Ethics by Design and Ethics of Use approaches for Horizon Europe artificial intelligence projects

[WP6, task 6.4]

Lead contributor	Philip Brey, *University of Twente* (p.a.e.brey@utwente.nl)
Other contributors	Brandt Dainow, *University of Twente*

Due date	Day, Month, Year
Delivery date	
Type	Report
Dissemination level	PU = Public

Keywords	Ethical issues; artificial intelligence; AI; robotics; robots; ethics; software design;

The SIENNA project - *Stakeholder-informed ethics for new technologies with high socio-economic and human rights impact* - has received funding under the European Union's H2020 research and innovation programme under grant agreement No 741716.

Abstract

This is a guidance document. It is intended at raising the awareness in the scientific community, and in particular with beneficiaries of EU research and innovation projects who wish to develop, deploy or use AI-based systems. This document offers guidance for the inclusion of an *Ethics by Design* approach. All AI-based systems (including robotics) proposed under the Horizon Europe framework will be subject to this review. It also offers ethics guidance for the deployment and use of AI-based systems in research. Adherence to the advice detailed in this document will ensure that your Horizon Europe proposal is 'AI'ethics ready' and will greatly facilitate your ethics compliance.

The Ethics by Design approach detailed here offers a way by which to include ethical principles and procedures into the design and development processes. Historically, ethical problems in AI have only been detected after the system has been deployed. Essentially, Ethics by Design seeks to make the ethical aspects of AI and robotics systems integral requirements of the system on the same level as reliability or security. The aim of Ethics by Design is to ensure ethical issues are addressed in the first place by using ethically-focused activities throughout the design, development and deployment phases of a project. These activities are detailed in this document, as are the ethical values these activities uphold, and to which all Horizon Europe AI and robotics projects must comply.

Document history

Version	Date	Description	Reason for change	Distribution
VX.X	dd Mmm YYYY	First Draft	VX.X	dd Mmm YYYY
VX.X	dd Mmm YYYY	Comments	VX.X	dd Mmm YYYY
VX.X	dd Mmm YYYY	Draft	VX.X	dd Mmm YYYY
VX.X	dd Mmm YYYY	Final Version		

Information in this report that may influence other SIENNA tasks

Linked task	Points of relevance
Number and title	What should be considered by linked tasks

Table of Contents

Abstract **2**

Table of Contents **3**

Executive summary **4**

List of figures **5**

List of tables **5**

List of acronyms/abbreviations **5**

1. Introduction **6**

2. Ethics by Design Principles **7**

2.1. The Ethics by Design approach 7

2.2. Values and Ethical Requisites of Ethics by Design 9

2.3. Conclusion 14

3. How to apply Ethics by Design in AI development **15**

3.1. Generic model for design 15

3.2. Design Phase: Specification of objectives 17

3.3. Design Phase: Specification of requirements 18

3.4. Design Phase: High-level Design 20

3.5. Design Phase: Data collection and preparation 22

3.6. Design Phase: Detailed design and development 24

3.7. Design Phase: Testing and evaluation 26

4. Ethical Deployment and Use **28**

5. Glossary of terms **30**

6. References and further reading **32**

Executive summary

This is a guidance document. It is intended for researchers who wish to develop or deploy AI-based systems under the Horizon Europe framework. This document offers guidance for the inclusion of an *Ethics by Design* approach for use in Horizon Europe applications. In addition, it offers ethics guidance for the deployment and use of AI systems in research. The adherence with the advice detailed in this document will greatly facilitate ethics compliance under the Horizon Europe Framework Programme.

The Ethics by Design approach offers a way to include ethical principles into the development processes. Historically, ethical problems in AI have generally been addressed only after deployment. It is now understood that it is possible to anticipate ethical issues in advance, rather than wait for harm to arise and then fix the problem, and to improve the design process by actively designing in support of moral values and principles. The Ethics by Design approach can be likened to the privacy by design approach, which has a longer history. Compared to this approach, it broadens the set of moral values included in design, including values like fairness, transparency and accountability.

The ethical values upon which Ethics by Design is based are grouped into six categories: Respect for Human Agency; Privacy and Data Governance; Fairness; Individual, Social and Environmental Well-being; Accountability and Oversight; Transparency. We then develop "ethical requisites," which are the conditions that an application must meet in order to achieve its goals ethically. Ethical requisites are instantiations of values within AI and robotics systems and development cycles. Asimov's Three Laws of Robotics are an example of ethical requisites. Ethical requisites may be met in many ways; through functionality, in data structures, in the process by which the system is constructed, and so forth. For example, one way the value of fairness can be met as an ethical requisite is to require that data sets and data processes do not exhibit racial bias. While many ethical requisites are aspects of the system itself, some are concerned with the way in which the system is developed. For example, the value of transparency requires that developers can explain how they tested for and removed bias from a dataset. It is not sufficient for developers to be satisfied there is no bias. Developers must be able to show what processes they used to remove bias and the analysis they undertook to determine why those processes, and not others, were used.

To support the incorporation of ethical requisites into the design process, the Ethics by Design approach contains a detailed proposal for steps to be taken at different phases in a design process. The approach presents a generic model of a design process, according to which the design of AI applications typically involves six tasks, which, depending on the precise design methodology chosen, may be carried out sequentially, in parallel or iteratively. They are the specification of objectives, specification of requirements, high-level design, data collection and preparation, detailed design and development, and testing and evaluation. For each of these tasks, detailed ethics guidelines are provided that recommend actions to be taken to support satisfaction of the ethical requisites, and thereby the ethical values. The text also contains steps for the inclusion of these detailed guidelines in one's preferred design methodology. Included as well in this text is an ethics checklist for easy verification of conformity to ethical requisites.

This document also contains a section with ethics guidelines for the deployment and use of AI systems in research projects. Based on the same set of values used in Ethics by Design, we present practical guidelines for inclusion of ethics considerations in project planning and management, acquisition, deployment and implementation, and monitoring.

List of figures

- **Figure 1:** The 5-layer model for Ethics by Design

List of tables

- **Table 1:** List of acronyms/abbreviations
- **Table 2:** Glossary of terms

List of acronyms/abbreviations

Abbreviation	Explanation
AI	Artificial Intelligence
GDPR	General Data Protection Regulation
IEC	International Electrotechnical Commission
IEEE	Institute of Electrical and Electronics Engineers
SHERPA	Shaping the ethical dimensions of smart information systems– a European perspective
XAI	Explainable Artificial Intelligence

Table 1: List of acronyms/abbreviations

1. Introduction

This document presents an approach to account for ethical issues in the development of AI applications (the *Ethics by Design* approach) and an analogous approach for incorporating such considerations in the deployment and use of AI applications (the *Ethics of Deployment and Use* approach). Ethics by Design aims for the systematic inclusion of ethical values, principles, requirements and procedures into design and development processes. The approach is intended to ensure that ethical issues are addressed as early as possible and followed up closely during the research activities. This requires specific ethically-focused activities at each stage of the design and development phases of a project. These activities are detailed in this document, as are the ethical values these activities uphold, which all Horizon Europe projects in the area of AI should comply with. Of course, for each particular ethical value or principle, its relevance and applicability will depend on the type of application, and hence needs to be carefully considered.

Using a generic model of the design process, we then offer detailed guidelines for meeting the requirements at each stage of the design process. We then explain how each of the six values can be embodied in AI and robotics systems as an "ethical requisite". For example, in order to be fair it is an ethical requisite that the system does not discriminate against particular racial groups. This document then explains how the application should detail the methods by which it will meet these ethical requisites at each stage of the design and construction process.

For researchers who deploy or use AI systems developed elsewhere, this document offers guidance for the inclusion of ethical features in the deployment and use of AI, robotics and big data systems. We present an Ethics of Deployment and Use approach for proper inclusion of such guidelines at different stages in the deployment and use of these systems.

2. Ethics by Design Principles[1]

The advice must be applied as appropriate to a degree matching the type of research being proposed (from basic to precompetitive).

2.1. The Ethics by Design approach

Ethics by Design is an example of the Value Sensitive Design approach (Friedman, Kahn, and Borning, 2002), which is the European Commission's recommended approach for EU-funded research in general, but with specific emphasis on AI. However, until recently, no detailed proposals for Ethics by Design approaches had been published. Our Ethics by Design approach is based on the findings of the EU-funded SHERPA project[2], which also takes an Ethics by Design approach. We moreover build on the ethics principles proposed in the High-Level Expert Groups on AI's *Ethics Guidelines For Trustworthy AI* (High-Level Expert Group on Artificial Intelligence, 2019), as well as from the SHERPA and SIENNA reports.

Many AI projects experience ethical issues only after they are deployed, while others reveal them during the development phase. Ethics by Design is intended to prevent ethical issues from arising in the first place, rather than trying to fix them after the damage has been done. Ethics by Design is intended to prevent ethical issues from occurring by proactively using moral principles as requirements of the system, termed "ethical requisites". Since many cannot be achieved unless the system is constructed in particular ways, ethical requisites sometimes apply to development processes and tools rather than the system being produced.

The Ethics by Design approach outlined in this document is now an indispensable part of the Horizon Europe evaluation process and is a mandatory requirement of all Horizon Europe applicants and beneficiaries. As part of the ethics review, adherence with all relevant ethical requisites will be assessed for all AI-related Horizon Europe applications.

5-Layer Model of Ethics by Design

Ethics by Design can be described as 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.

1. **Ethics by Design Values** – These are the principal ethical values that should guide AI or robotics systems. Where a system violates these values, it may be considered unethical. Values are to be upheld and enhanced. Privacy and fairness are examples of such values.
2. **Ethical Requisites** – Ethical requisites are the conditions that a solution or application must meet in order to achieve its goals ethically. In Ethics by Design, ethical requisites are instantiations of values within AI and robotics systems. Values may be instantiated in many ways; through functionality, in data structures, in the process by which the system is constructed, and so forth. For example, one way the value of fairness can be instantiated as

[1] *This section is intended for those who develop AI or robotics systems. It explains the principles by which ethical concerns can be factored into the design process. Chapter 4: Ethical Deployment and Use, p.40, is intended for those deploying AI or robotics systems (who may also be their developers).*

[2] Especially D3.2 Guidelines for the development and use of SIS (Brey et al., 2019) and D4.7 An Ethical Framework for the Development and Use of AI and Robotics Technologies (2020)

an ethical requisite is to require that a system does not exhibit racial bias. Asimov's Three Laws of Robotics are an example of ethical requisites.

3. **Ethics by Design Guidelines** Whereas ethical requisites are concerned with the system, guidelines are concerned with the steps by which it is created. Ethics by Design works on the basis that there are steps in the development process which are common to all design methodologies. The Ethics by Design approach offers a generic description of these phases in the development process and maps the ethical requisites onto these phases. This yields specific guidelines (usually formulated as tasks) at each phase which ensure that the final system instantiates the ethical requisites and therefore does not violate any ethical values. For example, the guidelines state that during the data gathering stage, data should be screened for fairness and any discriminatory biases that are found should be corrected.
4. **AI Methodologies** – There are a variety of methodologies used in AI and robotics projects. They are, at least partially, distinguished by the manner in which they organise the development process. Each methodology offers its own steps and sequence. Here Ethics by Design maps its principles onto the components of each individual methodology. If a project is using a different methodology, the researcher should return to the generic model (see *3: How to apply Ethics by Design in AI development , p.15*). By mapping the steps in the generic development process to their own methodology, they can allocate each guideline to the appropriate steps in their methodology.
5. **Tools & Methods** – The Tools and Methods layer accommodates specific programmatic artefacts and processes deployed within the development process to undertake Ethics by Design. It is possible some could be specific to a particular methodology and inapplicable to others, but at this stage, those which have emerged in the development community are tuned to ethical requisites and useable under any methodology. For example, Datasheets for Datasets (Gebru et al., 2020) are employed to assess the ethical characteristics of data, and so can be used at any stage which works with that data and for any norm relating to data. They can thus be deployed at multiple stages of the development process and are methodology-neutral.

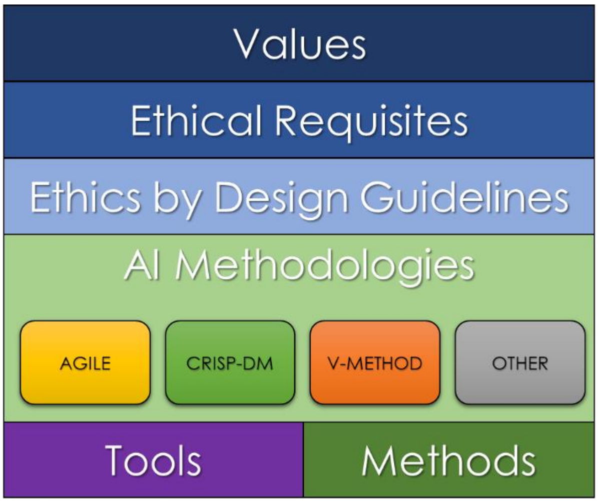

Figure 1:The 5-layer Model of Ethics by Design

2.2. Values and Ethical Requisites of Ethics by Design

Ethics by Design is based on ethical values such as fairness. These are then instantiated as concrete ethical requisites against which systems can be evaluated. This section will outline both the values and the ethical requisites that were derived from them. The requirements below are to be used as guides to what actions should be taken in the development process.

The requirements under the Ethics by Design approach can be grouped into six categories:

- Respect for Human Agency
- Privacy and Data Governance
- Fairness
- Individual, and Social and Environmental Well-being
- Transparency
- Accountability and Oversight

Under each category we will describe the values and provide examples of corresponding ethical requisites for AI and robotics systems.

Respect for Human Agency

Respect for Human agency encapsulates the values of autonomy, dignity and freedom. These are the fundamental rights upon which the EU is founded and are enshrined in the UN Declaration of Human Rights. Respecting autonomy means allowing people to think for themselves, decide for themselves what is right and wrong, and choose how they should live their life as a consequence. Human autonomy can take many forms because autonomy means each person deciding for themselves what their preferred form of autonomy is. Consequently what constitutes human autonomy is as varied as people are. As a result, 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. This is often done with the best of intentions because developers have not understood that other people may think differently. This is a particular problem with personalisation services which fail to respect cultural norms in other societies.

Dignity means every human being possesses an intrinsic worth which should never be compromised by others, including AI. This means they have the right not to be instrumentalized, objectified or dehumanized, and to be treated with respect. Respecting freedom, finally, means not constraining people in their actions that they should be able to pursue as autonomous persons, including freedom of movement, freedom of speech, freedom of information (the ability to access information), and freedom of assembly. It also includes freedom from constraints which conflict with one’s autonomy, such as coercion, deception and manipulation, within the limits of the law.

General ethical requisites

- 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.
- It should be ensured, as much as possible, that applications do not autonomously make decisions about vital issues that are normally decided by humans by means of free personal choices or collective deliberations, e.g., issues affecting life, health, well-being or individual rights, or economic, social and political decisions.
- It should be ensured, as much as possible, that end-users and others affected by the AI system are not deprived of abilities to make basic decisions about their own lives, have basic freedoms taken

away from them, are subordinated, coerced, deceived, manipulated, objectified or dehumanized, or that attachment or addiction to the system and its operations is being stimulated. This should not happen directly, through direct operations and actions of the system, and it also should be prevented, as much as possible, that systems can be used for these purposes.

Privacy & Data Governance

People have a right to privacy that must be respected. This includes both a respect for private and family life, home and communications and the protection of personal data. Both are subject to many legal protections, with personal data receiving extensive protection from the EU's General Data Protection Regulation (GDPR). Privacy rights are amongst the goods 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. It is important to note, that ethical issues can arise not only when processing personal data (where the data subject's rights and freedoms must be safeguarded) but also when the AI system uses non-personal data (e.g., racial bias).

General ethical requisites

- The processing of personal data requires careful consideration of the rights and freedoms of the data subjects. These should be safeguarded at all times. For more information and guidance please see the EU's *Guidance Note On Ethics And Data Protection*.[3]
- Whenever relevant, applications must explain how the proposed system supports the right of an individual to withdraw consent for the use of their personal data, and how they will be able to object to its use.
- In the case that personal data is processed by the developed AI systems, the proposal must demonstrate how it will ensure lawfulness, fairness and transparency of the data processing.
- Technical and organisational measures must be in place to safeguard the rights of data subjects through measures such as anonymization, pseudonymisation, encryption, and aggregation.
- Strong security measures to prevent data breaches and leakages must be set in place and described in the application (such as mechanisms for logging data access and data modification).
- 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 EU beneficiaries must apply GDPR principles.

Fairness

Fairness is used here in a philosophical sense, and is not to be confused with mathematical fairness or use of the term within computational modelling. Fairness means that people should be given equal rights and opportunities, and that people should not be advantaged or disadvantaged undeservedly. The first of these two principles, equal rights and opportunities, means that people are born equal and therefore should be in principle awarded the same fundamental rights, such as the referenced rights to autonomy, dignity, freedom, and privacy, and also the same fundamental opportunities, such as opportunities to get an education, apply and be considered for vacancies, and be adequately defended in court. It does not mean that the same outcomes must result, i.e., that people have equal wealth, income, status or success in life. In EU countries, these fundamental rights and opportunities are enshrined in law.

The second of these two principles, regarding undeserved (dis)advantages, particularly refers to the possibility of discrimination. Discrimination is a form of favouritism by which people are treated

[3] See https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection.pdf

unfairly on the basis of aspects of their identity which are inalienable and cannot be taken away from them. The most important of these are gender, race, age, sexual orientation, national origin, religion, health and disability. In EU countries, discrimination is not only unethical, but also against the law. Another principle related to fairness and nondiscrimination is diversity (or "respect for diversity"), which goes beyond nondiscrimination to include positive valuation of individual differences, recognition of differences in individual needs and support for the diverse composition of organisations and communities.

General ethical requisites

- *Avoidance of algorithmic bias:* AI systems should be designed to avoid bias in both input data, modelling and algorithm design. Algorithmic bias is a specific concern which needs specific mitigation techniques. Applications should specify the steps which will be taken to ensure data about people is representative and reflects their diversity. Similarly, applications should explicitly 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. This needs to consider inferences drawn by the system which have the potential to unfairly exclude or in other ways disadvantage certain groups of people.
- *Universal accessibility:* Whenever possible/relevant, AI systems should be designed so that they are usable by different types of end-users with different abilities. Applications are encouraged to explain how this will be achieved, such as by compliance with relevant accessibility guidelines. Moreover, AI systems should avoid functional bias in being designed to offer the same level of functionality and benefits to end-users with different abilities, beliefs, preferences and interests, to the extent possible.
- *Fair impacts:* Applications should demonstrate that possible negative social impacts on relevant groups, including impacts other than those resulting from algorithmic bias or lack of universal accessibility, have been considered and what steps will be taken to ensure the system does not cause them to be discriminated against or stigmatized, or otherwise have their interests affected in a negative way. This should be well documented in the research proposal.

Individual, Social and Environmental Well-being

It is desirable that AI systems contribute to, and do not harm, individual, social and environmental well-being. Individual well-being refers to the ability of individuals to lead happy, fulfilling lives in which they are able to pursue their own needs and desires. Social well-being refers to flourishing societies, whose basic institutions function well, including institutions of democratic government, the economy, healthcare, education, and others, and in which sources of social conflict are minimized. Environmental well-being refers to well-functioning ecosystems and absence of pollution and environmental degradation, enabled by sustainable processes of production, consumption, and energy use. At a minimum, AI systems should not contribute to harm, but preferably they should also make a positive contribution to these three forms of well-being.

General ethical requisites

- AI systems should take the welfare of all stakeholders into account and not reduce their well-being. It should be identified who the end-users and stakeholders will be of the application. It should then be assessed how the application could both enhance and harm their well-being, and documented choices should be made in development to support well-being and avoid harm to it.
- AI development should be mindful of principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects. There should be documented efforts

to consider the environmental impact of the system and, where needed, steps to mitigate negative impacts. In the case of robotics systems this must include the materials used and decommissioning procedures.

- AI systems with an application towards 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, for example by supporting uncivil discourse, amplifying fake news, segregating people into filter bubbles and echo chambers, creating asymmetric relations of power and dependence, and enabling political manipulation of the electorate. Mitigating actions should be taken to reduce the risk to such harms.
- AI and robotics systems should not reduce safety in the workplace. If relevant, your application should demonstrate consideration of possible impact on workplace safety, and compliance with IEEE P1228 (Standard for Software Safety).

Transparency

Transparency of AI systems refers to the idea that the purpose, inputs and operations of AI programs and algorithms should be knowable to its stakeholders so that they can understand how and for what purpose these systems function and how their decisions are arrived at. It is associated with other principles that address the understandability of AI systems, including explainability, traceability and interpretability. Transparency directly enables human agency, data governance, accountability, oversight and human governance. Transparency includes *all* elements relevant to an AI system: the data, the system and the processes by which it is designed, deployed and operated. Without transparency, a decision cannot be contested, or even understood. This would make it impossible to correct errors and unethical consequences.

The degree to which transparency is needed depends on the context and the severity of the consequences. It is important to note this is a judgement call, not a precise calculation, and others may not set boundaries or assess severities in the same manner as the researcher, so the precautionary principle dictates it is better to go too far than not far enough. This is why we recommend, if possible, that these decisions are made by a carefully constructed group, whose composition is sufficiently diverse so as to ensure a representative range of perspectives behind these decisions. Where the formation of a formal group is not possible, it is recommended researchers take steps to ensure they understand the full range of positions others may take.

General ethical requisites

- There is a general requirement for traceability across all areas of ethical AI and robotics. When building an AI solution one should consider what measures will enable the traceability of the AI system during its entire lifecycle, from initial design to post-deployment evaluation and audit.
- 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. Wherever it is necessary that people can audit, query, dispute or seek to change AI or robotics activities, your application must explain how this will be possible. It is not sufficient to merely consider the structure and functionality of the system in this respect. You must explain governance and other organisational processes by which your project will receive and assess requests from third parties.

- Whenever relevant, an application 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. However, with some systems this may not be possible. Nevertheless, the system (or those deploying it) should always have a mechanism by which to explain what the decision was and what data was used to make that decision. Explainability is especially a requirement for systems that make decisions and recommendations and perform actions for which accountability is required, such as decisions and actions that can cause significant harm, affect individual rights, or significantly affect individual or collective interests.
- The design and development processes will involve making decisions about ethical issues, such as how to remove bias from a dataset. The requirement for transparency means your development processes (and tools) will need components to keep records of such decisions so that it is possible to trace how these ethical obligations were met. This information may be required for audits, for disputing or resolving decisions made by the system, for correcting unexpected ethical issues which arise after system deployment and so that your own teams can learn and improve their handling of ethical issues.

Accountability and Oversight

Accountability for AI applications means that actors involved in their development, deployment or operation take responsibility for the way that these applications function and for the resulting consequences. Human oversight as a value requires that humans are able to understand, supervise and control the design, development, deployment and operation of AI and robotics systems. Oversight depends on accountability because one cannot hold someone or something (a system, a developer) to account unless one has an understanding of it and an ability to exert influence. Hence, to ensure accountability, developers must be able to explain how and why a system exhibits particular characteristics.

General ethical requisites

- AI systems should allow for human oversight regarding their decision cycles and operation, unless compelling reasons can be provided which demonstrate such oversight is not required. It should be explained how humans will be able to understand the decisions made by the system and what mechanisms will exist for humans to override them.
- The application should provide details of how ethically and socially undesirable effects of the system will be detected, stopped, and prevented from reoccurring.
- *To a degree matching the type of research being proposed (from basic to precompetitive) and as appropriate*, the application should include a formal ethical risk assessment for the proposed AI system. There should be documentation for the procedures for risk assessment and mitigation 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 they will be evaluated and actioned. The requirement for transparency means a mechanism should be included to communicate with these third parties has been done with their information.
- As a general principle, all AI systems should be auditable by independent third parties. The procedures and tools available under the XAI[4] approach support best practice in this regard. This is not limited to auditing the decisions of the system itself, but will need to discuss procedures and

[4] Explainable AI. See https://ieeexplore.ieee.org/abstract/document/8466590 for an overview.

tools used during the development process. Where relevant, the system should generate human accessible logs of the AI system's internal processes.

2.3. Conclusion

Ethics by Design is an approach for ensuring that an AI or robotics system complies with important ethical values. These values give rise to ethical requisites which a system must comply with. Some of these relate to the functionality while others relate to the processes by which systems are constructed. While many of these values are based on fundamental rights enshrined in EU charters and legislation, they are not specific to the EU alone, but reflect a growing global consensus. They are sometimes backed by legal requirements, but conformance cannot be achieved simply by adhering to legal obligations.

To a degree matching the type of research being proposed (from basic to precompetitive) and as appropriate, Horizon Europe applications must explain how the proposed project will ensure compliance with the relevant ethical demands. As with Privacy by Design, Ethics by Design calls for more than just specific features or functionality in the system. Supporting organisational processes are also required, as are specific features in development tools and methodologies (Cavoukian, 2009).

The following section uses a generic model of system development to detail the specific points any Horizon Europe AI or robotics application must cover in order to demonstrate how the project will meet the ethical demands of Horizon Europe projects.

3. How to apply Ethics by Design in AI development

Ethics by Design uses a generic model of the design process by which systems are produced. The ethical concerns are treated like reliability – as requirements any and all systems must achieve (*to a degree matching the type of research being proposed)*. Just like reliability, these ethical requirements place obligations on not just the system's features, but also on the development processes and tools themselves. This section explains how to embody these ethical factors in the design and development processes. We position ethical requisites as concrete tasks to be undertaken. By mapping one's own development methodology to the generic model used here, the relevant ethical requisites can be determined for each element of that development methodology. Once this has been accomplished, one has embedded Ethics by Design into one's development methodology as tasks, goals, constraints and the like. The chance of ethical concerns surfacing is minimised because each step in the development process will contain measures to prevent them arising in the first place.

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

- Because each individual has an inherent worth, AI systems should not negatively affect human autonomy, freedom or dignity.
- Because AI systems rely on data, it is important they do not violate the right to privacy and that the data used is representative and accurate.
- Systems should be developed with an inclusionary, fair, and non-discriminatory agenda.
- Because AI and robotics systems can have significant effects on individuals, society, and the environment, steps need to be taken to ensure they do not cause individual, social or environmental harm, rely on harmful technologies, or influence others to act in ways which cause harm.
- Systems should be as transparent as possible to their stakeholders
- Human oversight and accountability are required to ensure conformance to these principles and address non-compliance.

This chapter will describe the generic model, then outline the steps required to use it so as to incorporate Ethics by Design into one's development process.

3.1. Generic model for design

Ethics by Design is premised on the basis that development processes for AI and robotics systems can be described with a generic model containing six phases. While the six are presented here in a list format, *this is not necessarily a sequential process*.

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 by which to build 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.
4. ***Data collection and preparation.*** Data must be collected, verified, cleaned and integrated.
5. ***Detailed design and development.*** The actual construction of a full working system.

6. ***Testing and evaluation.*** Testing and evaluation of the system.

Specification of objectives

In the specification of objectives phase the system's objectives are evaluated against the ethical requisites presented in *Section 2.2: Values and Ethical Requisites of Ethics by Design*, p.9. Some objectives are not ethically permitted under any circumstance. For example, a system cannot be ethical if its objective is to subliminally manipulate people without their consent, or to assist in torture. If the aim is fundamentally incompatible with the ethical requisites, the project cannot proceed. **Not everything which can be done should be done.** It is possible that whether the system meets its ethical requisites or not depends on specific methods construction or the exact manner in which some functionality is implemented. If this is the case, proceed, but understand that some aspects of more detailed design will have ethical importance.

Specification of requirements

During this phase development requirements, resources and plans are assessed against the ethical requisites. One should determine how features of the system and construction processes will facilitate meeting the ethical requisites. For example, transparency may require that version control systems need additional components to record decisions taken regarding code changes. Ensure that the ethical requisites are included in the final list of product requirements. Not all ethical requisites may be relevant, so applications do not need to mention irrelevant ones.

High-level design

High-level design is concerned with the development of the requirements of the proposed system and the mechanisms by which this will be achieved. This often includes an initial ethical risk assessment. In many cases this will also include a hierarchical breakdown of the required sub-systems within the system, though some will consider this a part of detailed design. Ethical requisites should be treated just the same as any other requirements for the system. Design should include functionality by which to programmatically support ethical requisites, such as keeping logs of internal data manipulation by the system. The requirements for transparency and human oversight will typically require additional features beyond what is required to achieve the system's aim.

Data collection and data preparation

Data collection is an especially critical phase as far as ethics are concerned. Fairness and accuracy are the primary concerns here It should be assumed any data gathered is biased, skewed or incomplete until proven otherwise. 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.

Preparation of data itself may introduce issues. Steps should be taken to ensure testing, learning and algorithmic manipulation do not introduce new biases or other ethical issues (such as de-anonymisation). A frequent problem arises where testing does not accurately reflect the real-world use after deployment. For example, many facial recognition systems have poor performance with darker-skinned people due to testing on purely Caucasian populations. Consult the H2020 Guidance Note on Ethics and Data Protection at:
https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection.pdf.

Detailed design and development

In the Detailed Design and Development phase, actions which will incorporate the ethical requisites are added to the various tasks in the detailed design, as well as to the development infrastructure (tools, methodologies, procedures, and anything else which effects exactly *how* something is built).

Testing and evaluation

As part of the testing and evaluation phase, an ethical assessment is performed to see if the system meets its ethical requisites. It may be that the system achieves its functional requirements, but not the ethical requisites. 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, the Ethics by Design approach should prevent ethical issues at this stage of the development process. It is recommended that stakeholder involvement takes place during this phase.

We will now discuss each phase in detail and list the relevant tasks to be done.

3.2. Design Phase: Specification of objectives

While each project is unique, Ethics by Design lays down a set of standardised requirements which all AI and robotics should meet. An important first step is to ethically assess the objectives of a development projects against the ethical requisites. Sometimes, objectives are unethical or even illegal. For example, it cannot be an objective of a system to deceive people by collecting personal biometric data from them without their consent, using AI to hide this activity.

The two ethics guidelines for this design phase are the following:

- Assess whether the objectives for the design project will meet the relevant ethical requisites. It is recommended that a professional AI ethicist, if available, is enlisted to do the assessment of objectives, in collaboration with members of the development team.
- If your project has external stakeholders, such as researchers in other fields who will use the system, it is important your application shows how you plan to include them in the specification of objectives and specification of requirements phases. In particular, 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 they 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. In this way an appropriately diverse range of ideas and preferences will inform design choices.

Checking Design Objectives against Ethical Requisites

The objectives of the proposed system should be checked against the ethical requisites of section 2, in the manner suggested below. Potential violations differ in their degree of seriousness. Some violations may be only potential or less serious. Such concerns do not mean the objective should be abandoned, but that concrete steps will have to be taken to avoid unethical outcomes.

> When assessing objectives, 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.

To a degree matching the type of research being proposed (from basic to precompetitive) you should consider the following:

VALUE: Respect for Human Agency

- Check whether the objectives adhere to the human agency requirements. Serious ethical non-compliance is an issue for systems that limit human rights, subordinate, deceive or manipulate people, violate bodily or mental integrity, create attachment or addiction, or that hide the fact people are interacting with an AI system.

VALUE: Privacy & Data Governance

- Check whether the objectives are compatible with the privacy and data governance requirements. Non-adherence to any of these would result in serious non-compliance.
- Assess whether the plans for what data will be used are fair and appropriate. If the proposed data source is unfair or inappropriate, either change the data source or modify the objective so that that data source is not needed.

VALUE: Fairness

- Check whether the objectives are compatible with the requirement for fairness. Consider whether violations would 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, this would constitute serious non-compliance.

VALUE: Individual, and Social and Environmental Well-being

- Check whether the objectives are compatible with the well-being requirements. Particularly serious non-aherence applies to 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). 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 done for projects that are of sufficient scale.

VALUE: Transparency

- The ethical requisites for transparency do not usually apply to objectives, but have to be considered at this stage to determine the degree to which the system's objectives will allow for the required transparency to be built into the system.

VALUE: Accountability & Oversight

- Most of the ethical requisites for accountability and oversight do not apply to objectives, but rather to the architecture and detailed design of the system. However, all systems should have an objective of allowing for human oversight and intervention on--the system's decisions. If it does not, change the objectives or provide compelling reasons why such oversight is not required.

3.3. Design Phase: Specification of requirements

The primary function of the Requirement Specification phase is to arrive at a development plan that includes design specifications for the system, design the development infrastructure, determine staff resources required, set 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.

Changing these can be more challenging than building systems. Nevertheless, it cannot be assumed that any tool, process or organisational elements will be support Ethics by Design. Some of the ethical requisites 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 recognised it is possible development methods, tools or even organisational structures used on previous projects will need modification. As a result, applicants must recognise there is likely to be a need to adapt (or even replace) aspects of their customary development systems so that they become capable of delivering their project's ethical requisites.

In some cases, it may not be technically possible to meet every ethical requisite due to lack of suitable development tools. If this is the case, your application should explain this in order to justify non-compliance. However, you will need to be extremely rigorous in your investigations for suitable tools. The requirements here are not unique to Horizon Europe, but are common demands of many AI projects. Consequently, tools to meet these needs are developing rapidly. For example, Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2020) have been produced specifically to provide ethical documentation of AI development, while Explainable AI (XAI) is a rapidly developing set of methodologies and tools by which to build AI systems which allow for human governance.

The degree to which a technical inability to meet the ethical requisites blocks your project also depends on the particular ethical requisite 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 traffic lights has only a very limited impact on the life of individuals, so the need for transparency is lower. Where you believe it is technically impossible to meet a relevant ethical requisite, the importance of the requisite will be a factor in determining approval.

Ethics guidelines

To a degree matching the type of research being proposed (from basic to precompetitive):

- Similar to the ethical assessment of objectives, an ethical assessment should be done of the proposed design specifications, constraints, selected resources and infrastructure for compatibility with the ethical requisites. For example, some deep learning techniques may not to be the best choice for transparency. Check these specifications and constraints against the ethical requisites of section 2 and modify them if needed to ensure a good fit.
- Ensure that relevant ethical requisites are covered in the list of design specifications. For this purpose, consider inclusion of an Ethical Requisites document. At the Objectives phase this document will only cover ethical aspects of the overall system and the most obvious features of the development process. However, it can be refined and added to as the project proceeds.
- 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 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 later points in the development process as more information comes in. A professional AI ethicist, if available, should be able to perform such an assessment. Ethical risk assessment should be planned and budgeted

for. This assessment needs to be scaled to the nature of the project, the severity of ethical risks and the overall budget of the development project. A standard for ethical impact assessment is available at https://satoriproject.eu/media/CWA17145-23d2017.pdf.

- It is recommended that at this stage, that an *EbD implementation plan* is completed which specifies future steps to be taken to incorporate EbD in the development process and actors responsible for carrying them out and for monitoring them. This implementation plan should incorporate the ethical risk and impact assessment if one has been completed. We 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. Second, we recommend inclusion of an ethical governance model which includes organisational structures for governance of the EbD process, including, most likely, ethical review committees. The governance model needs to address the following issues: How will governance be exercised? What is the project's version of a supervisory authority to ensure the ethical requisites are met? What powers will it have? How will it be selected fairly and inclusively? What procedures will be used in the case of a conflict between the ethical governance authority and developers or engineers or clients? The governance mechanisms should include the steps which have to be taken to incorporate EbD into the development process, the actors responsible for carrying out EbD-related tasks and those who monitoring this.

3.4. Design Phase: High-level Design

In high-level design, the architecture for a system or software product is specified. To a degree matching the type of research being proposed (from basic to precompetitive) and whenever applicable, you should consider the following:

VALUE: Respect for Human Agency

- Verify that the design allows for an interface based on human-centric design principles which leave meaningful opportunities for human choice.

VALUE: Privacy & Data Governance

- Verify that the design supports the ethical requisites for privacy and data governance. Ensure development processes, procedures and tools do not expose personal data such that it violates the right to privacy. 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.
- Ensure there are formal processes 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 a mechanism to record how data selection was undertaken for external audits.
- It cannot be *assumed* that the data obtained is the data which you wanted. For example, datasets may be incomplete or methods of importing data may alter it in unexpected ways. Design formal processes to check for and correct bias or errors after importing any data.

VALUE: Fairness

- Undertake an accessibility assessment of the interface and other touchpoints. Ensure that, where relevant, the system meets accessibility standards.

- Does the high-level design suggest that some users of the system will obtain better functionality than others? If so, prepare a formal justification for this or modify the design.
- Examine the initial interface design and other touchpoints to see whether it is assuming a one-size-fits-all approach to users. If so, see if this makes using the system more difficult for some people. If this is the case, either modify the design or prepare a formal justification for it.

VALUE: Individual, and Social and Environmental Well-being

- 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 taken to more depth. Demonstrate how the system will be constructed in an environmentally friendly way.
- Evaluate whether the system could cause physical harm to people, animals or property. This is especially important with robotic systems. If this is possible, include design features to minimise this risk and/or the amount of harm which can be done. If the system will be able to respond to voice commands, you must include "emergency stop" vocal commands in the design.

VALUE: Transparency

- Design mechanisms to document how data acquisition, storage and use happen. This needs to be auditable. This must cover both the development process and use once operational.
- Design procedures and select and configure tools which can document development processes to a level that humans can understand and evaluate decisions made within the design and development processes. This will be required by anyone who is concerned the system is unethical in some way and wants to see if this was caused within the development process. This can include users, those responsible for ensuring the created system meets its ethical requisites, and external ethical auditors. We recommend a layered approach to this documentation, 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 documentation appropriate to their level of expertise and their specific concerns.
- Ensure the design includes mechanisms by which the AI system will record its own decisions so that they can be subject to human review. 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 for external audit.
- 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.
- Ensure ethical documentation systems are sufficient to make ethical issues identifiable and their resolution traceable and explainable.
- 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.
- Ensure there is no aspect of the AI system which could be mistaken for a human once the system is deployed. Bear in mind many people may not have an understanding of AI and can innocently 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 (Candello, Pinhanez, and Figueiredo, 2017; Castelo, Schmitt, and Sarvary, 2019).
- Ensure processes exist, and are actively maintained, by which both internal staff and third parties can report potential vulnerabilities, risks, or biases in the system during the development process.

VALUE: Accountability & Oversight

- To the extent possible and appropriate, design mechanisms for human oversight and external audit once the system is deployed. This may require additional functionality inside the system solely for reporting internal activity and which has no role in the system's functionality. Oversight after deployment will need access to the oversight work performed during development.
- Design a testing regime which can check that the system's internal operations meet the ethical requisites. This may require changes to the way functionality is achieved within the system so as to permit appropriate testing and remedial action.
- Ensure that the system is designed in a manner which permits external ethical auditing. If unsure, refer to existing ethical audit procedures and codes of conduct for ethical AI.

3.5. Design Phase: Data collection and preparation

To integrate ethical requisites into the data collection and preparation processes, assess how operations within each process might violate ethical or data protection requirements. Make necessary changes as a result. If appropriate changes are not possible, the design objectives may need to be altered. Bias, discrimination, diversity, privacy and data quality will be particularly important.

The processing of personal data is governed by GDPR. Personal data is any information that relates to an identifiable living individual. Personal data which has been de-identified or pseudonymised but can be used to re-identify a person is also personal data. Personal data that has been rendered anonymous to the degree that the individual is no longer identifiable is not personal data. However, the anonymisation must be absolutely irreversible. Special categories of data (also often called sensitive data) are a subset of personal that is particularly sensitive and must be treated with special attention. Such data includes data concerning racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic and health-related data, biometric data for the purposes of uniquely identifying a natural person, data concerning a natural person's sex life or sexual orientation and. Special rules may apply to the processing of data related to criminal convictions and offences.

VALUE: Privacy & Data Governance

- This document is not a definitive guide of your obligations regarding data processing. For more detailed information and guidance please see: *Guidance Note On Ethics And Data Protection*[5] and the relevant section *HE Guidance How To Complete Your Ethics Self-Assessment*.[6]
- 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.
- 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;
- Applications must demonstrate compliance with the principles of Privacy by Design and Privacy by Default. You should explain what internal policies and active processes the system and your organisation will implement to safeguard the rights and freedoms of the data subjects.
- If your planned system will process personal data, your application must demonstrate how you have incorporated the rights to data protection into your design. This includes how you will enable individuals to withdraw consent for the use of their personal data and what mechanisms you will

[5] https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection_0.pdf

[6] https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf

create which enable them to object to its use. You must demonstrate how your design will ensure that data controllers and processors are able to fulfil their data protection obligations.

VALUE: Fairness

- 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.
- Analyse your training data and ensure that your data is representative and value-aligned.
- 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. Assess the diversity and representativeness of users in the data, testing for specific populations or problematic use cases.
- Make sure data from one demographics 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 groups. 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.

VALUE: Transparency

- Prepare a data protection policy document which details how the project complies with data protection requirements. This will be needed for those concerned with ensuring compliance with the ethical requisites, data protection officers and regulators, and for ethical audits. This is a mandatory requirement under data protection regulations.
- You must carry out an analysis of the ethics risks related to the data processing and whenever needed, produce a risk mitigation plan.
- Ensure that you can explain how personal data is used, shared, and stored.

VALUE: Accountability & Oversight

- Many organisations processing personal data are required to have a data protection officer or similar, so if your organisation is one of these, it is highly recommended that they are consulted on the appropriate requirements for the project.
- Build a culture of shared responsibility for the organization's data assets and that the potential value of data assets is acknowledged. Ensure that employees understand the true cost of failing to implement a data quality culture.
- Make sure that roles and responsibilities are clear for governance and management of data assets and that all employees and stakeholders understand them.
- If using external organisations for data storage/analysis/collection ensure these are also compliant with data protection requirements. GDPR requires that you verify their practices are compliant yourself.
- Make sure you have clearly established what kind of sample you need, what kind of sample you have taken, and that you can articulate what it will be used for.

3.6. Design Phase: Detailed design and development

To a large degree this phase involves adding more detail to the ethical requisites 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 requisites, ensure that ethical guidelines are communicated to all developers and engineers, and that the design is evaluated relative to these ethical guidelines by them wherever they need to make relevant decisions. Issues that may be particularly relevant in this design are those relating to transparency, privacy and accountability.

To a degree matching the type of research being proposed (from basic to precompetitive) you should consider the following:

VALUE: Privacy & Data Governance

- If creating new personal data (e.g., through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets), make sure all newly created personal data is given at least the same protection and attracts the same rights as previously collected or held personal data.
- Ensure no new personal information is, or can be, collected or created during development of the system, unless necessary. If new personal data is collected or created, then have systems in place to impose access or use limitations which will protect individuals' privacy..
- Ensure there are processes to safeguard the quality and integrity of all pertinent data, including means of verifying that data sets have not been compromised or hacked. If you are in control of the quality of the external data sources used, assess to what degree you can validate their quality.
- Make sure that roles and responsibilities are clear for governance and management of data assets and that all relevant staff understand them.
- Ensure there are oversight mechanisms for data processing (including limiting access to only appropriate personnel, mechanisms for logging data access and making modifications).
- Be aware that once data is anonymized, it may be possible to de-anonymise it.
- Ensure there is an embedded process that allows individuals to access their data and remove it from the system and/or correct errors in the data where these occur. AI systems must support the right for someone to withdraw consent for the use of personal data or object to its use. If required by law, it should also support the right to be forgotten. You must therefore take steps to guarantee a person can access their personal data, and in a manner which protects other individual's privacy.
- Make sure no new personal data is, or can be, collected or created during regular use of the system, unless necessary (e.g., for the function of the system or realization of the business or research objectives).
- Institute both technical and organisational measures to achieve data protection by default (such as Privacy by Design methodologies), including through measures such as encryption, pseudonymisation, aggregation, anonymisation and data minimalization (especially for personal data).
- AI systems used for commercial purposes must respect data portability, meaning that a person can download their personal data and move it to a competitor. You must therefore ensure any individual's personal data can be exported from the system and that the loss of this record will not damage the system's functionality.
- Data can be manipulated, damaged, lost or inappropriately exposed within any system. Design processes to check for on-going degradation in the ethical 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.

VALUE: Fairness

- Check for algorithmic bias, particularly computational bias, during the detailed development phase. Data could be processed in a biased way, and therefore algorithms should be checked for this.
- 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.

VALUE: Individual, and Social and Environmental Well-being

- 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.

VALUE: Transparency

- Measurements to ensure traceability 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, 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.
- 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.
- 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.
- 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 (Doran, Schulz, and Besold, 2017) or Transparency by Design (Rossi and Lenzini, 2020) approaches and programmatic documentation, such as Model Cards (Mitchell et al., 2019).
- Consider if the system could present false or misleading information to people. Make sure that the system will not present false or misleading information to people. If so, add design requirements which will minimise this risk. In some cases, the risk is more likely once the system is operational. If this is the case, add documentation, functionality, or other steps to be used once the system is deployed to minimise misinformation.
- Consider if it is unavoidable that the system will manipulate data, or make decisions based on data, which cannot be traced or understood by humans. If so, add design requirements to expose data operations to scrutiny as much as possible. Alternatively, 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 (Aggarwal et al., 2019).

- Build tools and mechanisms into the development architecture to trap important 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 them.

VALUE: Accountability & Oversight

- Whenever possible, create mechanisms by which concerns raised by staff and third parties can be assessed and, if necessary, acted upon. Ensure any such steps are taken before development continues.
- 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.
- Refine and complete the project's ethical requisites 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 requisites.

3.7. Design Phase: Testing and evaluation

As part of the testing and evaluation phase you should use the project's ethical requisites document 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 requisites so the choice of testing methodology is important here. Implement this testing to determine whether the system meets all of its ethical requisites. Treat departures from the system's desired ethical characteristics just as seriously as a bug and undertake remedial work until the system meets its ethical requisites. 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.

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

VALUE: Transparency

- 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.
- 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.
- Ensure information to stakeholders, users and other affected persons about the system's capabilities and limitations is communicated in a clear, understandable and proactive manner, and which enables realistic expectations.

VALUE: Accountability & Oversight

- 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 action such reports.

- 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.
- Establish processes to obtain and consider users’ feedback and mechanisms exist to adapt the system in response as appropriate.
- 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.
- Develop and deliver training to users to help develop accountability practices (including teaching about the legal framework applicable to the system).
- Formally attempt to predict the consequences/externalities of the system’s operations.

4. Ethical Deployment and Use

In this section, we will present ethics guidelines for the deployment and use of AI systems. The ethics guidelines presented here are for the use of AI systems in research projects[7].

Our guidelines apply to four practices central to the use of AI systems in research projects: project management, acquisition, implementation and monitoring.

- *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. We address 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. An organisation is responsible for the ethical state of any AI system it uses, even if that system has been built by another.
- *Deployment and implementation* refers 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 requisites.
- *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 an common legal requirement.

We assume that all four of these processes take place when an AI system is deployed in a research project, and outline ethics guidelines for each of them.

Project planning and management

- Plan for Ethics of Use-related tasks. In budgeting and planning, take into account the potential ethical issues that were in the ethics self-assessment. Whenever appropriate, consider the appointment of independent ethics advisor with relevant expertise in ethics of new and emerging technologies and data protection.
- In the project plan 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 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 requisites provided for in the Ethics by Design objectives and requirements phases.
-

Acquisition

- If an AI system is externally required as an off-the-shelf solution, pick the system that is most capable of meeting the ethical requisites specified in *Section 2.2: Values and Ethical* Requisites *of Ethics by Design*, p.9. If the AI system is custom-built by an external developer, give preference to

[7] We have also developed ethics guidelines for their deployment and use in organisations (Brey et al., 2019). Those guidelines emphasize the organisational context in which AI systems are deployed and provide guidance for specific units and roles in the organisation.

a developer who uses an Ethics by Design approach or who is willing to adhere to the ethical requisites 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 them 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.

- If in-house development is chosen, follow the Ethics by Design methods in this document and verify that the resulting system adheres to the ethical requisites listed here.
- Ensure that any data collected and prepared for the system prior to deployment adheres to the data collection and preparation guidelines provided in *Section 3.5: Design Phase: Data collection and preparation* p. 22.
- 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, which should have examined these issues and so provide an existing framework for analysis. However, it is important to recognise that new issues may have arisen as the system evolved during development and you learn more about it. While the initial ethical assessment serves as a foundation, one should not be limited to only what it has identified.

Deployment and implementation

- Establish and implement plans and policies which support operational compliance with the ethical requisites for the system.
- Update data, access, security and risk management policies and procedures which apply to the system in order to account for the ethical requisites.
- 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.

5. Glossary of terms

Term	Explanation
Accountability	Accountability applies to both individuals and institutions. It means being able to explain the reasons behind your actions and a willingness be held responsible for them.
AI	Artificial Intelligence.
Algorithmic bias	Bias in computer systems which results in unfair consideration and treatment of individuals or social groups represented by the system or otherwise affected by the its outputs. Algorithmic bias can result from bias in inputs (biased or unbalanced data, especially data representing persons or groups), computational bias that results from choices in modelling and algorithm design, and output bias, which may arise from the feedback loops generated between an AI system and the environment affected by its decisions.
Auditability	Auditability refers to the ability of an AI system to undergo the assessment of the system's algorithms, data and design processes.
Autonomy	Ethical AI is concerned with human autonomy, of which there are three types. Moral autonomy is determining what is morally good and bad. Political autonomy refers forming one's own political opinions. Personal autonomy refers to deciding how one should live, especially by what values one should make decisions.
Bias	Bias is an unfair or unjustified prejudice towards or against a person, group of people, object, or position. See also *algorithmic bias*.
Discrimination	The act of making unjustified distinctions between human beings based on the groups, classes, or other categories to which they are perceived to belong, especially gender, race, age, sexual orientation, national origin, religion, income, property, health, or disability.
Diversity	Diversity is the organisation of people based on identity markers like gender, race, age, cultural heritage, ability, and education.
Ethics	Ethics are moral principles that govern a person's behaviour. It is also a branch of philosophy dealing with these principles. Applied ethics deals with the use of moral principles in real-life situations. AI Ethics is an example of applied ethics focused on the issues raised by AI.
Ethics assessment	The assessment, evaluation, review, appraisal or valuation of plans, practices, products and uses of research and innovation that makes use of ethical principles or criteria.
Ethical AI	Ethical AI refers to the development, deployment and use of AI that ensures compliance with ethical norms, including fundamental rights as special moral entitlements, ethical principles and related core values.
Ethical impact assessment	An approach for judging the ethical impacts of research and innovation activities, outcomes and technologies that incorporates both the means for a contextual identification and evaluation of these ethical impacts and the development of a set of guidelines or recommendations for remedial actions aimed at mitigating ethical risks and enhancing ethical benefits, typically in consultation with stakeholders.

Ethical requisite	A key term in this document. An ethical requisite is a requirement relating to ethical aspects of the system and the development thereof. Ethical requisites must be met in order to be compliant with the demands for responsible, trustworthy, ethical AI.
Explainability	Explainability is the extent to which the internal mechanics of a machine or deep learning system can be explained in human terms.
Informed consent	Permission freely given and granted in full knowledge of the possible consequences.
Oversight	The ability to oversee, supervise, and watch carefully over something – in this context, to oversee the functionality and output of AI systems.
Personal data	Information relating to an identified or identifiable natural person, directly or indirectly, by reference to one or more elements specific to that person. GDPR specifically mentions racial or ethnic origin, political opinions, religious beliefs, trade union membership, genetic data, biometric data, health, and sexual orientation.
Personal data processing	Any operation or set of programmatic operations to personal data.
Privacy by design	Privacy by Design is an approach taken when creating new technologies and systems. Privacy by Design encompasses IT systems, business practices and physical design. The approach is characterized by proactive anticipation of privacy invasive events so as to prevent them from occurring, rather than fixing them afterwards.
Profiling	According to Article 4(4) of the GDPR, 'profiling' means automated processing of personal data to evaluate personal aspects relating to a person, such as personal preferences, interests, or movements.
Pseudonymisation	According to Article 4 of GDPR, ‘pseudonymisation’ means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person
Reproducibility	Reproducibility describes whether an AI experiment exhibits the same behaviour when repeated under the same conditions.
Stakeholders	All those that research develop, design, deploy or use AI, as well as those that are (directly or indirectly) affected by AI – including but not limited to companies, organisations, researchers, public services, institutions, civil society organisations, governments, regulators, social partners, individuals, citizens, workers and consumers.
Traceability	Traceability of an AI system refers to the capability to keep track of the system’s data, development and deployment processes, typically by means of documented recorded identification.
XAI	Explainable AI. XAI refers to initiatives, including procedures and coding tools, in response to AI transparency and trust concerns. XAI aims to produce explainable models while also maintaining a high level of learning performance; and enable humans users to understand, trust and manage AI systems.

Table 2: Glossary of terms

6. References and further reading

Aggarwal, A., P. Lohia, S. Nagar, K. Dey, and D. Saha, 'Black Box Fairness Testing of Machine Learning Models', *Proceedings of the 2019 27th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering - ESEC/FSE 2019*, ACM Press, Tallinn, Estonia, 2019, pp. 625–635.

Brey, P., B. Lundgren, K. Macnish, and M. Ryan, *D3.2 Guidelines for the Development and Use of SIS*, *Shaping the Ethical Dimensions of Smart Information Systems– a European Perspective (SHERPA)*, SHERPA Project, 2019.

Candello, H., C. Pinhanez, and F. Figueiredo, 'Typefaces and the Perception of Humanness in Natural Language Chatbots', *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, ACM, Denver Colorado USA, 2017, pp. 3476–3487.

Castelo, N., B. Schmitt, and M. Sarvary, 'Robot Or Human? How Bodies and Minds Shape Consumer Reactions to Human-Like Robots', *ACR North American Advances*, 2019.

Cavoukian, A., *Privacy by Design: The 7 Foundational Principles*, Information and Privacy Commissioner of Ontario, Toronto, 2009.

D4.7: An Ethical Framework for the Development and Use of AI and Robotics Technologies, *SIENNA Project. WP4: Artificial Intelligence and Robotics - Ethical, Legal and Social Analysis*, European Union, Brussels, 2020.

Doran, D., S. Schulz, and T.R. Besold, 'What Does Explainable AI Really Mean? A New Conceptualization of Perspectives', *ArXiv Preprint ArXiv:1710.00794*, 2017.

Friedman, B., P. Kahn, and A. Borning, 'Value Sensitive Design: Theory and Methods', *University of Washington Technical Report*, No. 2–12, 2002.

Gebru, T., J. Morgenstern, B. Vecchione, J.W. Vaughan, H. Wallach, H. Daumé III, and K. Crawford, 'Datasheets for Datasets', *ArXiv:1803.09010 [Cs]*, March 19, 2020.

Gunning, D., 'Explainable Artificial Intelligence (Xai)', *Defense Advanced Research Projects Agency (DARPA), Nd Web*, Vol. 2, No. 2, 2017.

High-Level Expert Group on Artificial Intelligence, *Ethics Guidelines For Trustworthy AI*, European Commission, Brussels, 2019.

Mitchell, M., S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I.D. Raji, and T. Gebru, 'Model Cards for Model Reporting', *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 2019, pp. 220–229.

Rossi, A., and G. Lenzini, 'Transparency by Design in Data-Informed Research: A Collection of Information Design Patterns', *Computer Law & Security Review*, Vol. 37, 2020, p. 105402.

Xie, X., J.W.K. Ho, C. Murphy, G. Kaiser, B. Xu, and T.Y. Chen, 'Testing and Validating Machine Learning Classifiers by Metamorphic Testing', *Journal of Systems and Software*, Vol. 84, No. 4, April 2011, pp. 544–558.

Annex - Checklist for compliance to Ethics by Design requisites of a developed system or application

General requirements

- The application has been developed using the Ethics by Design approach, based on an EbD implementation plan that has been formulated early on in the project. This implementation plan incorporates an ethical risk and impact assessment that assesses special ethical issues and risks associated with the application that is to be developed.
- Stakeholders are consulted during the early stages of the development cycle (specification of objectives and requirements) as well as during later stages (testing and evaluation), and their preferences and values have been taken into account.

Respect for Human Agency

- The application gives system operators and, as much as possible, end-users the ability to control, direct and intervene in basic operations of the system.
- Applications do not autonomously make decisions about vital issues that are normally decided by humans by means of free personal choices or collective deliberations, e.g., issues affecting life, health, well-being or individual rights, or economic, social and political decisions.
- It is unlikely that operations and actions directly caused by the application deprive end-users and others affected by the AI system abilities to make basic decisions about their own lives, have basic freedoms taken away from them, subordinates, coerces, deceives, manipulates, objectifies or dehumanizes them, or stimulates attachment or addiction to the application and its operations, and uses of the application for these purposes are precluded as much as possible.

Privacy & Data Governance

- The rights and freedoms of data subjects are protected in data processing by the application, along the requirements set by the EU's *Guidance Note On Ethics And Data Protection*[8] and the GDPR.
- The applications explains, where relevant, how it supports the right of an individual to withdraw consent for the use of their personal data, and how they will be able to object to its use.
- Technical and organisational measures are in place to safeguard the rights of data subjects through measures such as anonymization, pseudonymisation, encryption, and aggregation.
- Strong security measures to prevent data breaches and leakages are set in place and described in the application (such as mechanisms for logging data access and data modification).
- Data is acquired, stored and used in a manner which can be audited by humans.

Fairness

- The application is designed to avoid algorithmic bias, including input bias, computational bias and output bias.

[8] See https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection.pdf

- To the extent possible, the system is designed according to principles of universal accessibility, so that it is usable by different types of end-users with different abilities, and functional bias is avoided, to the extent possible, so that same level of functionality and benefits is offered to end-users with different abilities, beliefs, preferences and interests.
- Possible negative, unfair and discriminatory impacts on stakeholder groups resulting from use of the application, including impacts other than those resulting from algorithmic bias or lack of universal accessibility, have been analyzed and mitigated where possible.

Individual, Social and Environmental Well-being

- The application has been designed to minimize any potential harms to end-users and other stakeholders, and to enhance their welfare to the extent possible.
- The application embodies principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects.
- AI applications with an application towards media, communications, politics, social analytics, and online communities have been designed so that potential negatively impacts on the quality of communication, social interaction, information, democratic processes, and social relations have been mitigated.
- AI and robotics applications intended for use in the workplace mitigate for potential negative impact on workplace safety, and and are compliant with IEEE P1228 (Standard for Software Safety).

Transparency

- Measures are in place to enable the traceability of the AI application during its entire lifecycle, from initial design to post-deployment evaluation and audit.
- The application is designed so that it is manifest to end-users that they are interacting with an AI system.
- The purpose, capabilities, limitations, benefits and risks of the AI system and of the decisions conveyed by it have been documented, and this documentation is made easily available to end-users and other stakeholders, including instructions on how to use the system properly. Wherever it is necessary that people can audit, query, dispute or seek to change AI or robotics activities, it is explained how that is possible.
- Whenever relevant, decisions made by the system are explainable to users, including the reasons for making the decisions. Explainability is especially a requirement for applications that are involved in decisions and actions for which accountability is required, such as decisions and actions that can cause significant harm, affect individual rights, or significantly affect individual or collective interests.
- Records have been kept of decisions during the design and development process about ethical issues, for later audits, dispute resolution, improvements to the application if new ethical issues emerge, and collective learning.

Accountability and Oversight

- The application allows for human oversight regarding its decision cycles and operation, including sufficient transparency of operations and abilities by end-users and operators to influence or override operations, unless compelling reasons can be provided which demonstrate such oversight is not required.

- It is documented how ethically and socially undesirable effects of the system will be detected, stopped, and prevented from reoccurring.
- *To a degree matching the type of research being proposed (from basic to precompetitive) and as appropriate*, the application includes a formal ethical risk assessment for the proposed AI system, with documentation for the procedures for risk assessment and mitigation after deployment.
- Whenever relevant, procedures and mechanisms are in place so that end-users, data subject and other stakeholders will be able to report complaints, ethical concerns or adverse events and for evaluation and mitigation of such concerns and events.
- The application is built to be auditable by independent third parties.