Ethics by Design and Ethics of Use in AI and Robotics

Lead contributor
Philip Brey, University of Twente (p.a.e.brey@utwente.nl)

Other contributors
Brandt Dainow, University of Twente

Due date

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.

© SIENNA, 2020
This work is licensed under a Creative Commons Attribution 4.0 International License
Abstract

This document offers guidance for an Ethics by Design approach when developing AI-driven systems, including robots. The Ethics by Design approach 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 problems are not generated in the first place by using ethically-focused activities throughout of the design, development and deployment phases of a project. We first detail foundational ethical values which all AI-driven systems should comply with. We then extrapolate these into specific features an AI system should possess. In order to make this document as useful as possible, these features are, as much as possible, presented as tasks to be performed. Finally we show how to apply these concerns within each stage of the design, development and deployment stage.

Document history

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Description</th>
<th>Reason for change</th>
<th>Distribution</th>
</tr>
</thead>
<tbody>
<tr>
<td>V.X.X</td>
<td>dd Mmm YYYY</td>
<td>First Draft</td>
<td>V.X.X</td>
<td>dd Mmm YYYY</td>
</tr>
<tr>
<td>V.X.X</td>
<td>dd Mmm YYYY</td>
<td>Comments</td>
<td>V.X.X</td>
<td>dd Mmm YYYY</td>
</tr>
<tr>
<td>V.X.X</td>
<td>dd Mmm YYYY</td>
<td>Draft</td>
<td>V.X.X</td>
<td>dd Mmm YYYY</td>
</tr>
<tr>
<td>V.X.X</td>
<td>dd Mmm YYYY</td>
<td>Final Version</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Information in this report that may influence other SIENNA tasks

<table>
<thead>
<tr>
<th>Linked task</th>
<th>Points of relevance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number and title</td>
<td>What should be considered by linked tasks</td>
</tr>
</tbody>
</table>
Table of Contents

Abstract .......................................................................................................................... 2
Table of Contents .......................................................................................................... 3
Executive summary ....................................................................................................... 4
List of figures ................................................................................................................ 6
List of tables ................................................................................................................ 6
List of acronyms/abbreviations .................................................................................... 6
1. Introduction ............................................................................................................. 7
2. Ethics by Design Principles .................................................................................... 8
   2.1. The Ethics by Design approach ........................................................................ 8
   2.2. Values and Ethical Requisites of Ethics by Design ........................................... 10
   2.3. Conclusion ......................................................................................................... 17
3. How to apply Ethics by Design in AI development - a practical guide for system developers 18
   3.1. Generic model for design .................................................................................. 19
   3.2. Design Phase: Specification of objectives ....................................................... 22
   3.3. Design Phase: Specification of requirements .................................................. 24
   3.4. Design Phase: High-level Design ..................................................................... 26
   3.5. Design Phase: Data collection and preparation .............................................. 29
   3.6. Design Phase: Detailed design and development ........................................... 31
   3.7. Design Phase: Testing and evaluation ............................................................. 34
4. Ethical Deployment and Use .................................................................................... 37
   4.1. Project planning and management .................................................................... 38
   4.2. Acquisition ....................................................................................................... 38
   4.3. Deployment and implementation ..................................................................... 39
   4.4. Monitoring ....................................................................................................... 39
5. Glossary of terms ................................................................................................... 41
6. References and further reading ............................................................................. 44
7. ANNEX: Incorporation with specific design methodologies .................................. 45
   7.1. Mapping the generic model onto your design methodology ........................... 45
   7.2. Ethics by Design in Agile ................................................................................. 47
8. Annex – Organisational Adoption of Ethics by Design ........................................ 51
Executive summary

The Ethics by Design approach detailed here offers a way by which to include ethical principles and procedures into the design and development processes of AI-driven systems. Historically, ethical problems in AI have only been detected after the system has been deployed. It is now understood that it is possible to anticipate and avoid ethical issues in advance, rather than wait for harm to arise and then fix the problem. As part of the “By Design” movement, Ethics by Design is intended to prevent systems being created with ethical issues at all, just as Privacy by Design seeks to prevent systems being created which have privacy issues. 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. In addition to better (and formal) ethical assessment of their operation and impact before they are built, this requires changes in the way systems are developed. The aim of Ethics by Design is to ensure ethical problems are not generated in the first place by using ethically-focused activities and tools throughout of the design, development and deployment phases of a project. These activities are detailed in this document, as are the ethical values these activities uphold. Where tools exist to programmatically handle ethical issues during development, these are indicated.

The ethical values upon which Ethics by Design is based are drawn from previous research into responsible innovation within the EU and from international standards such as the Universal Declaration of Human Rights. These values are grouped into six categories, such as fairness, accountability and transparency. Applying these to AI and robotics, we then develop “ethical requisites,” which are the conditions that an system 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 a system does 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. If others object that the system is biased, 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. We then derive from these ethical requisites sets of ethical guidelines to be followed at different stages of the design, development and deployment of the system. These guidelines are concrete tasks which must be performed in order to achieve the ethical requisites.

The model of development used here is necessarily non-specific, using a generic model of system development. In order to assist the reader in understanding how to convert the generic model into something within their own methodology, this document concludes with an Annex which demonstrates how to position the ethical requisites and their corresponding tasks under the Agile methodology.
The main ethical requisites for AI and robotics systems can be summarised as follows:

- Because each individual has an inherent worth, AI systems should not negatively affect human autonomy, freedom or dignity, nor should they limit participation in democratic processes.
- 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 directly cause harm, rely on harmful technologies or processes, or influence others to act in ways which cause harm to individual, societal or environmental well-being.
- Human oversight and accountability are required to ensure conformance to these principles and address non-compliance.
- Systems should be as transparent as possible because only then are accountability and human oversight possible.
List of figures

- **Figure 1**: The 5-layer model for Ethics by Design
- **Figure 2**: The values and ethical requisites of Ethics by Design.
- **Figure 3**: Mapping the Generic Model to Agile

List of tables

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

List of acronyms/abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td>AI</td>
<td>Artificial Intelligence</td>
</tr>
<tr>
<td>GDPR</td>
<td>General Data Protection Regulation</td>
</tr>
<tr>
<td>IEC</td>
<td>International Electrotechnical Commission</td>
</tr>
<tr>
<td>IEEE</td>
<td>Institute of Electrical and Electronics Engineers</td>
</tr>
<tr>
<td>SHERPA</td>
<td>Shaping the ethical dimensions of smart information systems– a European perspective</td>
</tr>
<tr>
<td>XAI</td>
<td>Explainable Artificial Intelligence</td>
</tr>
</tbody>
</table>

Table 1: List of acronyms/abbreviations
1. Introduction

This guidance document offers guidance for people who wish to develop AI-based systems (including robotics), and a set of potential assessment criteria and values for those who have a concern for the ethical status of AI systems. This document outlines an “Ethics by Design” approach, which aims for the systematic inclusion of ethical values, principles, requirements and procedures into design and development processes. Traditionally, ethical issues in AI systems have been discovered after the systems have been completed, usually only once they start to cause harm. The Ethics by Design approach is intended to ensure ethical problems are not generated in the first place. This requires specific ethically-focused activities at each stage of 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 AI and robotics projects should comply. However, it is important to bear in mind the importance of any particular ethical value will depend on the type of application and the relevance of that value to it. Not all values are equally important, so judgement must be exercised when considering them.

Using a generic model of the design process, we then offer detailed explanations of the factors which will ensure the system is ethical at each stage of the design process. We start with six ethical values, such as fairness. We then explain how each value can be embodied in AI and robotics systems as an “ethical requisite,” (a project requirement). 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 system developers can meet these ethical requisites at each stage of the design and construction process.

Finally, 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

This section explains the principles by which ethical concerns can be factored into the design process. Chapter 4: Ethical Deployment and Use, p.37, is intended for those deploying AI or robotics systems.

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 becoming the globally recommended approach for AI development. 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, 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 and start causing harm. 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. Ethics by Design is therefore something which affects the planning and creation processes by which to build AI-driven systems.

5-Layer Model of Ethics by Design

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

1. Ethics by Design Values – These are the primary ethical values by which we want to guide the ethical status of an AI or robotics system. 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 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, ethical 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 found 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. We offer an example in this document by applying Ethics by Design to the Agile development methodology. If a project is using a different methodology, a developer can refer to the generic model. By mapping the steps in the generic development process to their own methodology, they can then allocate each guideline to the appropriate steps in their methodology. Indeed, this is the method by which we have developed the specific guidance for the Agile methodologies in our example.

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 interrogate 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.
2.2. Values and Ethical Requisites of Ethics by Design

Ethics by Design is based on ethical values such as privacy and 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 and in the development process.

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

- Human Agency
- Privacy and Data Governance
- Fairness
- Well-being
- Accountability and Oversight
- Transparency

Under each category we will describe the values and provide examples of corresponding ethical requisites for AI and robotics systems.
The following format is used:

**Title of Value**

Several paragraphs of text to explain the value.

**General ethical requisites**

A generalised description of operational and other features which are required of AI systems in order to meet the ethical requisites.

**Human Agency**

Human agency encapsulates the values of autonomy, dignity and freedom. These are the fundamental rights upon which the EU is founded. They are also the rights enshrined in the UN Declaration of Human Rights. Respecting autonomy means allowing people to decide for themselves what is right and wrong and the way they should live their life as a consequence. Dignity means every human being possesses an intrinsic worth which should never be compromised by others, including AI. Humans derive dignity from their capacity to determine what is right and wrong for themselves (their autonomy). This means they have the right not to be treated as a tool in the service of others or as “a means to an end”, but as a unique entity of inherent worth.

Human autonomy can take many forms. This is because autonomy means each person deciding for themselves what their own personal form of autonomy is. Consequently what constitutes human autonomy is as varied as people. As a result, it does not require positive actions for a system to restrict autonomy. Systems can restrict human autonomy without doing anything - simply by not catering for the full range of human variation in lifestyle, values, beliefs and all the other aspects of our lives which make us unique. This is often done with the best of intentions – restricting choices or decisions to those which the developer’s consider optimal, often simply because they never realised other people might think differently. This is a particular problem with personalisation services, which may not cater for some lifestyle choices, or which fail to respect cultural norms in other societies.

Respecting freedom means leaving people free to exercise their autonomy and live with dignity. Most importantly, freedom requires individuals have the ability to decide for themselves any matter which they think is so important that they want to decide it for themselves. Respecting dignity and autonomy means no one can tell another person that an aspect of their freedom is not important if that person thinks it is. In addition to the freedom to act, this includes freedom from constraints which conflict with one’s autonomy, such as coercion, deception and manipulation.

**General ethical requisites**

- It should be clear to people whether they are interacting with an AI system. They should be informed about the system’s abilities and how to judge and interact with them. This means that if an AI system is interacting with people, it should have specific features which inform people of the system’s presence and abilities, including its limits.
• AI systems should not subordinate, coerce, deceive or manipulate people, and should not create attachment or stimulate addiction. Furthermore, AI systems should not limit freedom of expression, access to information, freedom of assembly and association, or any other rights.
• AI systems should not be designed for uses in which human beings are objectified or dehumanized.

Privacy & Data Governance

As a value, data governance means humans must actively manage their personal data and the manner in which the system uses it. Data governance includes issues relating to quality and accuracy of data, access to data, as well as other data rights such as ownership. Ethical issues can arise from both non-personal data (e.g. racial bias) and personal data (where the data subject’s rights and freedoms must be safeguarded).

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.¹
• AI systems should support the right of an individual to withdraw consent for the use of their personal data. This means there needs to be a mechanism in place to allow them to object to its use.
• In the case that personal data is processed, an AI system should process it lawfully, fairly and transparently, in line with data minimisation principle.
• GDPR and similar regulations require that technical and organisational measures be in place to safeguard the rights and freedoms of the data subjects through measures such as anonymization, pseudonymisation, encryption, and aggregation.
• Strong security measures to prevent unauthorised access, data breaches and data leakages should be set in place (such as limiting access to qualified personnel, mechanisms for logging data access and making modifications).
• Data should be acquired, stored and used in a manner which can be audited by humans.

Fairness

‘Fairness’ is used here in a philosophical sense, not to be confused with mathematical fairness or use of the term within computational modelling. Fairness in this context has three possible meanings, depending on the context; sameness, deservedness, and compliance. Sameness means that each

¹ See https://ec.europa.eu/info/sites/info/files/5._h2020_ethics_and_data_protection.pdf
person is treated the same. Deservedness means ensuring an equitable distribution so that each get what they deserve. Fairness as compliance means operating in compliance with relevant rules.

Fairness means that all people have the right to be treated appropriately and not on the basis of irrelevant characteristics. In this sense non-discrimination is the application of fairness in the context of human characteristics. In particular, people should not be treated 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.

### General ethical requisites

- **Avoidance of algorithmic bias**: AI systems should be designed to avoid bias in both input data and algorithm design. Bias is a specific concern which needs specific mitigation techniques. AI development should contain specific steps to ensure data about people is representative and reflects their diversity. Similarly, the development process should have formal plans to look for and avoid errors in the selection of 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 certain groups of people from consideration.
- **Universal accessibility**: AI systems should be accessible to, and usable by, different types of end-users with different abilities and means - taking into account relevant criteria such as cognitive ability, special needs and access to certain types of hardware or software. Wherever possible Universal Design principles should be used throughout the planning and development processes.
- **Fair impacts**: Before an AI system is created, developers should formally assess possible social impact on relevant groups. If necessary, steps should 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.

### Personal and Social Well-being

The term ‘well-being’ covers a range of properties. Something has well-being when its needs are met and it is able to function properly. The values of autonomy and freedom mean that people can only achieve well-being if they are able to work towards their ambitions and live whatever they consider to be a “meaningful” life.

### General ethical requisites

- AI systems should take the welfare of end-users and of other stakeholders into account and not, on balance, reduce the well-being of users and other stakeholders. An AI system developer should be able to identify who the end-users will be and any other possible stakeholders before constructing the system. Such planning should consider whether the system could reduce their well-being and, if necessary, how this risk will be mitigated.
- AI and robotics development should be mindful of principles of environmental sustainability, both regarding the system itself and the supply chain to which it connects. Negative environmental...
impacts should be avoided. When planning or purchasing a system one should consider the environmental impact of the system and, where possible, the steps which can be taken to reduce it. In the case of robotics systems this should include considerations of the materials used, their origins and what will happen to them when the device is decommissioned.

- AI and robotics systems should not reduce safety in the workplace. Robotic systems which will share an environment with humans or other animals, should possess appropriate safety features. AI systems which can control actuators, open or close doors or windows, activate lighting or signage, or make other changes to the physical environment should contain safety features to ensure the system does not trigger a change which could harm someone (such as opening a window while someone is leaning on it). Robots should have safety-aware collision avoidance mechanisms. Software systems may also need some form of safety planning where relevant. Ideally, such software should be compliant with IEEE P1228 (Standard for Software Safety).

**Accountability and Oversight**

Human oversight as a value requires 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 understand or control something unless one has information about it. Accountability means there are mechanisms 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 (human-in-the-loop, human-on-the-loop, human-in-command), unless the developer can clearly provide compelling reasons why such oversight is not required. Therefore, AI-driven systems should include concrete functionality which will enable humans to understand the decisions made by the system and allow humans to override them or correct erroneous learning outcomes.
- To ensure ongoing ethical status one needs to be able to detect ethically undesirable effects of the system on end-users or subjects. The organisation using it needs to have a plan for how to stop those effects. This means the system must also be designed with mechanisms for correcting the AI behaviour so that these effects do not recur.
- There needs to be a formal ethical risk assessment for any proposed AI system. There also needs to be a procedure in place for risk assessment and mitigation after deployment. This is largely the responsibility of the operating organisation, but the system will need to include features to implement this.
- The operating organisation will need formal procedures so that third parties (such as suppliers, end-users and workers) can report potential ethical concerns about the AI system. Mere reporting is not enough; it needs to be evaluated and actioned. The requirement for transparency of AI systems means there also needs to be a procedure by which to communicate with these third parties what has been done with their information.
- The operating organisation will need processes by which data subjects can complain if they feel they have been negatively affected by the system. There need be mechanisms for redress. The AI system needs to have functionality which can implement such redress if necessary.
- AI systems should be auditable by independent third parties, through the establishment of mechanisms which facilitate auditability. Increasingly, this is becoming a legal requirement. Ideally, your development processes should follow best practice in XAI\(^2\). Before a designer starts construction, they plan for the facility for ethical audit of the system. This is not limited to auditing the decisions (or other outcomes) of the system itself, but will need to consider tools and procedures used during the development process, including learning models, data sources, annotation processes and decisions made to address potential ethical issues (such as bias in datasets). Where relevant, AI systems should generate human accessible logs of the AI system’s internal processes, input, output, and positive and negative impacts.

**Transparency**

Transparency directly enables human agency, data governance, 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 this level of transparency, a decision cannot be contested, or even understood. This would make it impossible to correct errors and unethical occurrences. The degree to which transparency is needed depends on the context and the severity of the consequences. However, 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 you, 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. At minimum this calls for a mix of genders and ages. Where the formation of a formal group is not possible, it is recommended you take steps to ensure you understand the full range of positions others may take, rather than simply rely on your own opinions.

**General ethical requisites**

- There is a general requirement for traceability across all areas of ethical AI and robotics. A system design (and the processes of construction) should include measures to facilitate the traceability of the AI system during its entire lifecycle, from initial design to post-deployment evaluation and audit

---

- The purpose, capabilities, limitations, benefits and risks of the AI system and of the decisions conveyed by it should be openly communicated to end-users and other stakeholders, including instructions on how to use the system properly. This ethical requisite is fundamental to meeting a number of the requirements above and is referenced in those sections. Wherever it is necessary that people can audit, query, dispute or seek to change AI or robotics activities, such as decisions or learning processes, the system will need formal mechanisms to make this possible. However, this does not end with the system itself; it needs to include governance and other organisational processes (including those in the development stage, such as code documentation) by which to receive and assess requests from third parties.

- 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 AI development processes (and tools) 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 organisations can learn from the experience and improve their handling of ethical issues.

- It needs to be made clear to end-users that they are interacting with an AI system – especially for systems that simulate human communication, such as chatbots. An AI system should have specific features to do this. These should not depend on particular educational backgrounds, technical knowledge or other skills which cannot be assumed of all people.

- Decisions made by an AI system should 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.
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. 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).
3. How to apply Ethics by Design in AI development
- a practical guide for system developers

This following section uses a generic model of system development to detail the specific points any AI-driven system should address in order to achieve the ethical requisites discussed above. The aim of this section is to be of immediate practical use when designing a system. We therefore list these as concrete tasks a designer should undertake as much as possible.

Ethics by Design uses a generic model of the design process by which software systems are produced. Under this approach the ethical concerns to be addressed are treated as general system requirements, just like reliability – as requirements any and all systems must achieve. Just like reliability, these requirements of an ethical nature place obligations not only on the system’s features, but also on the development processes and tools themselves. Our model of the development process explains how to embody these ethical factors in the design and development processes. This model positions ethical requisites in the phase of the construction process to which they relate as concrete tasks to be undertaken. By mapping their own development methodology to the generic model, a developer can determine the relevant ethical requisites for each element of their methodology. Ethical requisites will then be instantiated in that methodology as tasks, goals, constraints and similar guidelines. If these are adhered to, the chances of ethical concerns surfacing is minimised because each step in the development process will contain measures to prevent them arising in the first place. This is the essence of Ethics by Design – don’t allow ethical issues to arise in the first place.

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

In this section we directly address developers. However, managers can use this material to understand what is required of their developers, while purchasers of AI systems can use it to understand what they should look for when seeking ethical AI products.

Preliminary Note for Development Managers

Moving to Ethics by Design is a form of business process reengineering, with all the attendant difficulties – political, organisational, managerial and financial.

Ethical AI cannot be achieved purely through the characteristics of the system produced. Ethics by Design is, by definition, embedded into the development process itself. This means developers must change how they work. This is likely to require alterations to team structures, such as adding additional roles, creating new communication channels, building additional review and decision processes. It will also require additional tools and alterations to existing ones. Some tools, such as version control systems, may be completely incapable of modification to the degree necessary and may need to be replaced. Some aspects of development will be more onerous than before because they will add additional documentation requirements or additional considerations. In many ways, moving development to Ethics by Design is similar to moving from one type of programming language to a different type, such as moving from a procedural language to an object-oriented one.
Many developers, especially senior ones, depend for their status amongst their workmates on their expertise in the way the organisation codes. It is well known that your best developers will be 10 or 100 times more effective than the average. No organisation should lose them to the competition because they are unhappy with the changes required. Best practice is to use these individuals to lead these changes, rather than impose them from above. Resistance is inevitable unless the political dynamics of the development teams are taken into consideration.

As a result, it cannot be assumed that moving to Ethics by Design is simply a question of adding some considerations to the design of a system. Significant changes in the work environment are inevitable. A manager’s responsibility is to ensure these changes are not disruptive, but rather enhance the motivation of staff and the productivity of the development processes. Only when the organisation’s culture values ethical AI to the same degree it values reliability, profitability or customer satisfaction can full compliance with Ethics by Design methodologies be expected of all staff.

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. This model involves six broadly described tasks (sometimes also called phases). While the six are presented here in a list format, this is not necessarily a sequential process. For example, some methodologies, such as Agile, use cyclic models.

The six tasks in the generic model are:

1. **Specification of objectives.** This is the determination of what the system is for and what it should be capable of doing.
2. **Specification of requirements.** This is development of the technical and non-technical requirements and constraints by which to build the system. This includes initial determination of required resources, together with an initial risk assessment and cost-benefit analysis, resulting in a design plan.
3. **High-level design.** This is the development of a high-level architecture and is sometimes preceded by the development of a conceptual model.
4. **Data collection and preparation.** Data must be collected, verified, cleaned, formatted and integrated.
5. **Detailed design and development.** This involves the actual construction of a full working system. For software development, this will involve programming and coding. Robotic systems will also include a manufacturing component.
6. **Testing and evaluation.** This is the process of testing of the system and evaluation against the original objectives and requirements.

We will now briefly indicate how the ethical requisites can be instantiated in procedures within each task.
**Specification of objectives**

As part of the specification of objectives task, the system’s objectives need to be evaluated against the ethical requisites presented in Section 2.2: *Values and Ethical Requisites of Ethics by Design*, p.10. Some objectives are not ethically permitted under any circumstance. For example, a system cannot be ethical if its objective is to destroy people’s freedom because the objective of the system itself is to directly violate an important value. If it is possible to adapt the objectives so as to make the system ethical, this should be done before proceeding further. 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 maintain an ethical watch over the rest of the process and understand that some aspects of more detailed design will have ethical importance.

**Specification of requirements**

During this phase, design requirements and constraints, selected resources and design plans are assessed against the ethical requisites. At this phase one should determine how features of the system and the construction process facilitate meeting the ethical requisites. For example, it may be found that transparency cannot be achieved using a particular coding methodology or that version control systems need additional components to record decisions taken regarding code changes. Make modifications to enable attainment of the ethical requisites. Ensure that the ethical requisites are included in the final list of product requirements. Ideally, stakeholders should be included in this process.

**High-level design**

High-level design is concerned with the development of the technical and non-technical requirements of the proposed system, and the mechanisms by which this will be achieved, such as version control systems. This often includes initial determination of required resources, together with an initial risk assessment and cost-benefit analysis. This frequently involves high-level architectural design, such as overall database and application layer architectures, perhaps some critical schemas, information flow and security requirements. In many cases this will also include a hierarchical breakdown of the required sub-systems and critical sub-functions within the system, though some will consider this a part of detailed design (Kission, Ding, and Jerraya 1994). Under the Ethics by Design approach, the high-level architecture is developed in accordance with the ethical requisites. Ethical requisites should be treated just the same as any other requirements for the system. Issues that may be particularly relevant in this design phase are those relating to transparency, autonomy, privacy and fairness. 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, employment patterns or criminal sentencing, can be assumed to reflect the biases in that society. Data can never be assumed to be accurate, representative or neutral; it must be demonstrated that it is.

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.

**Detailed design and development**

In the detailed design and development phase, actions which will incorporate the ethical requisites are added to the various subtasks within 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 all 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 getting to this stage of the development process. It is recommended that stakeholder consultation or involvement takes place during this phase.
3.2. Design Phase: Specification of objectives

**General Notes and ethics guidelines**

While each project is unique, Ethics by Design lays down a set of standardised requirements which all AI, robotics and big data systems should meet. For obvious reasons, an important first step is to ethically assess the objectives of a development project (i.e., what kind of technology is being developed and what it its intended functionality and purpose) against the ethical requisites, before any details of the individual project are considered. 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 and using AI to hide this activity.

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

- Assess whether the formulated objectives for the design project will permit the system to meet the relevant ethical requisites. An ethical risk assessment, to be performed later in the specification of requirements phase, can also be applied retrospectively to the objectives, as it may point to further potential ethical issues with them. 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 a project has external stakeholders it is important to plan how to include them in the early phases of the project, especially the specification of objectives and specification of requirements phases. The early inclusion of stakeholders increases the chance that their values, preferences and needs are taken into account, and thereby increases the likelihood that the resulting technology is successful and trusted and attains its ethics requisites. In particular, stakeholders may be aware of other ethical issues which could arise from the use of the system. Stakeholders should be consulted about their preferences regarding what the objectives and requirements should be, their beliefs about what ethical issues are at stake and their recommendations about how these ethical issues should be dealt with. Moreover, it is recommended that project members and stakeholders represent appropriate diversity in terms of gender, age, ethnicity, cultural heritage and viewpoints. In this way an appropriately diverse range of ideas and preferences will inform design choices.

**Ethical Requisites of Design Objectives**

The objectives of the proposed system should be checked against the ethical requisites listed below to see if they are potentially violated. Potential violations differ in their degree of seriousness. If it is possible to adapt the objectives so that the system does comply with the ethical requisites, this should be done before proceeding further. If the aim is fundamentally incompatible with the ethical requisites, the project should not proceed. It is likely that whether the system meets ethical requisites or not depends on specific methods of implementation or construction. If this is the case, proceed, but pass these concerns to those designing the development architecture and maintain an ethical watch over the rest of the development process. Other violations may be only potential violations or be less serious in nature. These concerns do not mean the objective should be abandoned, but that
concrete steps will have to be taken to avoid the system becoming unethical. For example, a voice recognition system which is trained only on people with a strong regional accent may be less reliable for people with a different accent. The solution in this case would be to ensure the training data includes a wide variety of accents.

In one’s assessment of objectives, also consider the proposed system’s potential for misuse. Where possible, modify the system’s objectives to reduce such potential. If the potential misuse is significant, conduct a social 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.

**VALUE: 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.
- An additional consideration is whether the initial plans for what personal and non-personal data will be used is fair and appropriate. For example, it would be both unfair and inappropriate to build a system which assesses people by irrelevant characteristics. If the proposed data source is unfair or inappropriate, either change the data source or modify the objective so that the unfair/inappropriate data source is not needed.

**VALUE: Fairness**

- Check whether the objectives are compatible with the fairness requirements. Particularly important is the consideration of whether violation of any of these requirements would cause people to be significantly disadvantaged socially or politically, reduce the power that they have over aspects of their lives, such as work or lifestyle, or would likely result in discrimination or stigmatisation, either through direct actions by the system, or by likely uses to which it would be put. If so, this would constitute serious non-compliance.

**VALUE: Well-being**

- Check whether the objectives are compatible with the well-being requirements. Particularly serious are those violations that cause people to suffer physical, psychological or financial harm, support processes that are known to cause significant environmental damage, or that are likely to cause significant damage to social processes and institutions (for example, by contributing to misinformation of the public). Less serious violations are, for example, systems that are likely to inhibit communication and impoverish relationships between people. If there is significant potential 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: 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 regarding decision cycles and operations. If it does not, change the objectives or provide compelling reasons why such oversight is not required.

**VALUE: Transparency**

- The ethical requisites for transparency do not usually apply to objectives, but rather to the requirements, architecture and detailed design of the system. So they only have to be considered at this stage to determine the degree to which the system’s objectives permit the required transparency to be built into the system.

### 3.3. Design Phase: Specification of requirements

**General Notes**

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 appropriate according to the ethical requisites of 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. It may even be necessary to document the reasoning which led to the use of a particular technique. Consequently, requirements such as the capacity for human oversight and audit may impose a need to document many internal processes to a greater degree than has previously been the case. For example, while documentation within code has always been considered best practice, it has rarely been necessary and unavoidable to the degree Ethics by Design requires.

It must therefore be recognised it is unlikely that development systems, methods, tools or even organisational structures used on previous projects will be suitable without modification. Systems like git can easily accommodate the additional documentation requirements with a little planning, but others may be completely incapable of delivering the ethical requisites required. Even with a system like git, additional management procedures will be required to ensure developers produce the required documentation, and this will require staff training. As a result, it must be recognised there is
likely to be a need to adapt (or even replace) aspects of customary development systems so that they become capable of delivering the 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. However, one should be extremely rigorous in investigations for suitable tools and cannot merely decide that the traditional methodologies are insufficient as an excuse not to bother. The requirements here 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 important AI development processes. Meanwhile DARPA’s Explainable AI (XAI) (Gunning, 2017) is a rapidly developing set of methodologies and tools by which build effective machine learning techniques which are also explainable to humans and allow for human governance. In these and other cases, such tools are Open Source and freely available to all³.

The degree to which a technical inability to meet the ethical requisites blocks a 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 profoundly affected by its decisions. By contrast, a system which manages a city’s traffic lights has only a very limited impact on the life of individuals, so the need for transparency is much lower. Where it is genuinely technically impossible to meet a relevant ethical requisite, the importance of the requisite for that particular system will be a factor in the ethical status of the product.

**Ethics guidelines**

- An ethical assessment should be done of proposed design specifications, constraints, selected resources and infrastructure. For example, an early choice of deep learning techniques for a system that requires transparency and explainability may be judged not to be the best choice. For example, a design specification that a system use authentication via facial recognition may be undesirable from a privacy point of view.
- Once a complete design plan has been produced, an ethical risk and impact assessment should be performed to assess specific ethical risks that may result from development, deployment and use of the system. Steps should be planned and carried out to mitigate ethical risks. The ethical assessment of objectives and requirements undertaken earlier can be important constituents of

this assessment, but these only assess individual elements of the plan, rather than the plan as a whole. This risk assessment should be updated at later points in the development process as more information relevant to it comes in. A professional AI ethicist, if available, should be able to perform such an assessment in collaboration with members of the development team. Ethical risk assessments are scalable; a simple assessment can often be completed within a time constraint of days and with limited resources, whereas a detailed assessment may involve extensive foresight analysis, stakeholder consultation, mapping of potential risks and development of mitigating actions. Ethical risk assessment should be planned and budgeted for at the appropriate point in the development processes. This assessment needs to be scaled to the innovative nature of the project, the severity of ethical risks that were already identified at the stage of ethics review, and the overall budget of the development project.4

- Ensure that relevant ethical requisites are covered in the list of design specifications. For this purpose, consider inclusion of an Ethical Requisites document for the project. At the Objectives stage 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 during the high-level design and detailed design stage of the project.

3.4. Design Phase: High-level Design

In high-level design, the architecture for a system or software product is specified. The following ethics guidelines apply to this phase in the development cycle.

**VALUE: Human Agency**

- Verify that the chosen architecture allows for an interface based on human-centric design principles which leave meaningful opportunities for human choice, and that it allows for freedom of expression and information.

**VALUE: Privacy & Data Governance**

- Verify that the chosen architecture supports the ethical requisites for privacy and data governance. Ensure the development architecture contains processes, procedures and tools to ensure that personal data is not exposed during development 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, or developers may be given direct access to database contents when all they need is the ability to query the database. It is especially important to ensure developers do not have access to identifiable personal information except where absolutely

---

4 A standard for ethical impact assessment has been developed in CEN working document CWA 17145:2017-2, retrievable at [https://satoriproject.eu/media/CWA17145-23d2017.pdf](https://satoriproject.eu/media/CWA17145-23d2017.pdf)
necessary. GDPR (or other regulations) require that who has such access is formally documented.

- 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 document and justify how the initial data selection was determined sufficient for external audit.
- Data imported into a system may not exactly match what was sought for. It cannot be safely assumed that the data obtained is the data wanted. It may be that the datasets do not contain the data they were supposed to, the data may be incomplete or corrupt, or methods of importing or normalizing the data alter it in unanticipated ways which render it less than optimal. Design formal processes to check for and correct bias (or errors) after importing any data.

**VALUE: Fairness**

- 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 differential access or modify the design because this could be challenged by stakeholders who feel disadvantaged or even legally challenged under disability access requirements.
- 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 to fix this or prepare a formal justification as to why this is impossible to defend against disability, discrimination or other access challenges.
- Undertake an accessibility assessment of the interface and other touchpoints. Ensure that, where relevant, the system meets accessibility standards.

**VALUE: Well-being**

- An initial rough environmental assessment should have been conducted during the objectives phase. Once high-level design of the system is complete, this assessment should be taken into more depth. In addition, documentation should be prepared to demonstrate how the system will be constructed in the most environmentally friendly way possible.
- Evaluate whether the system, as defined by the high-level design, could cause physical harm to people or property. This is especially important with robotic systems. If this is possible, ensure design features to minimise this risk and/or the amount of harm which can be done, such as safety buffers, emergency stop buttons. If the system will be able to respond to voice commands, include emergency stop vocal commands in the design.

**VALUE: Accountability & Oversight**

- Design the ethical governance model for the development process. This focuses on mechanisms which will enable human oversight during the development process. There are two main elements to design; a technical ethical compliance system embedded into the development architecture and a set of organisational structures and procedures. The ethical
compliance architecture will need to focus on tools and processes at the developer level, such as Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2020), but will also need mechanisms for external communication from end-users and other stakeholders during testing and evaluation. The ethical governance model will also need to include organisational structures for actually doing the governance, such as ethical review committees and/or ethical compliance officers at developer level. Ensure such mechanisms include management procedures which ensure these mechanisms are actually used, such as formal reporting and assessment by senior management. The governance model needs to address the following issues: How will governance be exercised? What is the project’s version of an authority to supervise and ensure the ethical requisites are met? What powers will that authority have? How will it be selected? How can that selection process be demonstrated to be fair and inclusive? What procedures will be used in the case of a conflict between the ethical governance authority and developers or engineers or clients?

- 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 purpose. It may be possible to design oversight during development which can also be used once the system is deployed, but this may be more difficult than simply designing a different mechanism tuned to oversight of the deployed system. In either case, 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.

- There is an increasing tendency to demand AI systems can be externally audited for ethical compliance. Even if this is not the case for your type of project now, audit requirements could arise during the lifetime of the system. Ensure that the system is designed in a manner which permits such auditing. If unsure, refer to existing ethical audit procedures, codes of conduct for ethical AI.

**VALUE: Transparency**

- Design mechanisms to document how data acquisition, storage and use happen. This needs to be auditable by any people who need to check for ethical compliance, including users and other stakeholders, those responsible for ensuring the data practices fulfil the ethical requisites, external ethical AI auditors, regulators (where regulations apply) and any other person who has a need to determine the ethical status of the system’s data and use thereof. This consideration must cover both the development process and use once deployed.

- Design procedures and select and configure tools to document development processes to a level that humans can understand and evaluate decisions made within the design and development processes. This will be required for any people who need to check for ethical compliance within the development process. This can be anyone who has a concern the system is unethical in some way and wants to determine if this was caused within the development process, or who simply wants to understand how ethical concerns were dealt
with while the system was being created. This can therefore include users and other stakeholders, those responsible for ensuring the created system meets its ethical requisites, external ethical AI auditors, regulators (where regulations apply) or managers assessing the development process in order to emulate it in other projects or to determine how to improve it. 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 and want justification, explanation, or alteration, as part of a normal internal ethical governance review, because this information is required by developers developing other aspects of the system, or for other reasons.

- 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. Such systems should be targeted for evaluation as part of the testing regime.

- 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 a nuanced or educated understanding of AI operations 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 internal staff and third parties (e.g. suppliers, consumers, distributors/vendors) can report potential vulnerabilities, risks, or biases in the system, during the development process.

3.5. Design Phase: Data collection and preparation

**General Notes**

For systems that involve data processing, data must be collected, verified, cleaned, formatted and integrated. Data collection involves the collection of initial data, its description and initial analysis, and verification of quality. To integrate ethical requisites into this process, assess how different steps in the process might support or 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 re-evaluated. In this phase, fairness (including bias, discrimination, and diversity), privacy and data quality will be particularly important.
The processing of personal data is governed by GDPR and the specific national and sectorial legislative frameworks. Personal data is any information that relates to an identifiable living individual. Items of information which have the capacity to be amalgamated and then identify a particular person also constitute personal data, whether being so used or not. Online identifiers and location data also constitute personal data. Personal data that has been de-identified, encrypted 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 data which is particularly sensitive and must be treated with special attention. Such data are: data concerning racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic, 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 detailed information and guidance consult your data protection officer(s).
- 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;
- All personal data must be processed in lawful, transparent and fair manner. If the planned system will process personal data, you must incorporate the rights to data protection into your design. This includes how to enable individuals to withdraw consent for the use of their personal data, and what mechanisms will enable them to object to its use.
- Within the limits of current technology, the design should ensure that data controllers and data processors are able to fulfil their data protection obligations.

VALUE: Fairness

- Is it possible some of the data gathered could be biased in its representation of different groups, persons, or social entities, for example by overrepresentation of some categories, a lack of diversity in representation, or implicit stereotyping? If so, modify the criteria by which data will be selected to reduce such bias and/or plan steps to rectify the datasets once they are in the system. The requirements for transparency and oversight will 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.
- Ensure that input, training and output data is all analysed for harmful bias (e.g., some data sets may contain harmful biases if they consist solely of the behaviour of subclasses of all people, e.g.,
young white men, and if the system is deployed in situations where groups other than those in the data set will be affected).

- Where it is determined that harmful bias is possible, build mechanisms to avoid or correct it.
- 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 (e.g., the cleaning of the data set may inadvertently remove data relating to certain minority or under-represented groups, leaving the data set as a whole biased). 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: 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, such as cloud services, ensure these are also compliant with data protection requirements. It is not safe to assume their assurances are sufficient. 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.

**VALUE: Transparency**

- Prepare a data protection 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 produce a risk mitigation plan.
- Ensure that you can explain to others how personal data is used, shared, and stored.

3.6. Design Phase: Detailed design and development

**General Notes**

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. Just as Ethics by Design calls for
ethical matters to be dealt with during the development phase, so existing development tools and processes will need adaptation to support this activity. To integrate ethical requisites into this process, 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 decisions regarding them. Issues that may be particularly relevant in this design are those relating to transparency, privacy and accountability.

**Detailed Design and Development**

**VALUE: Privacy & Data Governance**

- If creating new personal or sensitive data (e.g., through estimation of missing data, the production of derived attributes and new records, data integration, or aggregation of data sets), further informed consent may need to be acquired. Please remember this document does not offer definitive guidance on GDPR compliance, and more authoritative information in this regard should be sought.
- Make sure all newly created personal or sensitive information/data is given at least the same protection and attracts the same rights as previously collected or held personal or sensitive information/data.
- Ensure no new personal information is, or can be, collected or created during development of the system, unless necessary. If new personal information is collected or created, then have systems in place to impose access or use limitations which will protect individuals’ privacy or sensitive information/data, and further informed consent is acquired, if needed.
- 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 in control of the quality of the external data sources used, assess to what degree the quality can be validated.
- Establish a developer culture of shared responsibility for the organization’s data assets. Make sure this culture understands the potential value of data assets. Ensure the impact and risk of data loss is continuously communicated and 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 relevant staff understand them.
- 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 (which it is in the EU), it should also support the right to be forgotten (from internet searches and directories). Steps must therefore be taken to guarantee a person can access their personal data, and in a manner which protects other individual’s privacy.
- Make sure no new personal information 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. The design 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.
• Ensure there are oversight mechanisms for data processing (including limiting access to only appropriate personnel, mechanisms for logging data access and making modifications).
• 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 (i.e.: accurate, fairness, appropriateness, security) prior to its use by the system. This should include measures to prevent external corruption, such as hacking. Ensure the data integrity systems are designed to prevent unauthorised manipulation of data and to mitigate against silent and other forms of low-level data corruption.

VALUE: Fairness
• Check for algorithmic bias during the detailed development phase. Data could be processed in a biased way, and therefore algorithms should be checked for this.
• Ensure that interface design honours principles of universal accessibility, and avoid the introduction of functional biases in the detailed development phase which could make the system unequally functional for different types of user.

VALUE: Well-being
• Follow resource-efficiency and sustainable energy usage practices. In particular, decisions made by the system that will directly affect the non-human world around us need to be carefully factored in, with strong emphasis on the impact on these ecological externalities.

VALUE: Accountability & Oversight
• Create a developer culture in which it is seen as important to deal with ethical issues in a timely fashion. Do not allow a culture to develop in which dealing with ethical issues is seen as a hassle and something to be addressed after “more important” work is completed. Most importantly, make sure a culture does not arise in which departures from the ethical requisites are treated as something to be fixed after the entire system is completed.
• 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.
• 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.
• 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.

VALUE: Transparency

• Measurements to ensure traceability to the degree needed should be established within the following methods:
  o 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).
  o 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.
  o 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).
• Could the system 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.
• Is it 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 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).

3.7. Design Phase: Testing and evaluation

The following general and value-specific guidelines apply to the testing and evaluation phase.

As part of the testing and evaluation phase, perform an ethical assessment to assess how well the system meets the ethical requisites. Possible outcomes are that ethical issues have been dealt with in
a satisfactory way, that further development is needed, or that specific guidance for, or restrictions on, deployment and use need to be in place to mitigate ethical issues.

Use the project’s ethical requisites document to design a testing regime to check the system’s compliance with its ethical requirements. While some aspects of the ethical requirements are likely to be factors in normal testing, it is highly unlikely any standard testing regime will consider all of the system’s ethical requisites. The choice of testing methodology is important here. For example, metamorphic testing is popular with machine learning (Xie et al., 2011) and can easily accommodate testing against ethical requisites if suitably designed, whereas techniques such as unit testing will need significant work to be a suitable testing methodology for ethical compliance (and highly unlikely to be capable of testing all ethical requisites). Implement and evaluate this testing of the system to determine whether it meets all of its ethical requisites. Treat departures from the system’s desired ethical characteristics just as seriously as any other type of bug and undertake remedial work to make the system meet its ethical requisites.

It is highly recommended that stakeholder consultation or involvement takes place during this phase in order to collect their viewpoint on whether ethical requisites have been met in a satisfactory way and to discuss what should be done when this is not the case.

**VALUE: Accountability & Oversight**

- 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 and perception of the system’s functionality and behaviour by end-users and other directly affected stakeholders. Even simple items like interface messages can be misinterpreted by those without a nuanced technical understanding. It cannot be assumed others will understand the system or its output in the same way as developers. Test the understanding of users and other affected persons regarding what the purpose of the system is, 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.

**VALUE: Transparency**

- 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.

- 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 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.
4. Ethical Deployment and Use

In this section, we will present ethics guidelines for the deployment and use of AI systems. We make a distinction between the development process for an AI system and its deployment and use after development, and offer separate ethics guidelines for both. We take as our principal actors the project team, while also taking into account that they will be operating in an organisational context.

Our guidelines for Ethical Deployment and Use apply to four practices we consider central to the deployment and use of AI systems in research projects: project planning and management; acquisition, deployment and implementation; monitoring.

- **Project planning and management** refers to the planning of a new research project, normally reflected in a project plan, and the management of the planned activities after the project has begun. Our ethics guidelines address what steps should be taken by project management in project planning and general project management in order to ensure proper consideration of ethical issues in the deployment and use of an AI system.

- **Acquisition** refers to the process of acquiring an AI system which is to be deployed and used in the project. In some projects, the system will be acquired from an external developer or vendor. In others, it will be developed in the project itself. A combination of external acquisition and in-house development is also possible. An organisation is responsible for the ethical state of any AI system it uses, even if that system has been built by another. As a result, external acquisition imposes unique ethical tasks not required when the system is developed in-house.

- **Deployment and implementation** refers to the process of deploying the AI system into a user environment, as well as planning and implementing required changes in the organisational context to ensure its successful implementation. It normally involves the development of an implantation plan, the preparation and training of stakeholders, the development and implementation of an operation and use plan, the configuration of the system and its imbedding in IT infrastructure, the testing of the system in its new environment, the implementation of needed organisational changes and new policies, and post-implementation review. The manner in which a system is deployed or implemented may change the ethical characteristics of the system. For example, the system may be deployed to work with different datasets from that on which it was trained. As a result, it cannot be assumed that the ethical characteristics of the system will remain unchanged when it is deployed. Deployment and implementation therefore imposes its own tasks to ensure the system continues to meet its ethical requisites.

- **Monitoring** is the process within project of monitoring the performance of the AI system, its conformance and compliance with external requirements, and the development and implementation of plans for improving its performance. No matter how robust the testing regime, the full ethical characteristics of a system may not be apparent until the system is deployed “in the wild.” The most common (but not the only) concerns are that the system may have completely unexpected (and untested) effects on users; its own internal processes may change as it learns; or the data it uses may lose ethical integrity. As a result, all AI systems require perpetual on-going ethical monitoring and, where necessary, adjustment. This is typically done with an audit procedure, which is becoming an increasingly common legal requirement.
We assume that all four of these processes take place when an AI system is deployed and used in a research project, and proceed to outline ethics guidelines for each of them.

4.1. Project planning and management

- In the project plan, ensure that you budget for Ethics of Use actions and include tasks or subtasks for these actions. The budget should be sufficient to ensure proper adherence to the Ethics of Use guidelines in the project. In budgeting and planning, take into account the potential ethical issues that were revealed in the ethics self-assessment.
- In the project plan define roles, responsibilities and procedures for implementation of the ethics guidelines and for monitoring and assessment of their implementation. This could include the institution of an AI ethics officer and the assignment of specific responsibilities to implement ethics guidelines or monitor their implementation by researchers in the project. It should not be assumed that whoever managed ethical compliance during development, even if available, is the appropriate authority for this role.
- Ensure that the objectives for which you want to use the system and the design requirements and resource choices conform to the ethical requisites provided for in the Ethics by Design objectives and requirements phases.
- Your plan should include details of the procedures for inclusion of stakeholders in decisions regarding the acquisition, deployment, implementation and monitoring of the use of the system. These procedures must ensure that stakeholders are, at a minimum, consulted regarding their values and interests with respect to the deployment and use of the system.

4.2. Acquisition

- If an AI system is externally acquired as an off-the-shelf solution, consider available options and 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.10. If a system does not meet the ethical demands contained in this guidance document, consider whether adaptations can be made to the system or focus on acquiring a different system. If the AI system is custom-built by an external developer, then give preference to 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. If the developer cannot demonstrate these, it is unlikely they will be able to ensure the ethical requisites are being met in the system itself. Without sufficient transparency, it will not be possible to determine the ethical compliance of the system.
- If the AI system is custom-built by an external developer, then give preference to a developer who uses an Ethics by Design approach or who is otherwise willing to adhere to the ethical requisites as listed in Section 2. If possible, verify that the system adheres to these requirements. A simple way to start is to ask the developer to explain the mechanisms by
which they operate Ethics by Design, such as documentation and ethical governance procedures. For example, they should be able to show how they document their datasets and models used for machine learning, including how they check for and eliminate bias.

- If in-house development is chosen, then follow the Ethics by Design methods presented earlier in this document, and verify that the resulting system adheres to the ethical requisites listed here.
- Ensure that any data that is 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. 29.
- 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 planned and 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.

4.3. 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 and procedures and pay attention to ethical aspects within communication regarding 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.

4.4. Monitoring

After launch of the system, continuous or periodical monitoring is required to ensure successful ethical compliance over time:

- Verify that end-users use the system according to user policies which include ethical requisites, are vigilant about ethical issues in operation and use, and consult with senior staff on issues that are morally problematic or ambiguous.
- Ensure that monitoring goals and metrics are in place for compliance with the ethical requisites. Periodically monitor compliance and propose improvements if monitoring shows compliance to be below target.
- Ensure that stakeholders, users and subjects of the system have “ethical complaint” communication channels by which to alert you to their ethical concerns as they arise. Ensure that these channels are monitored regularly and concerns are processed appropriately by people with appropriate levels of seniority to ensure action if necessary. Ethical concerns should never just vanish into the system, but this requires formal management and reporting processes to avoid. In addition, ethical problems often occur because a system affects people
who were never expected to be impacted by the system in the first place. Consequently, you should ensure such communication channels are open in a manner which allows unexpected groups to approach you with their concerns, and that these are handled appropriately.
### 5. Glossary of terms

<table>
<thead>
<tr>
<th>Term</th>
<th>Explanation</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Accountability</strong></td>
<td>Accountability applies to both individuals and institutions. It means taking responsibility for your actions rather than trying to shift responsibility (or blame) elsewhere. This involves being able to explain the reasons behind your actions when necessary, and being prepared to discuss your actions and their consequences. It implies a willingness to accept and act on criticism of your actions where that is justified.</td>
</tr>
<tr>
<td><strong>Auditability</strong></td>
<td>Auditability refers to the ability of an AI system to undergo the assessment of the system’s algorithms, data and design processes. This does not necessarily imply that information about business models and intellectual property related to the AI system must always be openly available. Ensuring traceability and logging mechanisms from the early design phase of the AI system can help enabling the system’s auditability.</td>
</tr>
<tr>
<td><strong>Autonomy</strong></td>
<td>Autonomy is the ability to decide courses of action independently of a ruling body. In AI, a machine or vehicle is referred to as autonomous if it doesn’t require input from a human operator to function properly. However, ethical AI is more concerned with human autonomy, of which there are three types. Moral autonomy refers to the innate capacity of humans to determine for themselves what is morally good and bad. Political autonomy refers to the capacity of human beings to form their own political opinions. Personal autonomy refers to the innate capacity of human beings to decide how they should live, especially by what values they should make their decisions.</td>
</tr>
<tr>
<td><strong>Bias</strong></td>
<td>Bias is an unfair or unjustified prejudice towards or against a person, group of people, object, or position. Bias can arise in many ways in AI systems. It does not necessarily relate to human bias or human-driven data collection. It can arise, for example, through the limited contexts in which a system is used, in which case there is no opportunity to generalise it to other contexts. Bias can be intentional or unintentional, but is a danger because it frequently causes discriminatory and/or unfair outcomes in AI systems.</td>
</tr>
<tr>
<td><strong>Discrimination</strong></td>
<td>The act of making unjustified distinctions between human beings based on the groups, classes, or other categories to which they are perceived to belong. Principles of non-discrimination state that in regard to human rights, there should not be any differentiation that is based on inalienable parts of one’s identity, including gender, race, age, sexual orientation, national origin, religion, income, property, health, disability and opinions.</td>
</tr>
<tr>
<td><strong>Diversity</strong></td>
<td>Diversity is the inclusion of different types of people, based on identity markers like gender, race, age, cultural heritage, ability,</td>
</tr>
</tbody>
</table>
educational background, cognitive style and the like. The principle of respect for diversity goes beyond the principle of non-discrimination to include positive valuation of individual differences, recognition of differences in individual need and support for the diverse composition of organisations and communities. Applied in an AI context, respecting diversity accounting for designing for diversity in the composition of data sets that represent people, in user-centred design, in the inclusion of stakeholders and stakeholder perspectives, and in the composition of design teams.

**Ethics**
Ethics is an academic discipline which is a subfield of philosophy. Applied ethics deals with real-life situations, where decisions have to be made under time pressure, and often limited rationality. AI Ethics is generally viewed as an example of applied ethics and focuses on the issues raised by the design, development, implementation and use of 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. Informed consent must be appropriately documented, based on written or otherwise documentable records stemming from a person capable of giving consent or, where the person is not capable of giving consent, by his or her legal representative.

**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. Among these, special categories of data within the meaning of the General Data Protection Regulation concern personal data relating to racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, as well as genetic data, biometric data, data concerning health or concerning sex life or sexual orientation.

**Personal data processing**

Any operation or set of operations performed or not using automated processes and applied to personal data or sets of data, such as collection, recording, organisation, structuring, storage, adaptation or modification, retrieval, consultation, use, communication by transmission, dissemination or any other form of making available, linking or interconnection, limitation, erasure or destruction.

**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. (Cavoukian, 2009)

**Profiling**

According to Article 4(4) of the GDPR, ‘profiling’ means any form of automated processing of personal data consisting of the use of personal data to evaluate certain personal aspects relating to a natural person, in particular to analyse or predict aspects concerning that natural person’s performance at work, economic situation, health, personal preferences, interests, reliability, behaviour, location 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.

Table 2: Glossary of terms
6. References and further reading


7. ANNEX: Incorporation with specific design methodologies

This section explains how the Ethics by Design approach presented in Section 3, which is based on a generic model of the design process, can be incorporated within specific design methodologies. We provide an example of how to do this with the Agile development model. In addition, Annex 2 of D4.7: An Ethical Framework for the Development and Use of AI and Robotics Technologies provides a detailed example of how to do this for V-Model.

7.1. Mapping the generic model onto your design methodology

The Ethics by Design approach is an addendum to design methodologies. It is intended to be grafted into whatever methodologies are being used in the project. For this reason, the Ethics by Design approach is intentionally methodologically neutral. Ethics by Design provides ethical guidelines by which a system can be designed and developed in a manner which ensures it is ethically safe at every stage of its life cycle.

However, in order to be of practical use, the approach must be integrated into the design methodology. We offer an example of integrating the Ethics by Design approach into the Agile design methodologies here. However, Ethics by Design is not limited to this methodologies. The generic model for design identifies six classes of task which must be accomplished in the creation of any AI or robotics system. While they have been presented as a list, that does not imply that they necessarily form a sequence. The development of every system must accomplish these tasks or the system cannot be created, but some, like Agile, vary the sequence. Consequently, any design methodology must include these tasks in some way. Ethics by Design can therefore be integrated into any design methodology by reference to our generic model.

The steps for integrating the Ethics by Design approach to any design methodology are as follows:

1. Identify where each of the generic tasks is undertaken within the target methodology.
2. All values are relevant to all tasks. However, not all ethical requisites will be relevant to all elements of the target design methodology. Identify which of the ethical requisites provided here are relevant to which element of the design methodology.
3. This will result in lists of ethical requisites under each element of the design methodology. Cross-reference these lists against the values to determine whether additional ethical requisites are needed to fully cover the scope of the project. Formulate additional requirements as appropriate.
4. Review the project’s aim, including final functionality and output, data sources and other forms of input and the context in which the system will be used. Consider whether the lists of ethical requisites are sufficient to cover these or whether additional requirements are necessary. It is highly likely that the intended use will generate context-specific ethical requisites. Other defining aspects of the project may also generate the need for specific ethical requisites. This is an especially important consideration where the system will offer unprecedented capabilities or have significant impact on people’s lives.
5. If possible (and appropriate), develop formal systems for ensuring Ethics by Design within each element of the design methodology. At its most basic, this can consist of checklists containing the ethical requisites for each design methodology element. However, additional tools or systems may be required. Some of these may be publicly available “AI ethics” tools, such as Model Cards (Mitchell et al., 2019) or XAI components which are open source\(^5\). Others may be available as add-ons to existing development tools. Some may simply require additional configuration in existing development systems. For example, Git repos can simply be configured to include ethics-related documentation and tools.

**Mapping the Generic Model to Agile**

Agile development is not one methodology but an umbrella term for a collection of frameworks and practices which adhere to the Agile mindset, such as Scrum. Thus, “being Agile” means prioritizing certain ways of working, most importantly responding to change as the project evolves. The Agile mindset is well suited to Ethics by Design because Ethics by Design is also a dynamic approach.

---

\(^5\) Source code for Model Cards can be obtained from [https://github.com/tensorflow/model-card-toolkit](https://github.com/tensorflow/model-card-toolkit). The git repository for XAI contains a growing set of tools and methodologies at [https://ethicalml.github.io/xai/index.html](https://ethicalml.github.io/xai/index.html)
As Figure 3 illustrates, the elements of the generic model do not map one-to-one to those of Agile. Because of Agile’s dynamic and cyclic nature, most steps of the generic model map to more than one phase in Agile, and some Agile phases map to multiple generic ones.

### 7.2. Ethics by Design in Agile

In what follows, the ethical requisites guide software developers during the development process by making them more aware of ethical considerations during development and by providing a framework for handling them as an integral part of the Agile philosophy. When following a truly Agile spirit, ethical incorporation into Agile must come from the developer himself, rather than imposed externally. Nonetheless, the requirements for transparency and human governance dictate that the decisions made by the developers should be well-considered, documented and defensible. It is therefore important to have a work environment where the team feels comfortable communicating their concerns, is given opportunities to voice them, that they are heard, and their concerns acted upon.

**Applying ethics to the Agile approach**

**Requirements Gathering phase**

The requirements gathering phase develops requirements for the end-product. The project’s objectives are assessed to see if they are compatible with the ethical requisites. For example, hidden surveillance breaks the many of the ethical requisites for transparency and human oversight.
problematic requirements may be countered by adding additional features, while others may be changed so as to mitigate their ethical problems. It is highly likely requirements will need to be changed if it is discovered they will have undesirable ethical implications. In extreme cases, the central aim of the project, or an essential feature, may be inherently unethical, in which case the project cannot proceed without fundamental adjustment. It is even possible some projects cannot proceed at all because their central aim is fundamentally unethical.

The ethical requisites for transparency and human oversight are not reflected in the product’s features and not traditionally considered criteria for assessing the success of a software project. However, they can impose significant constraints and demands on the development processes and therefore consume resources. The ethical requisites will be best achieved if they are considered “features” of the product just like any other requirement.

Many AI systems have been judged unethical, not because of flaws in the product, but because of the effect they have on users or society. Often these effects are unanticipated only because no one thought to check for them. It is therefore important to consider the full range of possible post-deployment effects when designing the system. In addition to the impact of expected usage, one should consider the ways in which the system could be intentionally used to cause harm, the possibility of the system being hacked and what could be done with a hacked system, and the possible impact of intentional or accidental data corruption. Here one should also consider the “butterfly effect” in which small changes early in the system can create profound changes in behaviour. For example, a minor change in annotation schemes could profoundly affect learning and thence decisions made by the deployed system.

**Planning & Designing phase**

The planning & designing phase concerns the design of the final system. Note that design is an evolving process under Agile. The Agile methodology works on the basis it is difficult to fully envisage the final product in all details before any development work commences. Furthermore unexpected obstacles are almost certain to emerge during the development process which demand the product change to some degree. The Agile methodology therefore expects that refinements and changes will occur as the system emerges. Ethics by Design lays down the ethical requisites a system should meet. These should be incorporated as appropriate into the design requirements. As the system evolves, ethical requisites may need to be added or changed just like other aspects of the design. Ethics by Design calls for ethics to be treated just like any other system requirement in this respect. In practical terms this requires on-going assessment of compliance with ethical requisites in just the same way as bug-fixing requires on-going assessment and mitigation processes.

It is necessary to assess whether there are other limitations which may give reason for concern. This includes ensuring the methods used to design and construct the system are capable of achieving the ethical requisites. Each of the ethical requisites imposes its own set of constraints and demands. Consideration should therefore be given to the mechanisms and systems by which these ethical requisites will be achieved during the construction process. As indicated above, the requirements for
transparency and human oversight of the development process will require additional developer resources beyond those required to support the product’s own ethical requisites.

Development phase

Because the development phase includes use of data, all issues related to data should be carefully considered here. Issues such as bias, discrimination, fairness, diversity, privacy and data quality are of particular importance when constructing and using the data set.

Ethics by Design requires that ethical requisites are kept in mind while developing the system. The Agile model puts considerable responsibility for final product functionality on relatively low-level coders and engineers. Just as small changes in code or components can have a major impact on the final features of a system, so small programming decisions may have a significant impact on conformance to the ethical requisites. Consequently, it is important that such decisions are well-considered. Agile does not place as much emphasis on documentation as some other methodologies. However, the gathering and construction of datasets, models and other components of the system need to be transparent so that they can be subject to human oversight. This is similar to tracking down bugs in code, in that ethical issues in the product as a whole often occur because more basic components of the system fail to meet their own ethical requisites. Transparency requires information repositories regarding the ethical characteristics of these components. Model Cards (Mitchell et al., 2019) and Datasheets for Datasets (Gebru et al., 2020) are both examples of tools for the programmatic generation of such information.

Testing phase

The testing phase is important as it allows developers to discover potential defects in the product. This phase will include testing for conformance to ethical requisites. In this sense, ethical requisites are to be included in the features, functionality and effects of the system. Design of testing regimes should formally include “ethical testing” to determine whether the product meets the ethical requisites. Here it is important to understand that ethical testing goes beyond the mere product itself. Many AI systems have been judged unethical not because of flaws in the product, but because of the effect they have on users or society. Often these effects are unanticipated only because no one thought to check for them. Perhaps the system works well for one racial group but not for another. In situations where indeed gender, race, etc. play a role, make sure all groups are assessed equally to avoid discrimination. The demands of transparency and human oversight dictate that the design of the testing regime will need to be documented and defensible.

Evaluation phase

The evaluation phase is focused on evaluating whether the project meets the requirements of the client. Ethics by Design requires that the ethical requisites be included in the evaluation process. It is therefore important that the client and any other stakeholders understand and support these ethical requisites. All Agile approaches should integrate a retrospective meeting (or something similar) in their evaluation phase. Such a meeting allows for deeper discussions on what went right or wrong in
the iteration. These meetings thus provide a good place and moment for an ethical discussion, and are therefore strongly recommended.
8. Annex – Organisational Adoption of Ethics by Design

This section is intended for managers of development organisations, such as software development companies, engineering companies making robots, and development departments within larger enterprises. Ethics by Design requires organisational changes in the way systems are constructed, so this section summarises important organisational changes a manager will need to introduce so that their teams can engage in Ethics by Design. This chapter may also be of use for senior management and board members who wish to understand the way in which their organisation will need to change.

Organisational Impact of Ethics by Design

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. This is achieved by changing how systems are built. While much of this involves changes in the way a system’s functionality is determined, it also requires changes in development processes and tools. Ethics by Design will not speed development, make it easier, or reduce costs. However, the demands that AI systems act ethically will only grow, and many ethical requirements are likely to become legal requirements. The move towards Ethical AI is an unavoidable, and global, trend. Those organisations which can most quickly adapt their structures to society’s demands for ethical AI will have a significant advantage over those who resist this trend.

It is important to bear in mind there is no established best practice in this area — every company will be doing this for the first time. It is therefore highly recommended that steps be taken to ensure lessons learned are documented and that procedures are modified as lessons are learned. This may mean that introducing Ethics by Design is an on-going process for some years, and that some disruption is experienced by staff. It is therefore important to maintain active and positive communication between management and developers as their work environment and culture changes around them, offering them clear advantages to supporting these changes.

The most significant changes for a development team will be:

- A large range of additional considerations during the design and development processes.
- New communication channels for ethical concerns.
- New roles.
- Additional reporting requirements.

Governance

Ethical AI requires that humans can oversee the learning, decisions, and operation of AI-driven systems. Not only does this require that developers built mechanisms into the AI product which permit this, it means organisations must put in place teams and/or individuals with oversight responsibilities. Developers could be asked to justify programming decisions to a degree of depth they have never experienced before. This will require documentation of what coding or algorithmic choices they made. It may require documenting who made these decisions and why. Resistance to explaining these matters is likely by many developers. Some will resist because they resent “outsiders” intruding into what has been their private space, while others will simply lack the skills to explain such matters.
The ethical status of a system may change as data changes, as it learns, and as usage changes. This means the organisation needs new processes for ongoing monitoring of the system’s ethical status, reporting channels for both staff and outsiders to register concerns, formal processes for assessing those concerns, and suitable mechanisms for implementing remedies. Those responsible for these processes may experience resistance from developers or others because such remedies may interfere with project schedules and will certainly generate additional costs. It is therefore essential those responsible for maintaining a system’s ethical status have sufficient authority to enforce their decisions. At minimum, this requires backing from appropriate senior management. Some companies have already started to give board members responsibility for this in order to ensure full compliance within the organisation.

Those who report concerns need to know that they have been taken seriously, considered and what decision was made. This is especially important with external stakeholders. Procedures for communication therefore need to be in place. These may have legal implications, so it is important that legal departments are positively involved in designing these process, lest they become a barrier to effective communication. Furthermore, the lessons of ethical failure and remediation need to be retained at an organisational level. This way the organisation can avoid repeating the same errors, and can learn how to implement remedial actions more effectively.

It is likely the profession of AI Ethicist will arise. There are already formal certifications for this role. The range of concerns and the skills required are wide-ranging, and so the position justifies formal training and certification. This means it may be possible to hire specialists in these governance areas. It also means it is unreasonable to expect any current staff member to take on this role without suitable training.

External Audit

It is likely many AI systems will be required to undergo auditing by external parties. This is already becoming law in some regimes for some applications. It is therefore important to understand what audit procedures a system may be subject to and ensure suitable procedures and documentation are in place to support audit. Even if audit is not currently required, it is worth preparing for the possibility new regulations will require an audit during the product’s lifetime.

If the organisation has appointed a professional AI Ethicist, preparing for, and dealing with, external audit will need to be a core competency.

Culture

Ethics by Design embeds ethics into the design process. The essence of success is to build a culture which treats ethical issues with exactly the same importance as core, undeniable, values in any system, like reliability and bug-fixing. Engineers and computer scientists have traditionally treated ethics as something which happens after they have finished their work; nothing to do with them and not the type of thing which a coder or engineer should be asked to think about. They are highly unlikely to have any training or experience in these matters and so may find even thinking about them difficult or even unpleasant. Managers cannot assume their developers and engineers will rush to embrace
ethical concerns, or that they have the skills to handle them. They will need both training and encouragement. Senior managers and board members must bear in mind the same may be true of their subordinate managers, or even themselves. Some degree of education is likely to be required of most staff at all levels. As with other cultural changes in development and engineering organisations, success depends on visible and motivational leadership from senior management.

**Finances**

Development cycles will almost certainly slow while the organisation learns how best to use Ethics by Design. It is unlikely to return to former levels because Ethics by Design requires additional tasks in every project. This will affect costs. It is well understood that all developments work within three constraints – time, money and features. Resistance from those staff primarily concerned with finances is likely in many organisations. Those seeking to introduce Ethics by Design into the organisation may therefore need to plan for this possibility and take remedial steps in advance. This is another area where support and leadership from the most senior levels is the best path to success.

**Tools**

Many of the demands of the developer under Ethics by Design require new tools. It is common when altering development processes that resistance takes the form of insisting the change is impossible due to lack of suitable tools. Where this occurs it should be investigated thoroughly by someone with the suitable technical skills. The requirements of Ethics by Design are common demands of many AI projects, so suitable tools are developing very rapidly. The pace of development in this regard is so rapid it is possible a tool could appear to fulfil a need in the time between identifying the need and the point in the development process where the need must be handled. In particular, DARPA’s Explainable AI (XAI) (Gunning, 2017) is a rapidly developing set of methodologies and tools by which build effective machine learning techniques which are also explainable to humans and allow for human governance. In these and other cases, such tools are Open Source and freely available to all.

**Final Notes**

We have highlighted here the most obvious organisational changes which adopting Ethics by Design requires. There will be many others. Since AI is a new technology, and ethical AI even newer, no one knows what all the requirements are, nor what is best practice. It is therefore important that moving to Ethics by Design is understood as an significant organisational move, not just a minor change in a

---

few development processes. One cannot expect staff who are experienced in this area, or that anyone can adapt to Ethics by Design without suitable training. In the initial stages compliance may be inconsistent, such that it requires much closer management supervision than once the organisation is used to it.

In many ways the change is similar to that experienced by organisations when web technologies emerged in the mid-1990’s. Many technical characteristics of web technology rendered old programming languages obsolete and also introduced demands on IT systems, ways of constructing software, which were completely new. Many senior developers in many organisations, with decades of experience in the old paradigms, found this extremely difficult and many first-generation web applications failed because organisations did not make the required managerial and organisational changes suited to web technologies. Where organisations possess managerial or technical staff who experienced that time, these people should be to possess many lessons which can aid the move to Ethics by Design. And while that move was difficult in the 1990’s, it is worth remembering that most organisations got there successfully in the end.