D7.2

Ethical design requirements

DW

Funded by
the European Union

Revision v1.0

Work package	WP7
Task	T7.5
Due date	30-11-2023
Submission date	28-11-2023
Deliverable lead	DW
Version	v1.0
Authors	Piercosma Bisconti Lucidi (DW)
Reviewers	Anna Rizzo (DW), Gastone Castellani (UNIBO), Enrico Giampieri (UNIBO), Nathan Lea (i-HD)

Abstract

This deliverable presents the requirements for an ethical and trustworthy design of SYNTHEMA technologies. In the first section, this deliverable outlines the regulatory challenges for AI systems in healthcare, discussing the role of regulations as both a barrier and a driver of technology innovation. Then, the seven principles for Trustworthy AI are discussed: these introduce the forthcoming AI Act. The deliverable analyses the forthcoming regulation and the risk categories and gives an overview of the standardisation activities related to the AI Act in the EU. After that, the Value-Sensitive Design methodology is presented, and its implementation in the first year of SYNTHEMA is discussed. In the final section, the discussion on the ethical framework is narrowed down to seven key ethics requirements for the development of the SYNTHEMA technologies.

Keywords

Artificial intelligence, ethics, value-sensitive design, ethics by design, trustworthy AI, synthetic data generation, requirements

Document revision history

Version	Date	Description of change	Contributor(s)
v0.1	08-11-2023	1st version	Piercosma Bisconti (DW)
v0.2	10-11-2023	2nd version after first review	Anna Rizzo (DW)
v0.3	17-11-2023	First internal review by UNIBO and i-HD	Enrico Giampieri (UNIBO), Nathan Lea (i-HD)
v0.4	27-11-2023	Reviewed version, taking into account reviewers' comments	Piercosma Bisconti (DW), Anna Rizzo (DW)
v1.0	28-11-2023	Final version, after final review and proofreading	Piercosma Bisconti (DW), Anna Rizzo (DW)

Disclaimer

The information, documentation and figures available in this deliverable are provided by the SYNTHEMA project's consortium under EC grant agreement **101095530** and do not necessarily reflect the views of the European Commission. The European Commission is not liable for any use that may be made of the information contained herein.

Copyright notice

Document information

Nature of the deliverable		**R**
Dissemination level		
PU	Public, fully open. e.g., website	✓
CL	Classified information as referred to in Commission Decision 2001/844/EC	
SEN	Confidential to SYNTHEMA project and Commission Services	

***** Deliverable types:
R: document, report (excluding periodic and final reports).
DEM: demonstrator, pilot, prototype, plan designs.
DEC: websites, patent filings, press and media actions, videos, etc.
OTHER: software, technical diagrams, etc.

Table of contents

List of tables

Acronyms and definitions

AI	Artificial intelligence
CEN	European Committee for Standardization
CENELEC	European Committee for Electrotechnical Standardization
DoA	Description of action
EHR	Electronic health record
EC	European Commission
ETSI	European Telecommunications Standards Institute
EU	European Union
GDPR	General Data Protection Regulation
GAN	Generative adversarial network
GPAI	General Purpose AI
LLM	Large Language Model
AIHLEG	High-Level Expert Group on Artificial Intelligence
LLM	Large language model
MDSS	Medical decision support systems
SME	Small and medium-sized enterprise
SDG	Synthetic data generation
VSD	Value Sensitive Design
WP	Work package

Funded by
the European Union

1 Executive summary

This deliverable highlights seven pivotal ethics requirements, stemming from the integration of diverse AI ethics and governance approaches. These requirements are deeply entrenched in the current standardisation efforts of the EU CEN-CENELEC, providing a robust framework for practical implementation. The deliverable details guidelines for each ethical requirement, ensuring applicability and relevance. The seven requirements are summarised below.

1. **SDG standard requirements**: the project emphasizes adherence to upcoming standards from ISO and CEN-CENELEC related to AI, specifically focusing on generative AIs and SDG. A notable standard is the ISO/IEC AWI TR 42103, which provides an overview of synthetic data in AI systems. SYNTHEMA commits to incorporating this standard upon its availability. It is recommended that experts from the SYNTHEMA consortium participate in ISO national bodies to gain early access to the standard, contribute to its development, and align the project accordingly.

2. **Fairness**: to combat biases and ensure demographic representation, SYNTHEMA stresses training models to accurately represent disease incidence. This is crucial for SDG, as biases could be magnified in synthetic datasets. The project will apply at least the ISO/IEC TR 24027, addressing bias in AI systems and decision-making, and will document its application. Participation in CEN-CENELEC national bodies is encouraged for early standard access and contribution.

3. **Oversight and acceptance**: focus group discussions highlighted the impact of AI system design on physician trustworthiness. The project recognizes the need for the technology to be both technically and perceptually trustworthy for effective adoption and use. This involves informing physicians about the platform limitations and functionality, and testing UX/UI aspects for social acceptance using scientific measures like TAM and UTAUT. The model's interpretability and its influence on physician trustworthiness are also key areas of investigation.

4. **Building patient trust**: VSD sessions revealed concerns about patient mistrust towards SDG, potentially affecting therapy adherence. To address this, at least 10 patients will be interviewed to understand their concerns and perspectives, which will be discussed by clinical and technical partners. Quantitative assessment of technology acceptance and trustworthiness will be utilised.

5. **Explainability**: SYNTHEMA places a strong emphasis on explainability as a fundamental aspect to ensure the practicability of other requirements. It is crucial that the decisions

made by the system are understandable by physicians, which is key to maintaining transparency and accountability. To achieve this, the project commits to ensuring that the contribution of each feature to the model output is transparent and interpretable.

6. **Accountability**: accountability in AI is vital for maintaining trustworthiness and managing liability in scenarios involving errors or malfunctions. SYNTHEMA aims to establish a governance process to assess the accountability of any potential errors or malfunctions in the platform. This process will be defined and clarified before a specified milestone (M42) and will consider the possible integration of the SYNTHEMA platform with other technologies. Adhering to relevant standards on accountability and governance is crucial for the project's success and credibility.

7. **Value-sensitive design**: central to the development of these requirements is the conceptual analysis of SYNTHEMA ethical and social implications. This analysis is intricately coupled with the implementation of VSD. VSD stands out as a qualitative methodology focused on capturing the expectations and concerns of various stakeholders regarding technology. By prioritizing these values, it aims to harmonise the design process with stakeholder interests, thereby enhancing technology acceptance and ethical alignment.

SYNTHEMA has employed a collaborative and co-creative approach to establish its ethics requirements. This approach involved organising numerous sessions and conducting two focus groups. These interactive platforms enabled a comprehensive gathering of insights and perspectives from a diverse range of participants. The focus groups, in particular, played a crucial role in grounding the ethics requirements within a co-creative framework. This methodology not only fostered a participative environment but also ensured that the requirements were reflective of a wide array of stakeholder values and concerns.

Furthermore, the project engagement with the ongoing EU CEN-CENELEC standardisation activities signifies its commitment to aligning with broader regulatory and ethical frameworks. This alignment is critical in ensuring that the SYNTHEMA outcomes are both relevant and compliant with prevailing standards and practices in AI ethics and governance. The inclusion of practical implementation guidelines further demonstrates the project dedication to translating ethical principles into actionable strategies.

In summary, the deliverable offers a comprehensive and actionable set of ethics requirements. These requirements are the culmination of a methodical and inclusive process that combines conceptual analysis, VSD, and co-creative methodologies. The project alignment with EU CEN-CENELEC standardisation activities and its focus on practical implementation guidelines underscore its commitment to establishing a robust and ethically sound framework for AI development and application.

2 The regulatory challenges for an effective implementation of AI systems in the EU healthcare sector

In November 2021 the *European Union* (EU) Commission's DG Connect released a report that details the implementation status of *artificial intelligence* (AI) systems in the EU healthcare sector. The report is titled *"Study on eHealth, Interoperability of Health Data and Artificial Intelligence for Health and Care in the European Union"*. This report outlines the current trends and challenges for an effective use of AI systems in this domain.

A noteworthy trend in this discourse is the emphasis on the centrality of the patient. AI technologies, exemplified by tools like chatbots, are increasingly being designed and discussed in the context of their ability to cater to individual patient needs. The narrative often gravitates towards the latest innovations and clinical trials, underscoring the potential benefits for patients. However, it is interesting to note that the direct impact of these innovations on health outcomes is not always the focal point of these discussions.

The year 2020, marked by the onslaught of the Covid-19 pandemic, brought about a significant shift in this narrative. Between March and June, there was a noticeable uptick in mentions of AI in the context of the pandemic. The discussions revolved around how AI could be harnessed to detect early symptoms, track the spread of the virus, and even predict outbreaks.

Beyond the pandemic, the potential of AI in disease detection has been a recurring theme, with a particular emphasis on its applications in oncology. The exploration of AI-powered tools for cancer screening has generated considerable interest, pointing to a future where early detection could significantly improve patient outcomes.

Lastly, the broader medical community is becoming increasingly cognizant of the transformative role AI can play across various medical disciplines. From radiology and cardiac imaging to the precision of surgical interventions, the promise of AI is vast. On the other hand, the report also outlines some challenges to an effective implementation. For example, one outstanding challenge for an AI-driven healthcare sector is the analysis and treatment of biases on the dataset, and the likely reproduction of these biases in generated dataset, when adopting data augmentation techniques. On the more governance side, among others, it is interesting for the sake of this chapter to highlight the following two:

1. **Imprecise legislation around AI**. The dynamic nature of AI and its rapid evolution poses legislative challenges. Current regulations may not fully address the complexities and nuances of AI applications in healthcare. The absence of clear and precise legislation can lead to ambiguities, potentially stifling innovation and deterring investments due to uncertainties around compliance and liabilities.
2. **Balancing technical solutions with data protection**. AI in healthcare often requires vast datasets to function optimally. However, the EU places a strong emphasis on personal

data protection, especially with regulations like the *General Data Protection Regulation* (GDPR). Striking a balance between leveraging AI's potential and ensuring robust data protection is a persistent challenge.

The ethical, regulatory, and privacy considerations remain significant challenges and constitute substantial barriers to innovation and the successful integration of AI systems within the healthcare sector. However, the report underscores that the primary obstacle is not the regulation per se, but the ambiguity surrounding it. The aim of this document is to outline essential ethical guidelines and recommendations to facilitate compliance with existing frameworks. Prior to delving into these guidelines and recommendations, we explore foundational concepts that form the theoretical frameworks from which these requirements are derived.

3 The ethical challenges of SDG

3.1 Introduction

The field of AI has given rise to novel methods of generating synthetic data, which are artificially-produced data points that emulate the statistical properties of real-world datasets. In the health domain, the generation of synthetic data is a particularly active area of research due to the potential for such data to enable advances in medical research while circumventing some of the privacy concerns associated with using actual patient data. However, *synthetic data generation* (SDG) is not without its ethical and privacy challenges. This chapter provides a comprehensive review of the recent literature addressing these challenges, with a particular focus on the balance between data utility and the protection of individual privacy.

3.2 Privacy and data protection concerns

Privacy and data protection are paramount concerns in the health sector due to the sensitive nature of personal health information. The generation of synthetic health data poses a dual-edged sword; while it can be used to protect patient confidentiality, it can also entail risks of re-identification if not managed appropriately. Rocher et al. (2019)[1] demonstrated that even de-identified data could often be re-identified using machine learning techniques combined with auxiliary information. The implications for synthetic data are significant, as the potential for re-identification must be considered when generating and sharing such data.

3.3 Ethical considerations

Generating synthetic data for health research also raises ethical questions, particularly in terms of consent and the authenticity of data. The work of Zliobaite and Custers (2016)[2] highlighted the ethical dilemma of using patient data without explicit consent for the generation of synthetic datasets. While synthetic data itself does not contain real patient details, the source data used to create it often does, raising concerns about the need for informed consent.

3.3.1 Quality and trustworthiness of synthetic data

The quality and trustworthiness of synthetic data is another ethical concern. The work of Bellovin et al. (2019)[3] scrutinized the accuracy of synthetic data and how discrepancies could lead to incorrect medical conclusions or biased research outcomes. The need for high-quality,

[1] Rocher, L., Hendrickx, J.M. and De Montjoye, Y.A., 2019. Estimating the success of re-identifications in incomplete datasets using generative models. Nature communications, 10(1), pp.1-9.

[2] Žliobaitė, I. and Custers, B., 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. Artificial Intelligence and Law, 24, pp.183-201.

[3] Bellovin, S.M., Dutta, P.K. and Reitinger, N., 2019. Privacy and synthetic datasets. Stan. Tech. L. Rev., 22, p.1.

representative synthetic data is essential to ensure that it serves as a reliable substitute for real data in health research.

Complementing this perspective, Jordon et al. (2018)[4] introduced differential privacy techniques within the data synthesis process to generate datasets that maintain privacy while still being statistically similar to the original. Their research demonstrates the utility of such differentially private synthetic datasets in preserving individual privacy without substantial loss in data utility.

Extending the discussion to the representativeness of synthetic data, Xu et al. (2019)[5] explored the risks of bias in synthetic data generation. Their work highlights the importance of representativeness in synthetic datasets, emphasizing that biases in the training data can lead to skewed models which, in turn, produce biased synthetic data. They propose methods to detect and correct for such biases during the data generation process.

Ensuring that synthetic data can be trusted for health research applications requires careful consideration of the data generation methodologies. Choi et al. (2017)[6] showcased the capabilities of *generative adversarial networks* (GANs) in creating realistic and high-fidelity synthetic *electronic health records* (EHRs). Their work underscores the potential of advanced deep learning techniques in crafting synthetic datasets that mirror the complexity of real health data.

Lastly, Torfi et al. (2020)[7] contribute to the ongoing discussion with a focus on the practical applications of synthetic data. They investigate the use of synthetic data in facilitating healthcare predictive analytics, outlining how carefully crafted synthetic datasets can be used to train predictive models without compromising patient privacy, thus maintaining the balance between data utility and privacy.

3.4 Regulatory perspectives

Regulatory perspectives on synthetic data are evolving, as discussed by Jordon et al. (2018)[8], who examined the legal frameworks applicable to synthetic data in healthcare, particularly the GDPR in Europe. They argued that the GDPR's provisions on data protection by design could provide a guiding framework for the generation of synthetic health data yet noted that specific guidance for this novel data type is lacking.

[4] Jordon, J., Yoon, J. and Van Der Schaar, M., 2018, September. PATE-GAN: Generating synthetic data with differential privacy guarantees. In International conference on learning representations.

[5] Xu, L., Skoularidou, M., Cuesta-Infante, A. and Veeramachaneni, K., 2019. Modeling tabular data using conditional gan. Advances in neural information processing systems, 32.

[6] Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W.F. and Sun, J., 2017, November. Generating multi-label discrete patient records using generative adversarial networks. In Machine learning for healthcare conference (pp. 286-305). PMLR.

[7] Torfi, A., Fox, E. A., & Guha, R. K. (2020). Generating Synthetic Healthcare Records using Generative Adversarial Networks. Journal of the American Medical Informatics Association, 27(6), 856–864.

[8] Ibid., 4.

3.5 Balancing utility and privacy

The balance between data utility and privacy protection is a recurring theme in the literature on synthetic health data. Beam and Kohane (2018)[9] explored the trade-offs between data utility for research and the risk of individual re-identification. They noted that techniques such as differential privacy offer promising methods for creating synthetic data that is both useful and preserves privacy, but the choice of privacy parameters is critical in determining the usability of the data.

The generation of synthetic data in the health domain offers significant potential benefits for medical research and public health, but it is accompanied by complex ethical and privacy issues that must be navigated with care. Future research should continue to develop methods to optimize the balance between the utility of synthetic data and the privacy of individuals, with an emphasis on the quality and trustworthiness of the generated data. Furthermore, the development of regulatory guidance specific to synthetic data is essential to guide its ethical use in health research.

[9] Beam, A.L. and Kohane, I.S., 2018. Big data and machine learning in health care. Jama, 319(13), pp.1317-1318.

4 The ethics frameworks for AI in the EU

The rapid evolution of AI has brought forth opportunities and challenges. Recognising the transformative potential of AI and the need for a robust ethical framework, the *European Commission* (EC) took a proactive step by introducing the "Ethics Guidelines for Trustworthy AI." This document, crafted by the *High-Level Expert Group on Artificial Intelligence* (AI HLEG), serves as a beacon, guiding the development and deployment of AI in the EU towards a path that is not only innovative but also ethically sound.

At the heart of the guidelines is the concept of "trustworthy AI." The EC envisions an AI ecosystem where systems are not just advanced but also reliable, ethical, and aligned with human values. Trustworthy AI is characterised by three foundational pillars:

1. **Lawful**: this emphasises the importance of AI systems adhering to existing laws and regulations. Given the dynamic nature of technology and the evolving legal landscape, ensuring that AI systems remain compliant is crucial. This lawful adherence ensures that AI developers and users operate within the boundaries set by regulatory authorities, minimising legal risks and fostering public trust.
2. **Ethical**: beyond legal compliance, AI should resonate with moral principles and values. This means that AI systems should be designed and deployed in ways that uphold human dignity, respect human rights, and promote societal well-being. Ethical considerations ensure that AI serves humanity and does not inadvertently harm or marginalise individuals or groups.
3. **Robust**: trustworthy AI should be technically sound and should operate reliably in diverse conditions. This robustness extends to ensuring that AI systems are resilient against both intentional malicious attacks and inadvertent errors. Furthermore, the social robustness of AI, which pertains to its broader impact on society, is equally vital.

4.1 The seven key requirements for trustworthy AI

To operationalise the concept of trustworthy AI, the guidelines delve deeper, outlining seven essential requirements that AI systems should strive to meet:

1. **Human agency and oversight**: central to the European vision of AI is the idea that technology should augment human capabilities, not undermine them. AI should be a tool that empowers individuals, enhancing their decision-making without eroding their autonomy. Effective human oversight mechanisms should be in place to ensure that AI's actions align with human intentions.
2. **Technical robustness and safety**: as AI systems become more integrated into critical sectors like healthcare, transportation, and finance, their technical reliability becomes paramount. Systems should be designed to handle uncertainties, operate securely, and be resilient against both external attacks and internal system failures.

3. **Privacy and data governance**: In the digital age, data is a valuable asset. AI systems, which are inherently data-driven, should prioritise the protection of personal data. Robust data governance mechanisms should ensure that data is acquired, stored, and processed in ways that respect individual privacy and comply with data protection regulations.
4. **Transparency**: For users to trust AI, they need to understand it. The processes, algorithms, and decision-making mechanisms of AI systems should be transparent. This transparency ensures that users, regulators, and the broader public can understand and trust the actions and decisions made by AI.
5. **Diversity, non-discrimination, and fairness**: AI should be for everyone. This means that systems should be designed to be inclusive, catering to diverse user groups. Moreover, AI algorithms should be free from biases, ensuring that decisions made are fair and do not discriminate against any individual or group.
6. **Environmental and societal well-being**: The broader impact of AI on society and the environment cannot be ignored. AI systems should be sustainable, minimising their environmental footprint. Additionally, the societal implications of AI, from its impact on employment to its role in shaping public discourse, should be considered, ensuring that AI contributes positively to societal progress.
7. **Accountability**: As AI systems wield increasing influence over various aspects of society, mechanisms should be in place to hold developers, users, and other stakeholders accountable for the outcomes. This accountability ensures that in cases of errors, biases, or other issues, there are clear avenues for redress and corrective action.

The EC's *"Ethics Guidelines for Trustworthy AI"* is a testament to the EU's commitment to ensuring that AI serves humanity. By emphasising the principles of lawfulness, ethics, and robustness, and by outlining clear requirements for AI systems, the guidelines provide a roadmap for developers, policymakers, and users. As AI continues to shape the future, these guidelines ensure that this future is not only technologically advanced but also ethically sound and human-centric.

5 From Ethics Guidelines for Trustworthy AI to the EU AI Act

The *Ethics Guidelines for Trustworthy AI* serve as the foundational basis upon which the subsequent regulation on AI has been developed. In the subsequent sections, we will examine the pivotal concepts of the AI Act as it stands in November 2023.

In the following chapter, the reader will find an analysis of the AI Act. For what concerns the AI Act applications to SYNTHEMA, we remind that research and innovation has special derogations in the application of the forthcoming regulation. However, any exploitation of the SYNTHEMA technologies will fall under the EU AI Act, so to take into account the regulation's requirements would be exploitation-wise.

Since the AI Act is not in its final form at this moment, we cannot give a definitive answer on the risk category of SYNTHEMA technologies, when they will be exploited. However, since SDG impact directly the fundamental rights, namely health, it is safe to assume that SDG technologies will fall under the high-risk category, and therefore will need special risk and impact assessment and mitigation.

5.1 Introduction

The acceleration of technological development and especially the advent of AI systems have profoundly revolutionized industry and society in recent years.

With a view to unifying the single market and protecting fundamental rights, within the communication 'Artificial Intelligence for Europe', the EC drafted a strategy on AI in 2018. This strategy is in turn part of the more ambitious and broader European goal of 'Digital Europe', the action plan drafted by the Commission to make the digitisation process a priority at European level while achieving the goal of making the EU 'climate neutral by 2050'. Jointly with a number of member states, the EC presented the 'Coordinated Plan on Artificial Intelligence' in December of the same year, with the aim of 'promoting and strengthening development, research and investment in AI systems', primarily encouraging member states to develop and refine national strategies. To monitor the progress of the above-mentioned plan, the EC has also devised the 'AI WATCH' project, a tool that tracks the technological, research and industrial capacity of individual Member States, as well as their national policies. The project also offers a detailed map of the latest updates and technological developments in the field of AI systems worldwide. The first report on national strategies was published in February 2020.

The EC is also the author of the *'2021 Review of the Coordinated Plan on Artificial Intelligence'*, a document aimed at 'analysing and reporting on the financial initiatives undertaken by the Commission in the 2021-2027 multi-year plan'. This plan envisages funding of EUR 1 billion per year from 2020 onwards to be invested in projects:

- Horizon Europe, the European programme promoting innovation and research between 2021 and 2027;
- Digital Europe, the European funding programme that aims to bring digital services to the attention of public and private administrations and citizens.

The most recent and most important step taken by the European Union was the proposed EU Regulation on Artificial Intelligence. The *EU Regulation on Artificial Intelligence* (also the 'AI Act' or the 'Regulation'), proposed by the EU on 21 April 2021, is the most comprehensive and multi-layered piece of legislation in the world of AI regulation to date. This very wide-ranging and detailed piece of legislation aims to establish a uniform EU system of rules for the regulation of AI, encompassing all sectors except the strictly military field. It is important to emphasise how this regulation confers minimum enforceable rights on individuals subject to the actions of AI systems, while at the same time focusing on the regulation of the economic actors who in various capacities supply AI systems and employ them professionally.

At the time of writing, the AI Act has not yet been approved by the EU institutions and is in the so-called trialogue phase between the EU Parliament, the EU Commission and the EU Council (each of these three institutions has proposed its own version of the AI Act and now has to converge on a shared text, in order to be able to finally approve it). Consequently, although the general approach and structure of the regulation will not undergo substantial changes, a whole series of ancillary elements, however important, may be subject to changes, even significant ones. For example, as regards the definitions, the duration of the so-called grace period between final approval and effectiveness.

The scope of the AI Act is extensive, covering almost every type of AI system. In this way, the EU aims to address the multiplicity of applications of AI in today's rapidly changing technological landscape. This is crucial because AI is no longer limited to a specific sector or field; its influence is pervasive, touching virtually every aspect of our lives.

One of the key principles underlying the AI Act is the classification of AI applications according to their potential to cause harm. This classification framework is designed to ensure that regulation adequately matches the risks posed by different AI systems. It categorises AI applications into three main groups: (1) prohibited practices, (2) high-risk systems and (3) other AI systems.

1. **Prohibited practices** constitute the highest risk category, including AI applications that are inherently harmful or violate fundamental rights. Such practices are unequivocally prohibited, and the AI Act establishes strict measures to prevent their development, implementation or use. This includes AI systems that could be used for mass surveillance without consent or those that perpetuate discrimination and prejudice.

2. **High-risk systems** constitute the second category. These AI applications are considered to carry significant risks, such as potential harm to individuals, public safety or fundamental rights. They include medical AI systems, autonomous vehicles and other

advanced technologies where interests are high. The AI Act requires rigorous compliance assessments and safeguards for these high-risk systems. Developers and users are required to comply with strict rules to mitigate the associated risks.

3. **Other AI systems** include AI applications that do not fall into the categories of prohibited or high-risk practices. These systems are not subject to the same level of regulation, but must still comply with transparency and documentation requirements. This ensures that AI systems that are not considered high-risk are also developed and used in a responsible manner.

In summary, the Regulation represents an important and complex attempt to regulate the AI landscape. By classifying AI applications according to their potential to cause harm, it adopts a graduated approach to regulation. Indeed, on the one hand it prohibits inherently harmful practices and imposes strict controls on high-risk systems, while on the other hand it recognises the multiple applications of AI and seeks to strike a balance between innovation and security. This legislation represents a significant step in addressing the complex challenges posed by AI in our society and is likely to serve as a model for the regulation of AI in other parts of the world.
In any case, it is essential to wait for the final and definitive version of the Regulation in order to have access to a complete and comprehensive regulatory framework, especially with regard to the relevant definitions, the actors involved and the main obligations.

5.2 The actors of the AI Act

Among the many important aspects to consider when conducting a preliminary analysis of a set of compliance obligations is certainly the identification of roles. Below are the main roles, in the version of the proposed Regulation approved by the EU Parliament (2023).

Supplier
Supplier means any natural or legal person, public authority, agency or other body that develops an AI system or has an AI system developed with a view to placing it on the market or putting it into service under its own name or brand, whether for payment or free of charge.

User
User means any natural or legal person, public authority, agency or other body that uses an IA system under its authority, unless the IA system is used in the course of a non-professional personal activity.

Importer

Importer means any natural or legal person established in the Union who places on the market or puts into service an IA system bearing the name or trademark of a natural or legal person established outside the Union.

Distributor

Distributor means any natural or legal person in the supply chain, other than the supplier or importer, who makes an IA system available on the Union market without affecting its properties.

5.3 The Commission version (2021)

5.3.1.1 The risk-based classification system

A central aspect of the AI Act is its risk-based classification of AI systems. This classification assigns transparency obligations and compliance requirements to the various AI systems according to the level of risk posed by their potential use. According to Article 5 of the regulation, AI systems are classified into four distinct categories:

- 'Prohibited AI practices';
- 'High-risk AI systems';
- 'Limited risk AI systems';
- 'Minimal risk AI systems.

1. **Prohibited practices**: to determine the 'prohibited' character of a practice, the legislator does not consider specific AI systems but measures the degree of threat to EU values and in particular with respect to fundamental rights caused by the use of such systems. These practices are divided into four distinct categories: manipulation, exploitation of vulnerable groups, public social scoring and remote and real-time biometric identification. For these practices, Article 5 imposes a general prohibition.

2. **High risk**: the AI Act identifies two macro categories of high-risk practices. The first qualifies as high risk those AI systems that: 'create a high risk to the health and safety or fundamental rights of natural persons', while the second defines the scope by referring to a detailed list governed by Annex III.

3. **Limited risk**: within the category of limited risk fall three distinct AI systems covered by Art. 52. These are primarily systems 'intended to interact with natural persons', i.e. all chatbots and other personified AI systems. The second type of AI systems includes instead 'emotion recognition and biometric categorisation systems'. Such systems 'interpret biometric data and make assessments of individuals but, unlike the systems referred to in

Article 5(d) and Annex III, no. 1, are not aimed at identifying the natural person'. The last category instead regulates the practice and use of so-called deepfakes, i.e., fake audio, image, or video content created through generative intelligence systems capable of reproducing or simulating real places, objects, or persons.

In the case of medium- or low-risk AI systems, on the other hand, there is a requirement to 'make it known to the user that they are interacting with an AI system'. The addressees of these provisions are all 'natural persons' who in fact interact with the system and in particular 'end users'. For all these types, there is an exception to the obligation of transparency in the event that the systems are used for the investigation and detection of crimes.

4. **Minimal risk**: the last category, on the other hand, includes all those AI systems 'whose risk is considered minimal (for the security and rights and freedoms of citizens and which are to be developed and used in compliance with existing legislation without further legal obligations (such as voluntary adherence to Codes of Conduct)'. As was pointed out on the eve of the approval of the draft regulation, most AI systems in circulation in the year 2021 were probably to be considered as belonging to the category of minimal risk (consider predictive maintenance systems, video games with built-in AI systems, and telephone filters against spam).

5.3.1.2 The economic sanctions system

Since its first version, the draft regulation has provided for an accurate and detailed system of sanctions. This system was expressly regulated in Article 71, where the member states were entrusted with the burden of providing 'effective and dissuasive sanctions'. The sanctions provided for in that version of the Regulation, which have now undergone even more severe modifications, were as follows:

- "up to EUR 30 million or 6 % of the previous year's total annual worldwide turnover (whichever is higher) for violations related to prohibited practices or non-compliance with data requirements";
- "up to EUR 20 million or 4% of the total annual worldwide turnover in the preceding financial year for non-compliance with any other requirement or obligation of the regulation";
- "up to EUR 10 million or 2% of the total annual worldwide turnover of the previous year for supplying incorrect, incomplete or misleading information to notified bodies and national competent authorities in response to a request".

5.3.2 The European AI or Artificial Intelligence Board

In order to ensure a homogeneous and compliant application of the AI Act within the Member States and with a view to promoting cooperation between national supervisory authorities, the AI Act has expressly provided in Articles 56 and 57 for the establishment and organisation of a European Committee. This Committee will have an advisory and assistance function. The main tasks set out in Article 58 include the formulation of 'opinions, recommendations, advice or guidelines on matters relating to the implementation of the Regulation, including existing technical specifications or standards relating to the requirements laid down therein, as well as providing advice and assistance to the Commission on specific matters relating to AI'. The Regulation also expressly regulates the task given to each Member State to designate a national authority for the "supervision and enforcement (of the Regulation), as well as market surveillance activities". Finally, the Regulation provides for the drafting and adoption of national codes of conduct in order to ensure widespread and voluntary application of the Regulation's requirements, as well as to guarantee the measurement and verification of such application on the basis of pre-established indicators and determined objectives.

5.3.3 The Council version (2022)

On 25 November 2022, the Council adopted its own proposal for a regulation, building on the Commission's text of 2021.

This proposal acknowledged the complexity of the value chain through which artificial intelligence systems are developed and deployed. In this regard, the Council was primarily concerned to clarify the responsibilities attributed to the actors in this process. The Council also clarified these responsibilities by emphasising points of regulatory intersection with other European provisions such as Regulation (EU) 2016/679 (the 'GDPR').

In this version, important changes have been made to the definition of AI systems and risk categories.

5.3.3.1 Definition of AI systems

In order to distinguish the definition of AI systems from common, simpler software systems, the Council proposed a definition that takes into account 'systems developed through machine learning and logic and knowledge-based approaches'.

5.3.3.1.1 Banned AI practices

Within the category of prohibited practices, the Council proposal included AI systems aimed at social scoring, used by private actors. Artificial intelligence systems that 'exploit the vulnerabilities of a specific group of persons, including persons vulnerable because of their social or economic situation' are also prohibited.

As regards remote biometric identification systems carried out 'in real time' in publicly accessible spaces, on the other hand, the text of the proposal specifies that such use must be 'strictly necessary for law enforcement purposes and for which law enforcement authorities should exceptionally be authorised to use such systems'.

5.3.3.2 High-risk AI systems

A significant change proposed by the Council concerns the provision of requirements for high-risk AI systems. A change was also made concerning the structure of the classification of high-risk AI systems.

5.3.3.3 General-purpose AI systems

Dedicated provisions have been added for 'General-Purpose AI' systems (i.e. with no predetermined purpose), which regulate the cases in which AI systems are integrated into other high-risk intelligence systems, with the aim of preventing this type of system from falling outside the scope of the Regulation

5.3.4 The scope of the AI Act: openings and restrictions

With regard to the scope of application of the AI Act, the military, defence and national security spheres were explicitly excluded. On the other hand, the Council considered it appropriate to broaden and protect the scope of those who make use of AI systems for research purposes. To this end, an exception was made for the application of the provisions of the AI Act in the fields of research and development, with the exception of minimum transparency requirements. With regard to the use of AI systems in law enforcement activities, then, specific provisions have been made to protect the 'confidentiality of data sensitive to their activities'.

5.3.5 The Parliament version (2023)

On 14 June 2023, the EU Parliament approved its own version of the Regulation, which on the one hand holds firm to the structure already proposed by the Commission and the Council, and on the other hand comes at a very different time in history (post release of services such as ChatGPT on the market). For the latter reason, the legislator of the EU Parliament has made a number of important changes and amendments, both in terms of definition and interpretation of the risk categories, as well as numerous other minor changes.

5.3.5.1 A more technologically neutral definition

When drafting the amendments, one of the recurring themes for Parliament was that of guaranteeing citizens and businesses forms of protection that are transparent, traceable, non-discriminatory and respectful of the development of AI systems. The definition of AI systems constitutes a crucial aspect for regulatory clarity and relevance, since it first defines the scope of application. This definition probably constitutes the single most crucial point within the AI Act legislation, since attributing a specific meaning to such systems also implies, in the strictly technical and legal sphere, defining the perimeter of application. In this regard, it should be emphasised that a more neutral and all-encompassing definition has been adopted than the original one proposed in 2021. The 2021 definition qualified an AI system as 'software developed using one or more of the techniques and approaches listed in Annex I, which can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations or decisions that influence the environments with which it interacts'. The one adopted by the Parliament is the following: 'a "machine based" system designed to operate with different levels of autonomy and which can, for explicit or implicit objectives, generate outputs such as predictions, recommendations or decisions, which influence physical activities or virtual environments'. One of the most important reasons for this change is the need to align the definition profiles at the international level, in particular with the one drawn up by the OECD. In fact, a definition closely aligned with the work of international organisations dealing with AI is better able to guarantee legal certainty, harmonisation of standards and their wider acceptance.

5.3.5.2 Expansion of risk categories

In its amendments, the EU Parliament added the following systems to the category of prohibited practices:

- "Remote biometric identification systems carried out 'in real time' in publicly accessible spaces";
- "Ex post" biometric remote identification systems, with the exception of use by law enforcement agencies for the prosecution of serious crimes and subject to judicial authorisation;
- 'Biometric categorisation systems that make use of sensitive data (e.g. gender, race, ethnicity, citizenship, religion, political orientation)';
- 'Predictive policing systems (based on profiling, location or past criminal behaviour)';
- "Emotion recognition systems carried out in law enforcement, border management, workplaces and educational institutions".
- 'The indiscriminate collection of biometric data from social media or CCTV to create facial recognition databases'.

In addition, the conditions for an AI system to be considered high risk have been slightly reshaped in the parliamentary version. In fact, they are to be considered as such:

1. AI systems used in products covered by the General Product Safety Regulation 2022/2065. Medical devices, lifts, cars, toys, etc. fall into this category;
2. Artificial intelligence systems fall into eight specific categories:

2.1 'biometric identification and categorisation systems for natural persons',

2.2 "systems concerning the management and operation of critical infrastructures",

2.3 "systems concerning vocational education and training",

2.4 "systems concerning employment, management of workers and access to self-employment",
2.5 "systems concerning access to and use of essential private services and public services",
2.6 "systems concerning law enforcement",
2.7 "systems concerning the management and policies of asylum, border control and immigration in general",
2.8 "systems concerning assistance in legal translation and law enforcement".

In the latter case, an AI system is to be considered a high risk only where it poses a significant risk of harm to the health, safety or fundamental rights of natural persons.

It is recalled that all AI systems considered to be 'high-risk' will have to undergo conformity assessment before they are placed on the market and throughout their life cycle. With regard to the category of low-risk AI systems, transparency requirements have been included to enable users to be informed about the impact of use of such systems. Under the new provisions, the user can decide whether or not to continue using the application after having already interacted with it. Users will have the right to be informed when interacting with AI systems, including generative AI systems that create new audio or video content or manipulate existing content, such as so-called deepfakes.

5.3.5.3 Generative AI and foundational models

During Parliament's revision of the text, changes were made to the definition of so-called foundational models, General Purpose AI and generative AI systems. In fact, the strong media and social impact of some generative AI systems based on *Large Language Models* (LLM) such as ChatGPT led the European legislators to develop a dedicated category. In the negotiating draft, the parliamentary committees proposed a regulatory organisation based on three distinct terms: 'General Purpose AI', 'Foundation models' and 'Generative AI'. By Foundation models is meant an AI model trained on large amounts of large-scale data, designed for a generality of results and which can be adapted to a wide range of distinct tasks. At the same time, General Purpose AI means AI systems that can be used and adapted to a wide range of distinct tasks.

The category of General Purpose AI is to be seen as broader than both generative intelligence systems and foundational models. In fact, if generative intelligence systems represent a subset of foundational models, the same relationship does not apply between foundational systems and General Purpose AI. In this case, some foundational models belong to the General Purpose AI category, while others do not. The result worked out by the two commissions was a multi-level approach. The commissions thus envisaged that most of the obligations should fall on the economic operators that will integrate these systems in an application considered to be high-risk, thus excluding a predetermined classification of *General Purpose AI* (GPAI) systems. Suppliers of

GPAI systems are, however, instructed to support the compliance of downstream operators by providing all relevant information and documentation on the AI model used.
Generative AIs, on the other hand, are a practical application of foundational models and are therefore to be regarded as subsets of foundational models. With regard to the obligations relating to generative intelligence systems, Parliament has added certain information and transparency obligations. Additional obligations concern disclosing that the content was generated by AI and which data sources were used to train the algorithm, providing the model's own characteristics including potential risks of use and possible measures to mitigate negative effects or, conversely, the reasons why it is considered that mitigation is not possible. Furthermore, these models must be designed to avoid the creation and generation of illegal content. Finally, mandatory publication of the summary of data used for training purposes protected by copyright is required.
The parliamentary committees also provided for an obligation on the providers of such systems to provide data governance measures, to apply security controls and risk mitigation before placing the systems in question on the market, together with the obligation to consider foreseeable risks to health, security, fundamental rights, the environment, democracy and the rule of law. The amendments made by the committees then require manufacturers of generative models to reduce the energy consumption and resource utilisation of their systems and to register the systems in an EU database, to be set up. Suppliers of generative AI systems, on the other hand, are called upon to comply with the transparency obligations set out in the regulation (ensuring that users are informed that the content has been generated by the machine); to apply 'appropriate safeguards' in relation to the content generated by their systems; and to provide a summary of any copyright-protected material used to train their AI.

5.3.5.4 Derogation schemes for innovation

In order to ensure a proper balance between protection and innovation, the EU Parliament has included some exceptions to protect innovation for 'research purposes and for AI components provided under open-source licences'. The new amendments also promote 'regulatory sandboxes, or controlled environments, set up by public authorities to test AI before its deployment'. According to the legislator, the purpose of regulatory sandboxes should in fact be to promote safe AI innovation by creating a controlled environment for experimentation and testing in the development and pre-commercialisation phase in order to ensure the compliance of innovative AI systems with the AI Act and the enhancement of legal certainty for innovators and the oversight and understanding by competent authorities of opportunities, emerging risks and impacts of AI use, as well as accelerating market access, including by removing barriers for *small and medium-sized enterprises* (SMEs) and start-ups.

5.3.5.5 New principles

Six general principles applicable to all AI systems were also introduced in order to create technological systems that respect and comply with fundamental human rights. These principles are:

- Surveillance and supervision of the human being;
- Technical solidity and security;
- Privacy and data governance;
- Transparency;
- Diversity, non-discrimination and equity;
- Social and environmental wellbeing.

5.3.5.6 The rights of the individual

Parliament also focused on the rights of individuals. Whereas earlier versions of the act did not include them in the category of 'interested persons' (stakeholders), they are now expressly granted the rights to lodge complaints with supervisory authorities, to obtain explanations of the decision-making process by those employing high-risk systems, and the possibility of taking representative actions.

5.3.5.7 Transition period

After the entry into force of the regulation, currently estimated by the end of this year, the legislator is envisaging a transition period, aimed at allowing the obliged parties to adapt to the complex regulatory provisions (as well as, for example, to set up the AI Guarantee Authority). Initially 36 months, the transition period has been shortened to 24 by parliament and is expected to be further shortened in the final approval phase (to 12 or 18 months).

5.3.5.8 The EU standardisation request

In order to practically specify many of the requirements of the AI Act, the EU Commission has drafted a standardisation request to the CEN-CENELEC and prompted the national standardisation bodies in starting the definition of standards for AI systems. Among them, many are relevant for the health sector. These standards will define the practical and technical requirements in order to ensure compliance with the AI Act.

These 11 items represent the overarching map of the forthcoming standards on AI, and their relationships, as depicted in the EU Commission standardisation request.

5.4 The standardisation of AI in the CEN-CENELEC

The *European Committee for Electrotechnical Standardization* (CENELEC) is a European committee aimed at harmonising and developing technical standards in the fields of electrotechnics and electrical engineering, in collaboration with regulatory bodies at both national and supranational levels. Founded in 1973, this non-profit organisation operates under Belgian law and is headquartered in Brussels. CENELEC, together with *the European Telecommunications Standards Institute* (ETSI) and the *European Committee for Standardization* (CEN), forms the European system for standardisation. Although it is not an institutional structure of the European Union, CENELEC works closely with it.

CEN and CENELEC have established a new Joint Technical Committee named CEN-CENELEC 21 "Artificial Intelligence", following recommendations presented in response to the White Paper on Artificial Intelligence by the EC, the Road Map of the CEN-CENELEC Focus Group on Artificial Intelligence, and the German Standardization Roadmap for Artificial Intelligence. This Joint Technical Committee, with its Secretariat managed by the Danish standardisation body DS, is tasked with developing and adopting standards related to AI and associated data. Moreover, it provides guidelines to other technical committees dealing with AI. Specifically, the CEN-CLC/JTC 21 identifies and adopts international standards (**Table 1**) that either already exist or are under development by other organisations, such as the ISO/IEC JTC 1, subcommittees including SC 42 Artificial Intelligence. The CEN-CLC/JTC 21 is also engaged in producing standards that cater to the needs of the European market and society, aiming to support the legislation, policies, principles, and values of the European Union.

Project reference	Work item	Title
EN ISO/IEC 22989:2023	JT021004	Information technology - Artificial intelligence - Artificial intelligence concepts and terminology (ISO/IEC 22989:2022)
EN ISO/IEC 23053:2023	JT021005	Framework for Artificial Intelligence Systems Using Machine Learning (ISO/IEC 23053:2022)
prCEN/CLC/TR 17894	JT021001	Artificial Intelligence Conformity Assessment
prCEN/CLC/TR XXXX	JT021002	Artificial Intelligence - Overview of AI tasks and functionalities related to natural language processing
prCEN/TR XXX	JT021010	Information Technology - Artificial Intelligence - Green and Sustainable AI
prCEN/TR XXX	JT021009	AI Risks - Checklist for AI Risks Management
prCEN/TR XXX	JT021007	Data Governance and data quality for AI in the European context
prEN ISO/IEC 24029-2	JT021015	Artificial intelligence - Assessment of the robustness of neural networks - Part 2: Methodology for the use of formal methods
prEN ISO/IEC 25059	JT021014	Software engineering - Systems and software Quality Requirements and Evaluation - Quality model for AI systems
prEN ISO/IEC 42001	JT021011	Information technology - Artificial intelligence - Management system
prEN XXX	JT021012	Accuracy of natural language processing systems

prEN XXX	JT021008	Artificial Intelligence trustworthiness characterisation
prEN XXX	JT021019	Competence Requirements for AI ethicists professionals

Table 1. List of european standards on AI under development by the CEN-CENELEC JTC21.

6 Stakeholders’ perspectives in AI development: the VSD methodology

Value Sensitive Design (VSD) is a praiseworthy methodology employed in the creation of technology that accounts for human values in a comprehensive manner. Developed in the early 1990s, VSD has since been gaining popularity as a framework that insists on considering human aspects during the design process. This approach is especially vital in the development of AI systems, as these technologies are becoming more integrated into the daily lives of individuals and society at large.

The core principle of VSD is to identify and incorporate values that are relevant to the technology being created. These values might encompass privacy, autonomy, trust, fairness, accessibility, and well-being. VSD is proactive: rather than retroactively addressing issues once they surface, it encourages designers and developers to anticipate and address potential ethical concerns during the design phase. This is achieved through a tripartite methodology comprising conceptual, empirical, and technical investigations.

AI technologies are rapidly evolving, and their applications range from autonomous vehicles and healthcare diagnostics to recommendation systems and natural language processing. These systems, which often involve decision-making processes, can have far-reaching consequences on human lives. In the context of AI systems for medical applications, the importance of VSD cannot be overstressed. In fact, *medical decision support systems* (MDSS) impact multiple stakeholders in the delicate processes of diagnosis and therapy. To consider the expectations and concerns of every stakeholder is of paramount importance to ensure adoption and technology effectiveness.

One of the foremost values that VSD helps safeguard in AI systems is **fairness**. AI systems often rely on data to make decisions. However, if the data is biased, AI systems can perpetuate and even amplify these biases. For example, an AI-driven decision support system might have been trained on data that are not representative of the population, and therefore cause threats to patients. VSD prompts designers to identify such issues at the conceptual stage, to gather empirical evidence on potential biases, and to design technical solutions that mitigate these biases.

Privacy is another value that VSD highlights in AI systems. Many AI applications require access to sensitive personal information. Without the thoughtful integration of privacy values, these systems could inadvertently expose or misuse personal data. VSD ensures that privacy is not an afterthought, but an integral part of the system from its inception.

Autonomy is yet another critical value. As AI systems become more autonomous, there is a risk that they may inadvertently erode human autonomy. For instance, an AI system that makes healthcare decisions could diminish a patient’s ability to have a say in their own healthcare. VSD ensures that AI systems are designed with the preservation of human autonomy in mind.

In a world where AI technologies are increasingly prevalent, ensuring that these systems are aligned with human values is of paramount importance. VSD serves as a formidable approach to embedding ethics into the fabric of AI systems. By ensuring fairness, privacy, and autonomy, among other values, VSD contributes not only to the development of more ethical AI systems, but also to the nurturing of a society in which technology serves as a force for good.

6.1 The VSD process

The image depicts a flowchart that outlines the VSD process, which is an approach to designing technology that accounts for human values in a principled and comprehensive manner throughout the design process.

At the top, we have "Conceptual Investigations," which focuses on understanding and defining the values important to stakeholders before any design takes place. This step involves identifying heritage values as well as explicitly supported stakeholder values, analysing, and prioritising them. This ensures that the design process starts with a clear understanding of what values are important to the users and other affected parties.

The second step is "Empirical Investigations." Here, the key values that have been articulated by stakeholders are examined through empirical methods—such as observations, interviews, and surveys—to understand how these values manifest in the real world and how they can be evaluated in concrete scenarios. This helps in ensuring that the design is grounded in real-world conditions and user behaviours.

The last step is "Technical Investigations," which involves the actual development of the technology. It emphasizes that the properties of the technology should be thoroughly analysed to ensure that the design aligns with and supports the values committed to in the previous steps. The arrows between the steps suggest an iterative process, where insights from one stage may influence and lead to revisions in the others. For instance, technical constraints identified during Technical Investigations might lead to a re-evaluation of stakeholder values or adjustments in empirical methods.

7 The implementation of VSD in SYNTHEMA

Within the SYNTHEMA context, the VSD methodology has proven to be valuable for comprehending the diverse concerns held by stakeholders.

Methodologically, VSD facilitates the dual exploration of concerns and expectations that are inherent to a particular technology, as well as those perceived by stakeholders that may not be intrinsic to the technology itself. It is crucial to distinctly recognise this dichotomy: during the development of any technological artifact, developers make decisions concerning its functionality, scope, design, and interactions with the environment. Each decision plays a pivotal role in shaping how the technology interacts with its surroundings, thereby influencing its implications and impacts. These are the inherent values embedded within the technology.

Conversely, certain values, such as expectations and concerns, may be perceived as integral to the technology by one or more stakeholders, despite lacking a technical basis. This discrepancy can arise due to the technology's inability to effectively convey its purpose and functions through its design. A compelling example of this latter scenario can be observed in the case of 5G technologies, which are often associated with numerous perceived threats that are not substantiated by technical evidence. As a result, it becomes imperative to address not only the actual value choices made by developers but also those design elements that have the potential to engender misunderstandings and misinterpretations among one or more stakeholders.

7.1 The process

7.1.1 Stakeholder analysis

During the value-sensitive design sessions, it is paramount to acknowledge that diverse stakeholders hold distinct perspectives regarding concerns and potential applications of a given technology. In order to gain a comprehensive understanding of these varied viewpoints, the initial step involves the identification of these stakeholders.

In the initial four months of the project, a comprehensive stakeholder analysis was conducted, drawing upon pertinent literature pertaining to SDGs and MDSS (McParland 2019[10], Wijnhoven

[10] McParland, C.R., Cooper, M.A. and Johnston, B., 2019. Differential diagnosis decision support systems in primary and out-of-hours care: a qualitative analysis of the needs of key stakeholders in Scotland. Journal of Primary Care & Community Health, 10, p.2150132719829315.

2019[11], Vasey 2022[12], Webber 2023[13]). It is noteworthy that MDSS is the technology most likely to incorporate the outcomes of the SYNTHEMA SDG platform, as documented in authoritative sources.

Regular monthly meetings were convened with technical and medical partners, wherein deliberations were held to discern the key stakeholders crucial to the design process of the project. The culmination of this analysis has yielded the following list:

- AI developers
- Platform providers
- Physicians
- Clinical researchers
- Non-clinical and non-AI academics
- Legals
- Privacy experts
- Managers of clinics accessing the SDG platform
- Patients & patients associations
- Standardisation body
- Regulators
- Digital innovators

By undertaking this systematic stakeholder analysis, we are well-equipped to engage with a diverse range of perspectives, ensuring that the ensuing value-sensitive design sessions incorporate the multifaceted concerns and needs of all relevant stakeholders. This map will be useful in later stages of the project, in order to gain insights on values, expectations and concerns that are representative of the different stakeholders.

7.1.2 Focus groups

In the context of SYNTHEMA, two distinct focus groups were convened with the primary aim of eliciting the concerns and expectations of two distinct stakeholder groups: technical partners and clinical experts. It is imperative to emphasize that a comprehensive understanding of the perspectives held by these clinical experts is pivotal for the accurate interpretation of the concerns raised by other stakeholders.

[11] Wijnhoven, F. and Koerkamp, R.K., 2019, October. Barriers for adoption of analytical cdss in healthcare: Insights from case stakeholders. In 5th International Conference on Information and Communication Technologies in Organizations and Society, ICTO.

[12] Vasey, B., Nagendran, M., Campbell, B., Clifton, D.A., Collins, G.S., Denaxas, S., Denniston, A.K., Faes, L., Geerts, B., Ibrahim, M. and Liu, X., 2022. Reporting guideline for the early-stage clinical evaluation of decision support systems driven by artificial intelligence: DECIDE-AI. Nature medicine, 28(5), pp.924-933.

[13] Webber, C.M., Riberdy Hammer, A., Saha, A., Marinac-Dabic, D., Caños, D.A. and Tarver, M.E., 2023. Integrating Patient-Generated Health Data Throughout the Total Product Life Cycle of Medical Devices. Therapeutic Innovation & Regulatory Science, pp.1-5.

These focus groups were executed during the M6 in-person meeting held in Lisbon. The decision to conduct them separately stemmed from the recognition that segregating the different stakeholder categories during these sessions is crucial to fostering an environment where all participants feel comfortable expressing their viewpoints. Without such separation, there is a risk that discussions pertaining to technical aspects of the technology would predominantly involve developers and technical partners, potentially inhibiting other participants from voicing their perspectives in the presence of experts. Each of these focus group sessions had a duration of approximately one hour and was scheduled at the conclusion of each of the two days of the project meeting.

Each focus group consisted of approximately ten participants. The facilitator of the focus group initiated the discussion by presenting the "Ethics Guidelines for Trustworthy AI" established by the *High-Level Expert Group on Artificial Intelligence* (HLEG) in 2019. This served as a foundational reference to kick-start the dialogue and encourage participants to freely express their thoughts and impressions concerning the SYNTHEMA platform and its alignment with *Sustainable Development Goals*. To facilitate the discussion, the focus group facilitator posed a set of structured questions for consideration, corresponding to each of the ethical principles outlined in the HLEG's guidelines.

«Imagine a situation where one of these principles is negatively impacted by SYNTHEMA technologies»

«Imagine you are a patient/physician, what concerns would you have toward SYNTHEMA technologies?»

When relevant, «How will SYNTHEMA positively impact on Sustainable Development Goals? »

Sometimes the discussion overcame some of the questions related to ethics principles, but the objective of these initial focus groups was to gather the expert stakeholders' concerns and expectations. More specific issues will be enquired in a later stage, through semi-structured interviews.

The two focus groups conducted as part of SYNTHEMA shed light on the varying perspectives of technical and clinical partners concerning the use of AI in healthcare.

7.1.2.1 Focus Group 1 - Technical partners

Technical partners highlighted a range of ethical concerns:

- **Accountability**: Concerns were raised about the potential for AI predictions to bias clinician judgment, questioning who would be responsible in case of failure and emphasizing the need for clinicians to remain in control and make informed decisions.
- **Privacy**: The risk of data breaches was acknowledged, despite methods like Federated Learning (FL) and Secure Multi-party Computation (SMPC) that aim to mitigate such risks.

- **Transparency and Explainability**: There was an awareness of bias in data leading to distrust in AI among clinicians, with some relying too heavily on AI and others rejecting it outright.
- **Non-discrimination and Fairness**: Bias in health data representation was noted as a concern, especially for SDGs. SYNTHEMA's reliance on data augmentation could inadvertently perpetuate existing biases, a significant risk in the health sector. If augmentation algorithms amplify gender, ethnic, or other biases, the resulting datasets may lead to skewed AI predictions, affecting treatment outcomes. Ensuring representativeness in datasets is critical, requiring vigilant checks and balances.
- **General issues**: Fear among patients regarding their data use and the reliability of AI decisions, clinicians' fear of being replaced, and questions about the platform's validation were discussed.
- **Advantages**: Benefits identified included enhanced health, simplicity of use, more equal treatment, increased accessibility, and promotion of networking and collaboration.

7.1.2.2 Focus Group 2 - Clinical partners

Clinical partners focused on the importance of human oversight and raised specific issues:

- **Technical Robustness and Safety**: The necessity for human supervision of technology was emphasized.
- **Privacy and Data Governance**: The need for clear communication regarding data use and limitations was highlighted.
- **Transparency, Diversity, and Fairness**: The importance of informing physicians about the system's potential and pitfalls was discussed, alongside legal measures to control data use.
- **Societal and Environmental Well-being**: The discussion included barriers to technology adoption and the quality of data, with a focus on trust from both patients and physicians.
- **Accountability**: Recommendations for building trust through clear policy information were suggested. In fact, it is unclear who bears responsibility for eventual mistakes of the technology.
- **Physician-Patient Relationship**: It was discussed that patients might feel detached from healthcare, they might think they lose control over the data or that physicians are not dealing with them directly, therefore producing a sensation of abandonment.
- **Trust**: Patients may fear that the subscription or prediction based on synthetic data are not reliable, and therefore might be less likely to correctly follow the treatment. On the other hand, the fact that the data to be used for research is synthetic and not their, might be seen positively, for privacy reasons. An important part of the discussion focused on how to build patient trust on AI technologies for the healthcare sector.

It was emphasized that for some diseases, e.g. lifelong genetic conditions, there is long lasting relation of trust between patients and physicians, and hence the trust in the technology can be directly related to the trust they have in physicians.

Disease-Specific Concerns: During the discussion, it has been discussed that diseases like SCD and AML have different implications and should be addressed separately.

7.1.2.3 Differences between technical and clinical partners

Technical partners are primarily concerned with the technical and ethical implications of AI, such as privacy risks, accountability, and biases inherent in AI systems.

Clinical partners focus more on the human aspect of AI use, including maintaining human oversight, ensuring patient trust, and the practicalities of data governance and transparency.

Both groups express concern about bias and the potential for AI to replace human judgment, but technical partners delve into the specifics of technological safeguards, while clinical partners emphasize the human relationship and communication aspects.

7.1.2.4 Key takeaways

The following are some of the key takeaways from the focus groups. All these considerations contributed to shaping the ethics requirements for SYNTHEMA.

In the context of SYNTHEMA technology adoption, the inclination of physicians to place their trust in such advanced systems may hinge on subjective factors, including a predilection for or skepticism about technology. This dichotomy could manifest in their decision-making process; they may either uncritically accept the AI (AI) system's recommendations or entirely dismiss them, especially if the rationale behind the AI's conclusions is not sufficiently transparent.

To address this challenge, it is imperative to ensure that the intricacies of the platform's functionality and its inherent constraints are comprehensively communicated to medical professionals. This can be effectively accomplished through the meticulous design of the user interface, which should not only facilitate access to high-quality information but also enhance the interpretability of the system’s output. Such measures would serve to bridge the gap between the technological capabilities of the platform and the practical needs of its users.

Furthermore, the issue of bias replication within the sphere of data augmentation warrants serious consideration. The inadvertent reinforcement of existing societal biases — whether they pertain to gender, ethnicity, or other factors — could lead to the creation of datasets that fail to be truly representative. The opacity of the AI decision-making process, colloquially referred to as the "black box" issue, might consequently lead to flawed clinical decisions.

Addressing these concerns necessitates a twofold approach: firstly, the implementation of rigorous protocols to scrutinize and validate the quality and representativeness of data; secondly, the establishment of a framework for evaluating and mitigating the potential propagation of biases.

In parallel, the legal ramifications of utilizing data within the SYNTHEMA framework must be thoroughly evaluated. This encompasses identifying the entities accountable for the legal outcomes arising from the use of SYNTHEMA data. Establishing robust legal agreements with stakeholders who will access and utilize SYNTHEMA data and technology is paramount.

Lastly, the safeguarding of patient privacy against cyber threats remains a topic of paramount importance. Questions arise about the efficacy of Federated Learning and Secure Multi-party Computation in legally anonymizing data in compliance with stringent regulations, such as those enforced by data protection authorities. It is essential to quantify the extent to which these technologies can withstand adversarial threats and maintain the confidentiality and integrity of sensitive patient data.

In conclusion, the transition towards integrating SYNTHEMA innovative solutions into clinical practice must be underpinned by a thorough understanding of the system's capabilities, a commitment to fairness and legal compliance, and a relentless focus on protecting the sanctity of patient data.

Funded by
the European Union

8 Ethics requirements for SYNTHEMA

The ethics requirements for technology development in SYNTHEMA have been conceived to take into account all the EU's ethics for AI frameworks described above, including the relevant ISO/IEC JTC 21 standards for implementation where appropriate, and be tailored on the specificities of the SYNTHEMA technologies, as well as the needs, expectations and concerns of its stakeholders towards them, including AI developers', end users' and beneficiaries', investigated through a VSD approach.

These requirements have been conceptualised in its first version in M1-M12 of the project, through the processes described in this deliverable, and are included at the end of this section. These will guide the technical development of SYNTHEMA technologies and support its monitoring by the ethics AI experts in the consortium. In doing so, and through additional stakeholder engagement and feedback gathering activities, these are going to be refined and updated throughout the project implementation and will be released in their final version as part of the D7.3 - Ethics handbook (DW, M48).

8.1 The ethics requirement elicitation process

The requirement elicitation process on this preparatory phase of has been conducted under the lead of DW involving all partners in the consortium, as representatives of SYNTHEMA internal technology stakeholders. The process has encompassed the following phases:

1. **Ethical assessment framework building for SYNTHEMA (M1-M5)**: DW assessed SYNTHEMA GA with focus on the technologies to be developed in the project, and in relation to that conducted an analysis of the scientific literature and the applicable ethical frameworks for SYNTHEMA technologies (Guidelines for Trustworthy AI, AI Act) and their more recent developments.
2. **Stakeholder mapping (M5-M6)**: an initial mapping of stakeholders of SYNTHEMA technologies was realised to guide the following engagement phase, including the following groups:
    - Group A: AI developers, digital innovators.
    - Group B: physicians, clinical researchers and patients.
    - Group C: non-clinical and non-ai academics, legal and privacy experts.
    - Group D: standardisation bodies and regulators.

    These were used for a preliminary value extraction, mapping of value tensions, and scenario building to clarify different stakeholders' expected impacts, as described in the sections above.
3. **Focus groups (M6-M7):** 2 in-person focus groups have been conducted.
4. **Drafting of ethical requirements for SYNTHEMA (M10-M12)**: based on the applicable ethical frameworks, the preliminary analysis, and the focus groups discussions,

Funded by the European Union

a first version of the requirements were elaborated, reviewed and updated (also including the publication of the AI Act in November 2023) and released as part of the present deliverable.

5. **Next steps: monitoring, refinement and assessment of AI implementation**

- From M13 to M48 of the project, DW will continuously monitor the development of SYNTHEMA technologies, including assessing the risk categories of AI and technology systems developed in SYNTHEMA, and provide guidance for development and mitigation actions where necessary.
- In parallel, additional external stakeholders will be engaged and their feedback will be gathered through semi-structured interviews, and requirements will be refined accordingly.
- Once the technological development is completed, DW will conduct a final assessment of the technologies, and develop policy recommendations for the SYNTHEMA technologies, and a general risk management approach to ensure ethical and privacy compliance of SYNTHEMA technologies, for research and for market-oriented deployments, all to be incorporated in D7.3.

8.2 Ethics requirements for SYNTHEMA technologies

In the table below, the initial requirements for technology development in SYNTHEMA are listed. These include: in the top row, the reader finds (1) a rationale of the each principle considered; on the left column below, the (2) essential requirements to be complied with in relation to the applicable key requirements for trustworthy AI and other principles of relevance; on the right column below, where appropriate, some additional (3) recommendations, i.e., measures that are not mandatory, but nice to have, in relation to those principles.

<table>
<tr><td colspan="2"><b>(1) SDG standard requirements</b>: Many standards are forthcoming both from ISO and CEN-CENELEC on Artificial Intelligence, and some of them address generative AIs, and synthetic data generation specifically. Specifically, ISO/IEC AWI TR 42103 - Information technology — Artificial intelligence — Overview of synthetic data in the context of AI systems.</td></tr>
<tr><td><b>Requirement</b></td><td><b>Recommendation</b></td></tr>
<tr><td>When available, ISO/IEC AWI TR 42103 will be taken into account by the SYNTHEMA project</td><td>The participation of experts of the SYNTHEMA consortium inside one of ISO national bodies is advised. This would allow early access to the standard under development, and the possibility to engage with the standard contents and contribute in drafting it.</td></tr>
</table>

<table>
<tr><td colspan="2"><b>(2) Fairness</b>: To avoid biases, or underrepresentation of some demographic categories, it is paramount that the models have been trained to ensure that the data represents the incidence of the disease. For SDG, this requirement needs even more attention, since the bias might be replicated in synthetic dataset, leading to consequences of concerning magnitude.</td></tr>
<tr><th>Requirement</th><th>Recommendation</th></tr>
<tr><td>At least <b>ISO/IEC TR 24027 Information technology — Artificial intelligence (AI) — “Bias in AI systems and AI aided decision making”</b> will be applied during the project, and its application documented and shared with partners. If published in time for application in the project, <b>every</b> EU standard pertaining to data quality will be applied.</td><td>The participation of experts of the SYNTHEMA consortium inside one of CEN-CENELEC national bodies is advised. This would allow early access to the standard under development, and the possibility to engage with the standard contents and contribute in drafting it.</td></tr>
</table>

<table>
<tr><td colspan="2"><b>(3) Oversight and acceptance:</b> The discussion on the implication of the AI system design on the perceived trustworthiness of physicians has been key during focus groups. Certainly the technology should be technically trustworthy to be effective, but should also be perceived as such to be adopted and used in the right way. Literature on DSS shows clearly the importance of design in this area in order to ensure social acceptance, adoption and therefore technology effectiveness.</td></tr>
<tr><th>Requirement</th><th>Recommendation</th></tr>
<tr><td>The physicians will be clearly informed about the platform limitations and the functioning of SDG, in order to ensure the adequate level of trust in the platform. UX and UI aspects, if any, will be tested through scientific measures for social acceptance (TAM, UTAUT).</td><td>The <b>implications of the model’s interpretability on the physician's perceived trustworthiness</b> should be thoroughly investigated.</td></tr>
</table>

(4) Building patient trust: during the value-sensitive design sessions, an outstanding issue discussed with clinical partners was the possible mistrust of patients in the use of SDG, eventually leading to worse therapy adherence.

Requirement	Recommendation
At least **10 patients will be interviewed** in order to understand their concerns related to SDGs and how to address them. This perspectives will be discussed by clinical and technical partners before M36	Use of quantitative assessment on technology acceptance and trustworthiness.

(5) Explainability: Explainability is a key concept to ensure that all the other requirements can be put into practice. Without explainability the decision of the system cannot be interpreted by physicians, cannot be logged and then be transparent for stakeholders, and accountability cannot be fully achieved.

Requirement	Recommendation
Explainability must be ensured: the **contribution of each feature to the model's output** shall be tested through methods like Shapley values, LIME, and others.	Current methods for ensuring explainability could be overcome and improved in the near future. The consortium should follow **up-to-date approaches and standardised methodologies from the CEN-CENELEC or ISO**

(6) Accountability: AI should be accountable in order to ensure trustworthiness and liability, in case of errors or malfunctions.

Requirement	Recommendation
A governance process for assessing accountability of eventual errors or malfunctions of the SYNTHEMA platform should be clarified before M42, to consider inclusion in D7.3. This process should consider the possible integration of the SYNTHEMA platform in other technologies	The relevant **standards on accountability and governance** should be followed.

(7) Value-sensitive design: in order to guarantee that developed AI systems are trustworthy and adopted, the inputs of stakeholders should be carefully taken into account, to ensure that concerns are mitigated and expectations are met.	
Requirement	**Recommendation**
Concerns, expectations, and values of stakeholders are **mapped and taken into account both before and during systems development**.	The consortium should **reach out to different stakeholders, also not comprised among consortium partners**, to include each stakeholder point of view in the development and design of technologies

9 Conclusions

Integration of ethical principles in AI development: the document underscores the importance of integrating ethical principles in the development and deployment of AI systems. These principles include fairness, accountability, transparency, and privacy. The aim is to ensure that AI technologies respect human rights and contribute positively to societal well-being.
This deliverable touched on many fundamental aspects of ethics of AI. We briefly summarise below the most important take-home messages.

Stakeholder perspectives and VSD: the document highlights the significance of considering diverse stakeholder perspectives through the VSD methodology. This approach ensures that the values of various stakeholders, including developers, end-users, and beneficiaries, are incorporated into the technology development process.

Focus on fairness and non-discrimination: special emphasis is placed on addressing biases in AI systems. The document suggests that ensuring fairness and non-discrimination in AI applications, especially in case of SDG, is crucial to prevent the perpetuation of existing societal biases. In the case of data augmentation, the impact of biases might be of unprecedented magnitude.

Regulatory compliance and standardisation: the document discusses the need for AI systems to comply with existing and emerging regulations and standards. This includes adherence to the European AI Act, GDPR, and relevant CEN-CENELEC and ISO/IEC standards. Ensuring compliance will aid in harmonizing ethical considerations across different jurisdictions and applications.

Building trust through transparency and explainability: the document stresses the importance of making AI systems transparent and explainable to build trust among users, particularly in healthcare settings. Clear communication about the functionalities and limitations of AI systems is essential for their acceptance and effective use.

Patient-centred approach in healthcare AI: in the context of healthcare, the document emphasizes the need for AI systems to prioritize patient trust and privacy. This includes addressing patients' concerns about data use and ensuring that AI-assisted decisions in healthcare are made with a clear understanding of the underlying algorithms and data.

Next steps and policy recommendations: looking ahead, the document proposes a roadmap for the continuous development and assessment of ethical AI technologies. This includes engaging additional stakeholders, refining ethical requirements based on feedback, and developing comprehensive policy recommendations for SYNTHEMA technologies.

Emphasis on legal and societal implications: the document acknowledges the legal and societal implications of AI technology. It recommends establishing robust legal frameworks and agreements to manage the use of AI data and technologies, ensuring that ethical and privacy compliance is maintained.

Next steps: SYNTHEMA will be continuously supervised by an ethics manager for the entire duration of the project. This supervision will imply monitoring the effective implementation of the seven ethics requirements above, carrying out interviews with stakeholders to further apply value-sensitive design, the monitoring of ethical and societal implications of the technical features designed and implemented. In M42, a final deliverable will summarize the work done in SYNTHEMA to ensure ethical design of AI systems, the lessons learnt, and the policy recommendation for generated datasets in the medical domain.