Respect for human agency encapsulates three more specific principles, which define fundamental human rights: **autonomy, dignity** and **freedom**.
2. **Privacy and Data governance:** people have the right to privacy and data protection and these should be respected at all times;
3. **Fairness:** people should be given equal rights and opportunities and should not be advantaged or disadvantaged undeservedly;
4. **Individual, Social and Environmental Well-being:** AI systems should contribute to, and not harm, individual, social and environmental wellbeing;
5. **Transparency:** the purpose, inputs and operations of AI programs should be knowable and understandable to its stakeholders;
6. **Accountability and Oversight:** humans should be able to understand, supervise and control the design and operation of AI based systems, and the actors involved in their development or operation should take responsibility for the way that these applications function and for the resulting consequences.

In order to embed these six ethical principles in their design, the proposed AI systems should comply with the general ethics requirements, listed below.

Respect for Human Agency

Respect for human agency encapsulates three more specific principles, which define fundamental human rights: **autonomy, dignity** and **freedom**.

[3] The ethical principles described in this document are informed by the work of the Independent High-Level Expert Group on AI (AI-HLEG) set up by the European Commission. They are also based on value frameworks proposed by the European Group on Ethics in Science and New Technologies, Statement on Artificial Intelligence, Robotics and 'Autonomous' Systems, 2018, the Institute of Electrical and Electronics Engineers (IEEE), the Organisation for Economic Co-operation and Development (OECD) and UNESCO.

Autonomy: Respecting autonomy means allowing people to think for themselves, decide for themselves what is right and wrong, and choose how they want to live their lives. It is important to note that AI systems can restrict human autonomy without doing anything - simply by not catering for the full range of human variation in lifestyle, values, beliefs and other aspects of our lives which make us unique. Hence, such restrictions may arise without any malevolent intent, but solely as a result of a lack of understanding about peoples' lives, beliefs, values, and preferences. This is a particular problem with personalisation services which do not take into consideration varying cultural norms. In addition, personalisation services, even when taking into consideration varying cultural norms, might restrict the information and options provided. AI based systems should avoid restricting unreasonably individual decision-making context (see also below under "freedom").

Dignity: Dignity entails that every human being possesses an intrinsic worth which should never be compromised. Hence, people should not be instrumentalized, objectified or dehumanized, but must be treated with respect at all times, including when using or being subjected to AI-based systems.

Freedom: Respecting freedom requires that people are not constrained in taking those actions which they should be able to pursue as autonomous persons, such as freedom of movement, freedom of speech, freedom of access to information, and freedom of assembly. In addition, freedom requires the absence of constraints which undermine peoples' autonomy, such as coercion, deception, exploitation of vulnerabilities, and manipulation. However, freedom is not absolute but limited by law.

Human Agency: General Ethical Requirements

- End-users and others affected by the AI system MUST NOT be deprived of abilities to make basic decisions about their own lives or have basic freedoms taken away from them.
- It MUST be ensured that AI applications do not autonomously and without human oversight and possibilities for redress make decisions: about fundamental personal issues (e.g. affecting directly private or professional life, health, well-being or individual rights), that are normally decided by humans by means of free personal choices; or about fundamental economic, social and political issues, that are normally decided by collective deliberations, or similarly significantly affects individuals.
- End-users and others affected by the AI system MUST NOT be in any way subordinated, coerced, deceived, manipulated, objectified or dehumanized.
- Attachment or addiction to the system and its operations MUST not be purposely stimulated. This should not happen through direct operations and actions of the system. It also should be prevented, as much as possible, that systems can be used for these purposes.
- AI applications should be designed to give system operators and, as much as possible, end-users the ability to control, direct and intervene in basic operations of the system.
- End-users and others affected by the AI system MUST receive comprehensible information about the logic involved by the AI, as well as the significance and the envisaged consequences for them.

Privacy & Data Governance

The rights to privacy and data protection are fundamental rights which must be respected at all times. AI systems must be built in a way that embeds the principles of data minimisation and data protection by design and by default as prescribed by the EU's General Data Protection Regulation (GDPR). For more information, please consult the Guidance Note on Ethics and data protection.

Privacy rights must be safeguarded by data governance models that ensure data accuracy and representativeness; protect personal data and enable humans to actively manage their personal data and the way the system uses it. Appropriate personal data protection can help developing trust in data sharing and facilitate data sharing models uptake. Data minimisation and data protection should never be leveraged to hide bias or avoid accountability, and these should be addressed without harming privacy rights.

Importantly, ethical issues can arise not only when processing personal data but also when the AI system uses non-personal data (e.g. racial bias).

Can we think of adding requirements on data trusts that may help research participants negotiate data use?

I think that the section will be of added value to AI researchers if it provides tailored recommendations that are based on the different types of data and data models used in AI systems as each category raises different challenges: training data, model data, production data, knowledge data or analysis-to-data, data-to-analysis and data-and-analysis-to-lake models.

Privacy and Data Governance: General Ethical Requirements

- The AI systems MUST process personal data in a lawful, fair and transparent manner.
- The principles of data minimisation and data protection by design and by default MUST be integrated in the AI data governance models.
- Appropriate technical and organisational measures MUST be set in place to safeguard the rights and freedoms of data subjects (e.g. appointment of data protection officer, anonymization, pseudonymisation, encryption, aggregation). Strong security measures MUST be set in place to prevent data breaches and leakages. Compliance with the Cybersecurity Act[4] and international security standards may offer a safe pathway for adherence to the ethical principles.
- Data should be acquired, stored and used in a manner which can be audited by humans. All EU funded research must comply with relevant legislation and the highest ethics standards. This means that all Horizon Europe beneficiaries must apply the principles enshrined in the GDPR.

Fairness

Fairness entails that all people are entitled to the same fundamental rights and opportunities. This does not require identical outcomes, i.e., that people must have equal wealth or success in life. However, there should be no discrimination on the basis of the fundamental aspects of one's own identity which are inalienable and cannot be taken away. Various legislations already acknowledge a number of them, such as gender, race, age, sexual orientation, national origin, religion, health and disability. Procedural fairness requires that the procedure was not designed in a way that disadvantages single individuals or groups specifically. Substantive fairness entails that the AI does not foster discrimination patterns that unduly burden individuals and/or groups for their specific vulnerability.

[4] Regulation (EU) 2019/881 of the European Parliament and of the Council of 17 April 2019 on ENISA (the European Union Agency for Cybersecurity) and on information and communications technology cybersecurity certification and repealing Regulation (EU) No 526/2013 (Cybersecurity Act).

Fairness can also be supported by policies which promote diversity. These are policies that go beyond non-discrimination by positively valuing individual differences, including not only characteristics like gender and race, but also people's diverse personalities, experiences, cultural backgrounds, cognitive styles, and other variables that influence personal perspectives. Supporting diversity means accommodating for these differences and supporting the diverse composition of teams and organisations.

Fairness: General Ethical Requirements

- *Avoidance of algorithmic bias:* AI systems should be designed to avoid bias in input data, modelling and algorithm design. Algorithmic bias is a specific concern which requires specific mitigation techniques. Research proposals MUST specify the steps which will be taken to ensure data about people is representative of the target population and reflects their diversity or is sufficiently neutral.
- Similarly, research proposals should explicitly document how bias in input data and in the algorithmic design, which could cause certain groups of people to be represented incorrectly or unfairly, will be identified and avoided. This necessitates considering the inferences drawn by the system which have the potential to unfairly exclude or in other ways disadvantage certain groups of people or single individuals.
- *Universal accessibility*: AI systems (whenever relevant) should be designed to be usable by different types of end-users with different abilities. Research proposals are encouraged to explain how this will be achieved, such as by compliance with relevant accessibility guidelines. To the extent possible, AI systems should avoid functional bias by offering the same level of functionality and benefits to end-users with different abilities, beliefs, preferences, and interests,.
- *Fair impacts:* Possible negative social impacts on certain groups, including impacts other than those resulting from algorithmic bias or lack of universal accessibility, may occur in the short, medium and longer term especially if the AI is diverted from its original purpose. This MUST be mitigated. The AI system MUST ensure that it does not affect the interests of relevant groups in a negative way. Methods to identify and mitigate negative social impacts in the medium and longer term should be well documented in the research proposal.

Individual, Social and Environmental Well-being

Individual well-being means people can live fulfilling lives, in which they are able to pursue their own needs and desires in mutual respect. Social well-being refers to the flourishing of societies, whose basic institutions, such as healthcare and politics, function well, and where sources of social conflict are minimized. Environmental well-being refers to the well-functioning of ecosystems, sustainability, and the minimization of environmental degradation.

AI systems should not contribute to any harm to individual, societal or environmental well-being, but instead AI systems should strive to make a positive contribution to these forms of well-being. To realize this goal, possible research participants, end-users, affected individual and communities and relevant stakeholders should be identified at the very early stage, to allow for a realistic assessment of how the AI system could enhance or harm their well-being. Documented choices should be made during development to support well-being and avoid harm.

Well-being: General Ethical Requirements

- AI systems MUST take into account all end-users and stakeholders and must not unduly or unfairly reduce their psychological and emotional well-being.

- AI systems should empower and to advance the interests and well-being of as many individuals as possible
- AI development MUST be mindful of principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects. Whenever relevant, there should be documented efforts to consider the overall environmental impact of the system and the Sustainable Development Objectives, where needed, steps to mitigate it. In the case of embedded AI this must include the materials used and decommissioning procedures.
- AI systems that can be applied in the area of media, communications, politics, social analytics, behavioural analytics online communities and services MUST be assessed for their potential to negatively impact the quality of communication, social interaction, information, democratic processes, and social relations (for example by supporting uncivil discourse, sustaining or amplifying fake news and deepfakes, segregating people into filter bubbles and echo chambers, creating asymmetric relations of power and dependence, and enabling political manipulation of the electorate). Mitigating actions must be taken to reduce the risk of such harms.
- AI and robotics systems MUST not reduce safety in the workplace. Whenever relevant, the application should demonstrate consideration of possible impact on workplace safety, employee integrity and compliance standards, such as with IEEE P1228 (Standard for Software Safety).

Transparency, explainability and objection

Transparency requires that the purpose, inputs, and operations of AI programs are knowable and understandable to its stakeholders. Transparency therefore impacts *all* elements relevant to an AI system: the data, the system and the processes by which it is designed and operated, as stakeholders must be able to understand the main concepts behind it (how, and for what purpose, these systems function and come to their decisions).

IP rights, confidentiality or trade secrets claims cannot prevent transparency as long as they can be preserved appropriately, for instance by way of selective transparency (e.g. confidentially to trustworthy third parties), technology or confidentiality commitments. Transparency is essential to realize other principles: respect for human agency, privacy and data governance, accountability, and oversight. Without transparency (meaningful information the purpose, inputs, and operations of AI programs), AI outputs cannot be understood, much less contested. This would make it impossible to correct errors and unethical consequences.

Transparency: General Ethical Requirements

- It MUST be made clear to end-users that they are interacting with an AI system (especially for systems that simulate human communication, such as chatbots).
- The purpose, capabilities, limitations, benefits, and risks of the AI system and of the decisions conveyed by it MUST be openly communicated to end-users and other stakeholders, including instructions on how to use the system properly.
- When building an AI solution, one MUST consider what measures will enable the traceability of the AI system during its entire lifecycle, from initial design to post-deployment evaluation and audit or in case its use is contested.
- Whenever relevant, the research proposal should offer details about how decisions made by the system will be explainable to users. Where possible this should include the reasons

why the system made a particular decision. Explainability is a particularly relevant requirement for systems that make decisions or recommendations or perform actions that can cause significant harm, affect individual rights, or significantly affect individual or collective interests.

- The design and development processes MUST address all the relevant ethical issues, such as the removal of bias from a dataset. The development processes (methods and tools) MUST keep records of all relevant decisions in this context to allow tracing how ethical requirements have been met.

Accountability by design, control and Oversight

Accountability for AI applications entails that the actors involved in their development or operation take responsibility for the way that these applications function and for the resulting consequences. Of course, accountability presupposes certain levels of transparency as well as oversight. To be held to account, developers or operators of AI systems must be able to explain how and why a system exhibits particular characteristics or results in certain outcomes. Human oversight entails that human actors are able to understand, supervise and control the design and operation of the AI system. Accountability depends on oversight: To be able to take responsibility and act upon it, developers and operators of AI systems must understand and control the functioning and outcomes of the system. Hence, to ensure accountability, developers must be able to explain how and why a system exhibits particular characteristics.

Accountability and Oversight: General Ethical Requirements

- It MUST be documented how possible ethically and socially undesirable effects (e.g. discriminatory outcomes, lack of transparency) of the system will be detected, stopped, and prevented from reoccurring.
- AI systems MUST allow for human oversight and control over the decision cycles and operation, unless compelling reasons can be provided which demonstrate such oversight is not required. Such a justification should explain how humans will be able to understand the decisions made by the system and what mechanisms will exist for humans to override them.
- To a degree matching the type of research being proposed (e.g. basic or precompetitive) and as appropriate, the research proposal should include an evaluation of the possible ethics risks related to the proposed AI system. This should include also the risk assessment procedures and the mitigation measures after deployment.
- Whenever relevant, it should be considered how end-users, data subjects and other third parties will be able to report complaints, ethical concerns, or adverse events and how these will be evaluated, addressed and communicated back to the concerned parties.
- As a general principle, all AI systems should be auditable by independent third parties (e.g. the procedures and tools available under the XAI approach[5] support best practice in this regard). This is not limited to auditing the decisions of the system itself, but covers also the procedures and tools used during the development process. Where relevant, the system should generate human accessible logs of the AI system's internal processes.

[5] https://github.com/EthicalML/xai

2. Part II - How To Apply Ethics By Design In AI Development: Practical Steps

Ethics by Design is an approach which can be used to ensure that the ethical requirements are properly addressed during the development of AI system or technique. *It is not the only possible approach.* However, this approach has been specifically designed to make it as clear as possible what is required, to explain the specific tasks which must be undertaken, and to help developers think about ethics while they are developing an AI system.

Why Ethics by Design?

For many AI projects, the relevant ethical issues may only be identified after the system's deployment, while for other projects these might be revealed during the development phase. Ethics by Design is intended to prevent ethical issues from arising in the first place by addressing them during the development stage, rather than trying to fix them later in the process. This is achieved by proactively using the principles as system requirements. What is more, since many requirements cannot be achieved unless the system is constructed in particular ways, ethical requirements sometimes apply to development processes, rather than the AI system itself.

Ethics by Design is described with a five-layer model. This model is similar to many others in Computer Science: higher levels are more abstract, with increasing levels of specificity going down the levels.

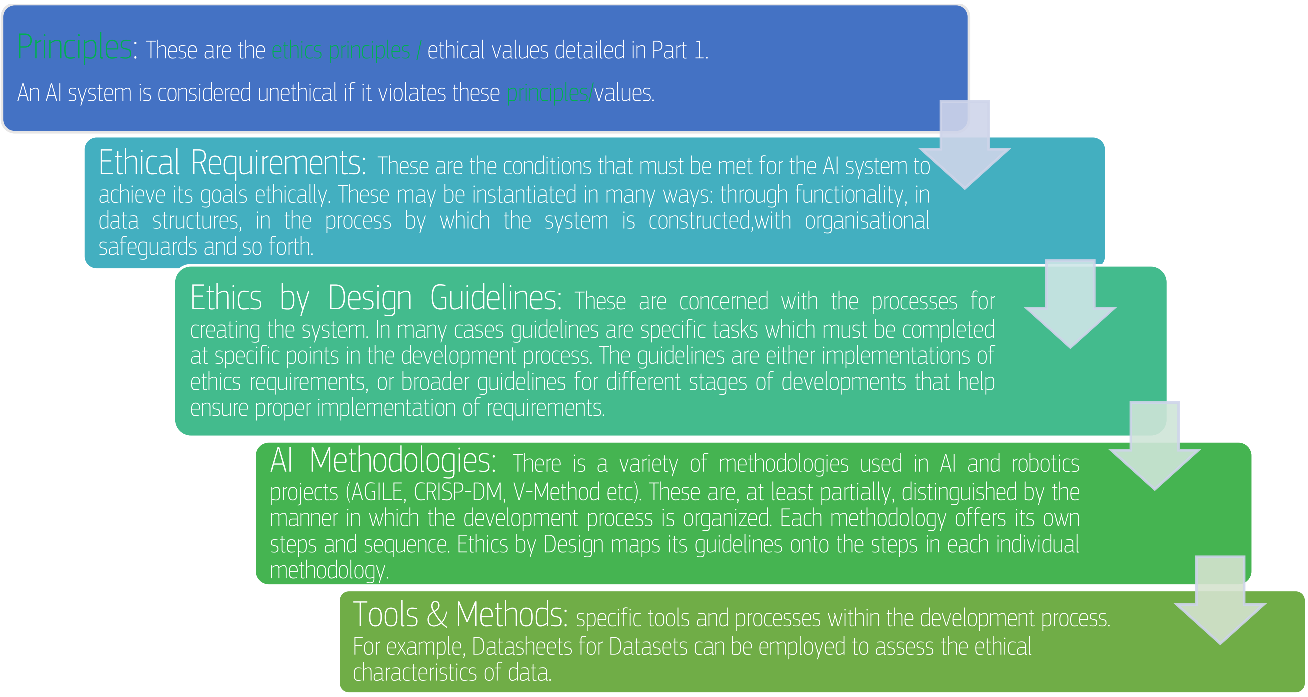

Figure 1: The 5-layer Model of Ethics by Design

The Development Process

The aim of Ethics by Design is to make people think about and address potential ethics concerns, while they are developing a system.

The main ethical requirements for AI and robotics systems above can be summarised as:

- AI systems must not negatively affect human autonomy, freedom or dignity.
- AI systems must not violate the right to privacy and to personal data protection. They MUST use data which is necessary, non-biased, representative and accurate.
- AI systems must be developed with an inclusive fair, and non-discriminatory agenda.
- Steps must be taken to ensure that AI systems do not cause individual, social or environmental harm, rely on harmful technologies, influence others to act in ways which cause harm or lend themselves to function creeps.
- AI systems should be as transparent as possible to their stakeholders and to their end-users.
- Human oversight and accountability are required to ensure conformance to these principles and address non-compliance.

Ethics by Design is premised on the basis that development processes for AI and robotics systems can be described using a generic model containing six phases. This section will describe the generic model, then outline the steps required to use it so as to incorporate Ethics by Design into your development process.

By mapping your own development methodology to the generic model used here, the relevant ethical requirements can be determined. Once this has been accomplished, the Ethics by Design will be embedded into your development methodology as tasks, goals, constraints and the like. The chance of ethical concerns surfacing is thus minimised because each step in the development process will contain measures to prevent them arising in the first place.

While the six phases of the generic model are presented here in a list format, this is not necessarily a sequential process. Steps may be iterative, such as in V-Model or CRISP-DM, or may deviate from waterfall models and may be incremental, such as in Agile. In each case, it should not be difficult to recognize similar tasks in one's favoured development methodology, and then map the tasks of the generic model onto steps or tasks in one's own approach.[6]

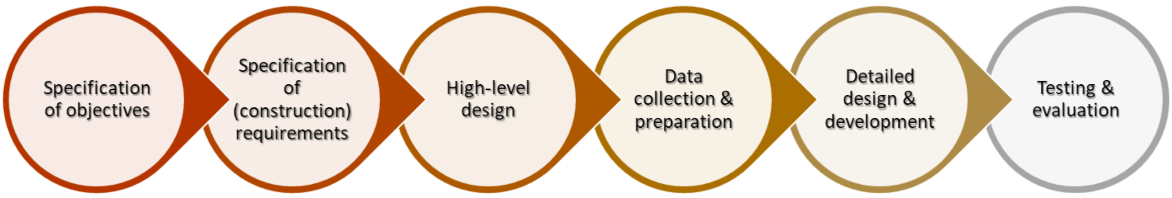

Figure 1: The Generic Model for AI Development

The six tasks in the generic model are:

1. ***Specification of objectives:*** The determination of what the system is for and what it should be capable of doing.
2. ***Specification of requirements*:** Development of technical and non-technical requirements for building the system, including initial determination of required resources, together with an initial risk assessment and cost-benefit analysis, resulting in a design plan.
3. ***High-level design:*** Development of a high-level architecture. This is sometimes preceded by the development of a conceptual model.

[6] In earlier versions of the Ethics by Design approach, such mappings have been provided for CRISP-DM, V-Model and Agile. See Brey et. al. (2019) for CRISP-DM and Brey et. al. (2020) for V-Model and Agile.

4. ***Data collection and preparation:*** Collection, verification, cleaning and integration of data.
5. ***Detailed design and development:*** The actual construction of a fully working system.
6. ***Testing and evaluation:*** Testing and evaluation of the system.

Each of these phases are discussed in detail below along with the relevant tasks to be undertaken to ensure ethics compliance.

Annex I provides an exemplary checklist, based on the requirements and ethical principles listed in Part 1 I. The ethics requirements and guidance should be applied to a degree matching the type of research being proposed (from basic to precompetitive).

1. Specification of Objectives

In this phase, the system's objectives are evaluated against the ethical principles and requirements presented in *Part 1.*

Potential ethical violations differ in their degree of seriousness. Some violations may only be hypothetical, highly unlikely, or less serious. If this is the case, the objectives of the system should not necessarily be abandoned, but concrete steps need to be taken to avoid unethical outcomes and their early and timely discovery and mitigation.

Be mindful that certain ethics issues may raise more serious ethics concerns than others. For example: systems that limit human rights, subordinate, deceive or manipulate people, violate bodily or mental integrity, create attachment or addiction, take away human control over major personal, moral or political decisions. Some of these systems may be deemed prohibited under EU law. Hence, some objectives are not ethically permitted under any circumstance. If the aim of a system is fundamentally incompatible with the ethical values and requirements, the project should not proceed. *Not everything that can be done should be done.*

Objectives that are likely to result in physical, psychological or financial harm to persons, environmental damage, or damage to social processes and institutions (for example, by supporting misinformation of the public) are not ethically acceptable either. If there is significant potential for social or environmental damage which could result from use of the technology, a social and/or environmental impact assessment should be conducted.

Furthermore, consider if the AI system may unintentionally (or intentionally) cause people to be disadvantaged socially or politically, or could result in unfair discrimination, either by the system, or by the way it will be used. If so, the system objectives should be modified and appropriate measures must be set in place to mitigate these risks.

NB! The ethical requirements for transparency, accountability and oversight do not usually apply to the objectives, but to the architecture and detailed design of the system. They usually have to be considered at this stage to determine the degree to which the system's objectives will allow for the required transparency and/or accountability and oversight to be built into the system.

⚠ When assessing objectives, always consider the potential for **intentional or accidental misuse**. Where possible, modify the system's objectives to reduce such potential. If the potential misuse is significant, conduct a risk assessment outlining the risks, the elements of the design which will need to be included to mitigate this, and any procedures required to reduce this risk once the system is deployed and operational.

The following two broad guidelines apply to the Specification of Objectives phase:

1. Assess whether the objectives for the design project will meet the relevant ethical requirements (see Part 1 and Annex 1). It is recommended that AI ethicist (if available), is enlisted to assess the objectives, in collaboration with members of the development team.
2. Include external stakeholders in both the specification of objectives and specification of requirements phases. Stakeholders may be aware of wider ethical issues which could arise from the use of the system. Whenever possible and relevant, stakeholders should be consulted about what ethical issues they believe are at stake and how these issues should be dealt with. Stakeholders should be appropriately diverse (gender, age, ethnicity, etc.) and include the major stakeholder groups that will be affected by the system – directly or indirectly. In this way an appropriately diverse range of ideas and preferences will inform design choices.

2. Specification of Requirements (technical and non-technical)

During this phase, development requirements, resources and plans are assessed against the ethical requirements.

The primary function of the Requirement Specification phase is to arrive at a development plan that includes design specifications for the system, a definition of the development infrastructure, a determination of staff resources required, the setting of milestones and other deadlines and so forth.

Most organisations have a standardised set of development tools used for all projects. The organisational and management structures and procedures are usually tuned to these tools, as are the development methodologies. Nevertheless, it cannot be assumed that any tool, process or organisational element will support Ethics by Design. Some of the ethical requirements present new problems during development. For example, it is no longer sufficient to merely correct datasets for bias, developers also need to document *that* this has been done and *how*. Consequently, requirements for human oversight and audit may impose a need to document many internal processes to a greater degree than has previously been the case. It must therefore be taken into account that previously relied on development methods, tools and organisational structures used on previous projects might need modification.

In some cases, suitable development tools that allow to meet the ethical requirement may not be readily available. However, it is required to be extremely rigorous in searching for suitable tools. The requirements set here are not unique to Horizon Europe, but are increasingly common demands of many AI projects. Consequently, the tools necessary to meet the relevant ethical requirements are being developed rapidly.

Not all ethical requirements may be relevant! The degree to which a technical inability to meet an ethical requirement can prevent a project from moving forward will depend on the particular ethical requirement in question and the system's functionality. For example, a system which approves personal loans must be able to explain each individual decision in a human-readable format because individual people will be affected by its decisions. By contrast, a system which manages appointments might only have a limited impact on the life of individuals, so the need for transparency is lower. Where it is believed it is technically impossible to meet a relevant ethical requirement, the importance of the requirement will be a factor in determining approval.

To a degree matching the type of research being proposed (from basic to precompetitive), the following broad guidelines apply to the Specification of Requirements phase:

1. The proposed design specifications, constraints, selected resources and infrastructure must be assessed for compatibility with the ethics requirements. For example, some deep learning techniques may not to be the best choice for transparency. It is appropriate to check these specifications and constraints against the ethical requirements of Part 1 and modify them if necessary to ensure a good fit.
2. Once a complete design plan has been produced (and to the degree matching the type of research being proposed), an *ethical risk and impact assessment* is recommended to assess specific ethical risks in development, deployment and use of the system. This should include an assessment of the risks associated with unintended uses and consequences of the system. Steps should be planned to avoid risks or mitigate those that are unavoidable. This risk assessment should be updated at regular points in the development process as more information becomes available. The ethical risk assessment should be planned and budgeted for. If available, it is recommended that it is conducted by an ethicist with suitable professional background. The assessment needs to be scaled to the nature of the project, the severity of ethical risks and the overall budget of the development project[7].
3. It is recommended that at this stage an *Ethics by Design (EbD) implementation plan* is composed, specifying all future steps to be taken to incorporate EbD in the development process and indicating the actors responsible for carrying out and monitoring the tasks. Where relevant, this implementation plan should also incorporate the ethical risk and impact assessment if one has been completed.
4. It is recommend, in addition, that it includes an ethical compliance architecture embedded into the development infrastructure and a set of organisational structures and procedures. The ethical compliance architecture will need to focus on tools and processes at the developer level, but will also need mechanisms for external communication from end-users and other stakeholders during testing and evaluation. A simple way forward is to relate the ethical compliance architecture with the legal compliance one.
5. It is recommended to include an ethical governance model[8] which incorporates organisational structures for governance of the EbD process (e.g. ethical review committees/ethics adviser/ ethics auditor).

3. High-level Design

High-level design is concerned with the development of an overall architecture of the system.[9]

Ethical requirements should be treated just the same as any other requirements for the system. Design should include functionality by which to programmatically support ethical requirements, such as keeping logs of internal data manipulation by the system. The requirements for transparency and

---

[7] Example of a standard for ethical impact assessment is available at https://satoriproject.eu/media/CWA17145-23d2017.pdf

[8] Ethics by design requires "ethical governance" - mechanisms for monitoring and maintaining adherence to the ethical requirements of the system during the design process. The exact manner in which ethical governance is done will depend on the size of the development organisation and the nature of the system being produced. The key principle is that the organisation has identified people responsible for ethical governance, clear mechanisms by which they work and the authority to enforce their decisions. In a small organisation, ethical governance could be a part-time role for a senior developer. By contrast, in a large organisation, ethical governance could be undertaken by a dedicated department spread vertically through the organisation, from board level representation to ethical compliance officers embedded in developer teams. Each type will have appropriate mechanisms for detecting ethical issues, evaluating them, requiring changes, enforcing their decisions and monitoring the results.

[9] In many cases this will include a hierarchical breakdown of the required sub-systems within the system, though some will consider this a part of detailed design.

human oversight will typically require additional features beyond what is needed to achieve the system's aim.

To a degree matching the type of research being proposed (from basic to precompetitive) and whenever applicable, the following broad guidelines apply to the High-level Design phase:

1. Verify that the design allows for an interface based on human-centric design principles which leave meaningful opportunities for human choice.
2. Design features and functions which will enable the capabilities and purpose of the system to be openly communicated to users and anyone else who may be affected by it.
3. Design mechanisms so that people will know when they are being subject to the decisions of the system. This may include operational procedures to be used once deployed.
4. Ensure there is no aspect of the AI system which could be mistaken for a human, without notice of it, once the system is deployed. Bear in mind many people may not have an understanding of AI and can unknowingly assume they are interacting with a person. For example, even when labelled as such, chatbots can be mistaken for humans by those who do not know what the term 'chatbot' means.
5. Verify that the design supports the ethical requirements for privacy, personal data protection, and data governance. Ensure development processes, procedures and tools do not expose personal data, such that it violates the fundamental rights of individuals. For example, error logs may needlessly include the personal data being accessed when a bug is encountered. It is especially important to ensure developers do not have access to identifiable personal information except where absolutely necessary and with GDPR compliant instructions.
6. Establish a formal process to guarantee the selection of data for the system will be fair, accurate and unbiased. Plan for an initial assessment of data sources before they are brought into the system. Design an auditable mechanism that records how data selection, data acquisition, storage and use happen (covering both the development process and use once operational). It cannot be *assumed* that the data obtained is the data one wanted. For example, datasets may be incomplete or methods of importing data may alter it in unexpected ways. Include in the design formal processes to check for and correct bias or errors after importing any data;
7. Evaluate whether the system could cause any physical harm to people, animals, property or environment. If this is possible, include design features to minimise the risk and/or the prevalence and severity of the potential harm (e.g. if the system will be able to respond to voice commands, you must include "emergency stop" vocal commands in the design).
8. Consider all possible negative impacts on relevant groups, including impacts other than those resulting from algorithmic bias or lack of universal accessibility and devise steps to mitigate any potential negative impact (e.g. stigmatisation, discrimination).
9. AI systems associated with media, communications, politics, social analytics, and online communities should be assessed for their potential to negatively impact the quality of communication, social interaction, information, democratic processes, and social relations. Where relevant, research proposals should detail mitigating actions which will be taken to reduce the risk of such harms.
10. Whenever relevant, you should consider the environmental impact of the system and, where needed, you should take steps to mitigate negative impacts. An initial environmental assessment should have been conducted during the objectives phase. Once high-level design of the system is complete, this assessment should be elaborated to demonstrate how the system will be constructed in an environmentally friendly way.

11. If relevant, demonstrate consideration of possible impact on safety and compliance with appropriate standards such as the IEEE P1228 (Standard for Software Safety), and include in the design features to ensure safety and compliance.
12. Consider how decisions made by the system will be explainable to users and other stakeholders. Where possible this should include the reasons why the system made a particular decision. The system (or those deploying it) should always have a mechanism by which to explain what the decision was and what data/features were used to make that decision. Explainability is particularly relevant for systems that make decisions, recommendations, or perform actions that may cause significant harm, affect individual rights, or significantly affect individual or collective interests.
13. Design procedures and select and configure tools which can document development processes in such a way that humans can understand and evaluate decisions made within the design and development processes. Research proposals should also explain how humans will be able to understand the decisions made by the system and what mechanisms will be designed for humans to override them. A layered approach to this documentation is recommended, so that it offers a range of technical detail, commencing with basic overviews, such as executive summaries, down to detailed schemas and other technical models. In this way people can be provided with documentation appropriate to their level of expertise and their specific concerns.
14. Ensure the design includes mechanisms by which the AI system will record its own decisions so that they can be subject to human review and accountability. Such review could occur through a post-deployment audit if data subjects or end-users question system behaviour, as part of an internal ethical governance review or external audit.
15. To the extent possible and appropriate, design mechanisms for accountability, human oversight and external audit after deployment. This may require additional functionality inside the system solely for reporting internal activity that have no role in the system's functionality. Mechanisms for oversight after deployment will need access to the oversight work performed during development.
16. Ensure in the design an ethical documentation system, sufficient to make ethical issues identifiable and their resolution traceable and explainable.
17. Include in the design a testing regime which can check whether the system's internal operations meet the ethical requirements. This may require changes to the way functionality is achieved within the system so as to permit appropriate testing and remedial action. The research proposal should provide details of how ethically and socially undesirable effects of the system will be detected, stopped, and prevented from reoccurring.
18. Whenever possible and relevant, ensure that there is a process by which both internal staff and third parties can report potential vulnerabilities, risks, or biases in the system, both during the development process and after deployment.
19. Ensure that the system is designed in a manner which permits external ethical auditing ensuring accountability is possible. If unsure, consult existing ethical audit procedures.
20. Design features and functions in a way that end-users are aware of both capabilities and limits of the AI system once deployed also to avoid function-creep issues, automation or translational biases.
21. Ensure that developers have sufficient training to develop awareness and future accountability practices (including knowledge about the legal framework applicable to the system).

In addition, research proposals are encouraged to explain how the design will accommodate universal accessibility, such as by demonstrating compliance with relevant accessibility guidelines.

This may include accessibility assessment of the interface and other touchpoints. The examination of the initial interface design and other touchpoints will reveal whether a one-size-fits-all approach is applied to users. If this is the case, you may need to modify the design or prepare a formal justification for it.

4. Data Collection and Preparation

Data collection and preparation is an especially critical phase as far as ethics are concerned.

It should be assumed that any data gathered is biased, skewed or incomplete until proven otherwise. Fairness and accuracy are the primary concerns here. In general, data gathered from human activity within any society, such as written communication or employment patterns, may reflect the biases in that society. It must therefore be actively demonstrated that data is accurate, representative or neutral before it can be trusted as such.

What is more, the preparation of the data itself may give rise to ethical issues as well. Steps should be taken to ensure that testing learning and/or algorithmic manipulation do not introduce new biases or give rise to other ethical issues (such as de-anonymization). Problems frequently arises where testing does not accurately reflect the real-world use after deployment, for instance creating automation or translational biases. For example, many facial recognition systems have poor performance with darker-skinned people due to testing on purely Caucasian populations.

Additional information on the ethics obligations related to personal data processing can be found it the Note on Ethics and Data Protection and the relevant section HE Guidance How to Complete Your Ethics Self-Assessment.

Bear in mind that when processing personal data, all EU funded research must comply with international, European, and national legislation and with the highest ethics standards. This means that EU beneficiaries must apply GDPR principles unless they are bound by higher standards of data protection. If using external organisations/service providers for data storage/analysis/collection ensure these are also compliant with data protection requirements.

To a degree matching the type of research being proposed (from basic to precompetitive) and whenever applicable, the following broad guidelines apply to the data collection and preparation phase:

1. Assess how operations within each process might violate ethical and/or data protection requirements. Make necessary changes as a result. If appropriate changes are not possible, the design objectives may need to be altered.
2. Ensure that the processing of personal data complies with the General Data Protection Regulation (GDPR) and other relevant national and EU legislation. Prepare a specific data protection policy which details how the project complies with data protection requirements. It is highly recommended to consult your data protection officer or person with relevant data protection expertise (if available). Be mindful that, whenever your system is processing personal data, you must comply with the data minimisation principle. This means that you must ensure that only data which is relevant, adequate and limited to what is absolutely necessary is processed by your system. You should also comply with the principles of data protection by design and by default and safeguard the rights and freedoms of the data subjects.
3. Specify the steps which will be taken to ensure data about people is representative and reflects their diversity or is sufficiently neutral. Similarly, you should document how errors will be

avoided in input data and in the algorithmic design which could cause certain groups of people to be represented incorrectly or unfairly.

4. Ensure that input, training and output data is all analysed for input bias (bias in data that results in unfair, unrepresentative or prejudiced representation of individuals and groups). In particular, verify, to the extent possible, that personal and group data accounts for diversity in gender, race, age, sexual orientation, national origin, religion, health and disability, and other social categories that are relevant to the task, and does not include prejudiced, stereotyping or otherwise discriminatory assumptions about people in these categories. Where it is determined that bias is possible, build mechanisms to avoid or correct it. If so, modify the criteria by which data will be selected or plan to rectify the datasets once they are in the system. The requirements for transparency and oversight demand that such rectification is documented.
5. Analyse your training data and ensure that your data is relevant and representative. Undertake a formal bias assessment of the data imported into the system. Do not assume any data imported into the system is unbiased – test it to the extent possible accordingly to the risks posed by the projected AI to fundamental rights and liberties. Assess the diversity and representativeness of users in the data, testing for specific populations or problematic use cases. Make sure data from one demographic group is not used to represent another unless it is justifiably representative. Evaluate the potential for harmful bias being introduced during the data preparation stage, such as inadvertently removing data relating to a minority group. Take steps to mitigate any such risk. Ensure that, whenever possible, there is an ability to go back to each state the system has been in to determine or predict what the system would have done at time t and, whenever possible, determine which training data was used.
6. Ensure the involved staff has appropriate training also on the applicable legal framework and the ethics concerns.

5. Detailed Design and Development

In the Detailed Design and Development phase, a fully working system is constructed. Actions which will incorporate the ethical requirements are added to the various tasks in the detailed design, as well as to the development infrastructure (tools, methodologies, procedures, and anything else that may affect exactly how something is built).

To a large degree, this phase involves adding more detail to the ethical requirements of the system, and to designing and implementing an ethical development architecture. Ethics by Design calls for ethical matters to be dealt with during the development phase, so existing development processes will need to support this activity. To integrate ethical requirements, ensure that ethical guidelines are communicated to all developers and engineers, and that the design is evaluated against the ethical implementation plan by them wherever they need to make relevant decisions. Issues that may be particularly relevant in this phase are those related to transparency, privacy and accountability. Appropriate training and awareness raising might be of help.

To a degree matching the type of research being proposed (from basic to precompetitive) and if relevant, the following broad guidelines apply to the detailed design and development phase:

1. If creating new (inferred) personal data (e.g., through estimation of missing data, the production of derived attributes and new records) or aggregating or creating new data sets, make sure all newly created personal data or data sets are given the same level of protection as previously collected or held personal data. Inferred personal data should not be able to alter the output of

the dataset or otherwise significantly affect data subjects' rights. Newly created personal data should not be misleading.

2. Ensure no new personal data is, or can be, collected or created during development of the system or during regular use of the system, unless necessary. If new personal data is collected or created, ensure mechanisms are in place that impose access or use limitations which will protect data subjects.
3. Ensure there are processes to safeguard the quality and integrity of all pertinent data, including processes for verifying that data sets have not been compromised or hacked. If you have control over the quality of external data sources used, assess to what degree you can validate their quality.
4. Make sure that roles and responsibilities for implementing the ethics requirements and architecture are clear and that all relevant staff understand the requirements.
5. Ensure there are oversight mechanisms for data processing (including limiting access to only appropriate personnel, mechanisms for logging data access and making modifications).In developing your ethics by design approach, consider that even when the data are anonymised, it may still be possible to de-anonymise it[10].
6. If you are mixing databases composed by personal and non-personal data, be aware that these should be processed as personal data, no matter the proportion of each type of data in the melted dataset. Ensure there is an embedded process that allows individuals to access their data and remove it from the system and/or correct errors in their data. You must therefore take steps to guarantee individuals can access their personal data, and in a manner which protects other individual's privacy.
7. Institute both technical and organisational measures to achieve data protection by design and by default (such as Privacy by Design methodologies), including through measures such as encryption, pseudonymisation, aggregation, anonymization and data minimization.
8. Data can be manipulated, damaged, lost or inappropriately exposed within any system. Design processes to check for on-going degradation in the quality of the data prior to its use by the system. This should include measures to prevent external corruption and to mitigate against silent and other forms of low-level data corruption;
9. Check for algorithmic bias during the detailed development phase. Data could be processed in a biased way, and therefore algorithms should be checked for this. (e.g. by using counterfactual evaluation methods)
10. Ensure that interface design honours principles of universal accessibility and avoid the introduction of functional biases in the detailed development phase that make the system unequally functional for different end-users.
11. Measurements to ensure traceability and accountability to the degree needed should be established within the following methods:
    - Methods used for designing and developing systems, such as the models built, the training methods, which data was gathered and selected, and how this occurred.
    - Methods used to test and validate systems, such as the scenarios or cases used to test and validate; the data used to test and validate; outcomes of the system (outcomes of,

[10] 'To determine whether a natural person is identifiable, account should be taken of all the means reasonably likely to be used either by the controller or by another person to identify the natural person directly or indirectly. To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments. The principles of data protection should therefore not apply to anonymous information' recital 26, GDPR.

or decisions taken by, the system); other possible decisions that would result from different cases, e.g., for other subgroups of users.
  - A series of technical methods to ensure traceability (such as encoding the metadata to extract and trace it when required). There should be a way of capturing where the data has come from, and the ability to construct how the different pieces of data relate to one another.

12. Make sure the code is actively explained and documented within the software program (as appropriate to the language(s) and methodology) and in appropriate ancillary documentation. Make sure documentation is understandable to fellow programmers and accessible by them. Confidentiality agreements can help overcome IP, confidentiality or trade secrets concerns.
13. Make sure you know to what degree the decisions and outcomes made by the system can be understood, including whether you have access to the internal workflow of the model.
14. Use formal methodologies and tools to ensure explainability wherever possible and if considered desirable for the particular system that is designed, such as the XAI or Transparency by Design approaches, and programmatic documentation, such as Model Cards.
15. Consider whether the system could present false or misleading information to people, and, if necessary, take measures to prevent that or add design requirements which will minimise this risk. In some cases, this risk will increase once the system is operational. If this is the case, add documentation, functionality, or other mechanisms to minimise this risk of misinformation.
16. Consider whether the system will unavoidably manipulate data, or make decisions which cannot be traced or understood by humans. If this is the case, add design requirements to expose data operations to scrutiny as much as possible, and/or prepare formal justification to explain why data operations cannot, and should not, be audited. Note that intellectual property concerns are not sufficient. Black box and “test track” testing regimes can be used to externally assess internal data operations.
17. Build tools and mechanisms into the development architecture to record all information relevant to ethics assessment, such as the source of datasets and the nature of models used. Ensure staff are trained and encouraged to use the relevant tools and mechanisms.
18. Whenever possible, create mechanisms to assess and, if necessary, act upon, concerns raised by staff and third parties. Ensure any such steps are taken before development (or deployment) continues.
19. Audit controls may need to be deeply embedded into the system. Ensure that audit controls are built to report performance and log the decisions made by the system.
20. Refine and complete the project’s ethical requirements document. This is likely to be an iterative process. As much as possible, record any decisions taken regarding how the system was made compliant with its ethical requirements.
21. Whenever possible follow sustainable energy usage practices. In particular, decisions made by the system that will affect the non-human world need to be carefully factored in.

6. Testing and Evaluation

As part of the testing and evaluation phase, an ethical assessment is performed to see if the system meets its ethical requirements.

It may be that the system achieves its functional requirements, but not the ethical requirements. If this is the case, the system cannot be considered to have been successfully completed. However, the whole point of Ethics by Design is to avoid such an outcome. If rigorously applied, Ethics by Design should prevent ethical issues at this stage of the development process.

As part of the testing and evaluation phase you should use the project's ethical requirements checklist to design a testing regime which can check the system's ethical compliance. It is highly unlikely any standard testing regime will consider all of the system's ethical requirements so the choice of testing methodology is important here. Implement this testing to determine whether the system meets all of its ethical requirements. Handle departures from the system's desired ethical characteristics just as seriously as a bug and undertake remedial work until the system meets its ethical requirements. It is highly recommended that stakeholder involvement takes place during this phase. Ask the stakeholders whether they are satisfied that the system adequately accounts for their values and needs (which are likely to have been discussed already at the beginning of the project) and make adjustments where needed.

The testing process should include testing the understanding of the system's behaviour by end-users. It cannot be assumed others will understand the system's output in the same way as developers. Test the understanding of affected persons regarding the purpose of the system, who or what may benefit from it, and (most importantly) what its limits are.

In addition to checking for compliance with the ethical requirements and engaging stakeholders, and to the degree matching the type of research being proposed (from basic to precompetitive) you should also consider the following steps:

1. Test whether users understand that they are interacting with a non-human agent and/or that a decision, content, advice or outcome is the result of an algorithmic decision in situations where not doing so would be deceptive, misleading, or harmful to the user.
2. Ensure audit controls are built into the system to check performance, record decisions made about the purpose and functioning of the system (including reporting on the impacts in general, not just occurrences of negative impacts). Ensure mechanisms are established to inform organisational users and end-users (if dealing directly with them) about the reasons behind the system's outcomes.
3. Ensure information about the system's capabilities and limitations is communicated to stakeholders, users and other affected persons in a clear, understandable and proactive manner, which enables realistic expectations.
4. Whenever possible, ensure practical processes exist for third parties (e.g. suppliers, consumers, distributors/vendors) or workers to report potential vulnerabilities, risks, or biases in the system. Ensure mechanisms exist to examine and act upon such reports.
5. Establish processes to obtain and consider users' feedback and mechanisms exist to adapt the system in response as appropriate.
6. Ensure users and stakeholders are given explanations they can understand as to why the system took a certain choice resulting in a certain outcome during testing so they can assess it accurately.
7. Develop and deliver training to users to help develop accountability practices (including the legal framework applicable to the system).
8. Formally attempt to predict the consequences/externalities of the system's operations.

3. PART III: ETHICAL DEPLOYMENT AND USE

This part of the guidelines apply to four practices central to the use of AI systems in research projects[11]: project management, acquisition, implementation and monitoring. By considering all these processes, these ethical guidelines aim also at preventing function creeps when the AI is deployed after the project end.

- *Project management* refers to the planning of a new research project, normally in a project plan, and the management of the planned activities during the project. For these activities, this part addresses the steps which should be taken in project planning and management to ensure proper consideration of ethical issues in the deployment and use of an AI system.
- *Acquisition* refers to process of acquiring an AI system if the system is not developed in the project itself. An organisation is responsible for the ethical state of any AI system it uses, even if that system has been built by another. If the system is developed in the project itself, then the Ethics by Design guidelines of Part 2 of this document apply.
- *Deployment and implementation* refer to process of deploying the AI system into a user environment, as well as planning and implementing changes in the organisation. The manner in which a system is deployed may change the ethical characteristics of the system. Implementation therefore must ensure the system continues to meet its ethical requirements.
- *Monitoring* is the process of monitoring conformance with requirements and the development and implementation of plans for improving performance. The full ethical characteristics of a system may not be apparent until the system is deployed "in the wild." As a result, all AI systems require on-going ethical monitoring and, where necessary, adjustment. This is typically done with an audit procedure, which is becoming a common legal requirement.

Project planning and management

- Plan for Ethics of Use-related tasks: In budgeting and planning, take into account the potential ethical issues that may occur. Whenever appropriate, consider the appointment of independent ethics advisor(s) with relevant expertise in ethics of new and emerging technologies and personal data protection.
- Define roles and procedures for implementation of the ethics guidelines, and for monitoring their implementation. This could include the institution of an AI ethics officer or team with responsibilities to implement ethics guidelines or monitor their implementation. It should not be assumed that whoever managed ethical compliance during development is the appropriate authority for this role.
- Ensure that the objectives for which the system will be used, the design requirements and resource choices conform to the ethical requirements provided for in the Ethics by Design objectives and requirements phases.

[11] Ethics guidelines for the deployment and use of AI systems in organisations have been developed as well. See Brey et al., 2019. Those guidelines focus on the organisational context in which AI systems are deployed and provide guidance for specific units and roles in the organisation.

Acquisition

- If an AI system is externally acquired as an off-the-shelf solution, choose the system that is most capable of meeting the ethical requirements specified in *Part 1*. If the AI system is custom-built by an external developer, give preference to a developer who uses an Ethics by Design approach or who is willing to adhere to the ethical requirements as listed in this guidance. To the degree possible, verify yourself that the system adheres to these requirements. At minimum, the vendor should be able to provide much of the required information. Since Ethics by Design calls for transparency and human oversight, it may be sufficient at first to ask the vendor to explain the developer's ethical oversight mechanisms and show samples of their transparency documentation. Without sufficient transparency, it will not be possible to determine the ethical compliance of the system nor to provide evidence for accountability reasons.
- If in-house development is chosen, follow the Ethics by Design method elaborated in Part II of this document and verify that the resulting system adheres to the ethical requirements listed.
- Ensure that any data collected and prepared for the system prior to deployment adheres to the data collection and preparation guidelines (Part II, Section 4).
- When appropriate, an ethical risk assessment and impact assessment should be performed to assess specific ethical risks in the use of the system. Mitigating actions should be carried out to mitigate any ethical risks detected. It may be possible to build this on top of the initial ethical assessment made when the project was first designed. However, it is important to recognise that new issues may have arisen as the system evolved during development and you learn more about it.

Deployment and implementation

- If the deployed AI tools contains personal data, delete them unless you can justify why the deletion is likely to render impossible or seriously impair the achievement of its specific purposes.
- Establish and implement plans and policies which support operational compliance with the ethical requirements for the system (see Part 1).
- Update data, access, security and risk management policies and procedures which apply to the system in order to account for the ethical requirements.
- In training for the operation and use of the system, include the new ethics policies. Pay attention to ethical aspects within communication about the launch of the system.
- Monitor the implementation of ethics guidelines for the system throughout the implementation phase, identify issues and risks and make adjustments where needed.

Monitoring

- Establish mechanisms to ensure in a verifiable way that end-users use the system according to user policies, are vigilant about ethical issues in operation and use, and consult with senior staff on issues that are morally problematic or ambiguous.
- Ensure that monitoring goals and metrics are in place for compliance with the ethics requirements. Periodically monitor compliance and propose improvements if monitoring shows compliance to be below target.
- When deploying AI system during the lifetime of the project, ensure that the AI stakeholders, users and subjects of the system have "ethical complaint" communication channels by which

to alert you to their ethical concerns as they arise. Keep in mind that ethical problems often occur because a system affects people who were never expected to be impacted by the system in the first place. Ethically compliant AI are designed to ensure appropriate technical and organizational safeguards to prevent function creep and intervene in case its risks could materialize.

Annex I Checklist: Specification of Objectives against Ethical Requirements

This checklist is a supporting tool and does not constitute an exhaustive list of all ethics requirement that may be applicable to the development of each specific AI system. It has to be used in conjunction with Part 1-3 of the current guidelines and applied to a degree matching the type of AI system and the research being proposed (from basic to precompetitive).

Specification of Objectives against Ethical Requirements	**Yes**	**No (how potential risks will be mitigated?)**
Respect for Human Agency		
End-users and others affected by the AI system are not deprived of abilities to make all decisions about their own lives, have basic freedoms taken away from them,		
End-users and others affected by the AI system are not subordinated, coerced, deceived, manipulated, objectified or dehumanized, nor is attachment or addiction to the system and its operations being stimulated.		
The system does not autonomously make decisions about vital issues that are normally decided by humans by means of free personal choices or collective deliberations or similarly significantly affects individuals,		
The system is designed in a way that give system operators and, as much as possible, end-users the ability to control, direct and intervene in basic operations of the system (when relevant)		
Privacy & Data Governance		
The system processes data in line with the requirements for lawfulness, fairness and transparency set in the national and EU data protection legal framework and the reasonable expectations of the data subjects.		
Technical and organisational measures are in place to safeguard the rights of data subjects (through measures such as anonymization, pseudonymisation, encryption, and aggregation).		
There are security measures in place to prevent data breaches and leakages (such as mechanisms for logging data access and data modification).		
Fairness		
The system is designed to avoid algorithmic bias, in input data, modelling and algorithm design. The system is designed to avoid historical and selection bias in data collection, representation and measurement bias in algorithmic training,		

aggregation and evaluation bias in modelling and automation bias in deployment		
The system is designed so that it can be used different types of end-users with different abilities (whenever possible/relevant)		
The system does not have negative social impacts on relevant groups, including impacts other than those resulting from algorithmic bias or lack of universal accessibility,		
Individual, and Social and Environmental Well-being		
The AI system takes the welfare of all stakeholders into account and do not unduly or unfairly reduce/undermine their well-being		
The AI system is mindful of principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects (when relevant)		
The AI system does not have the potential to negatively impact the quality of communication, social interaction, information, democratic processes, and social relations (when relevant)		
The system does not reduce safety and integrity in the workplace and complies with the relevant health and safety and employment regulations		
Transparency		
The end-users are aware that they are interacting with an AI system		
The purpose, capabilities, limitations, benefits and risks of the AI system and of the decisions conveyed are openly communicated to and understood by end-users and other stakeholders along with its possible consequences		
People can audit, query, dispute, seek to change or object to AI or robotics activities (when applicable)		
The AI system enables traceability during its entire lifecycle, from initial design to post-deployment evaluation and audit		
The system offers details about how decisions are taken and on which reasons these were based (when relevant and possible)		
The system keeps records of the decisions made (when relevant)		
Accountability & Oversight		
The system provides details of how potential ethically and socially undesirable effects will be detected, stopped, and prevented from reoccurring.		

The AI system allows for human oversight during the entire life-cycle of the project /regarding their decision cycles and operation (when relevant)		