
D2.2 / Trustworthy AI development and evaluation framework (fundamental version)

Editor

Ali Saad (AING)

Contractual delivery

January 2024

Actual delivery

February 2024

Deliverable type

R - Document, report

Dissemination level

PU – Public

Version - date

1.0 - 05/02/2024



Funded by the
European Union

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency. Neither the European Union nor the European Health and Digital Executive Agency can be held responsible for them.

Deliverable ID

Project acronym	AI-PROGNOSIS
Project full title	Artificial intelligence-based Parkinson's disease risk assessment and prognosis
Grant Agreement ID	101080581
Deliverable number	D2.2
Deliverable title	Trustworthy AI development and evaluation framework (fundamental version)
Work package	WP2 - Foundation, data curation and co-creation
Deliverable type	R - Document, report
Dissemination level	PU – Public
Version - date	1.0 - 05/02/2024
Contractual delivery	January 2024
Actual delivery	February 2024
Lead partner	AING
Editor	Ali Saad (AING)
Contributors	Ioannis Gerasimou, Stelios Hadjidimitriou (AUTH); David Lyreskog (UOXF); Ioannis Drivas (DBC); John Zaras, Charis Giaralis (SQD); Christos Chatzichristos (KUL); Andreas Stergioulas (CERTH)
Reviewed by	Dorine Matzakou-Karvouniari (INTRA) John Zaras (SQD)
Approved by	Leontios Hadjileontiadis (AUTH, Project Coordinator)
Keywords	Accountability; Artificial intelligence; Explainability; Fairness; Generalisation; Machine learning; Parkinson disease; Privacy protection; Reproducibility; Robustness; Transparency; Trustworthy AI.

Document history

Version	Date	Contributors	Action / status
0.1	17/10/2023	AING	Document structure (table of contents) draft
0.2	16/11/2023	AING	Document structure (table of contents) updated draft
0.3	12/12/2023	AING	Updated draft
0.4	13/12/2023	AING, UOXF, and DBC	Updated draft
0.5	15/12/2023	AING and AUTH	Updated draft
0.6	09/01/2024	AING, AUTH, DBC, CERTH, and SQD	Updated draft
0.7	22/01/2024	AING, SQD	Updated draft
0.8	24/01/2024	SQD	Reviewed by John Zaras (SQD)
0.8	01/02/2024	INTRA	Reviewed by Dorine Matzakou-Karvouniari (INTRA)
0.9	01/02/2024	AING, AUTH	Document revised; Minor edits; Approved for submission by the Project Coordinator
1.0	05/02/2024	AUTH	Submitted version

Contents

List of figures	7
List of tables.....	8
List of abbreviations	9
Executive summary.....	10
1 Introduction	11
1.1 Document scope.....	11
1.2 Document structure	11
2 Trustworthy AI foundation	11
2.1 Defining trustworthy AI dimensions.....	12
2.1.1 Robustness	12
2.1.2 Generalisation.....	12
2.1.3 Explainability	13
2.1.4 Transparency and accountability.....	13
2.1.5 Reproducibility	14
2.1.6 Fairness	14
2.1.7 Privacy.....	14
2.2 Socio-ethical dimension.....	14
2.2.1 Socio-ethical aspects of trustworthy AI	15
2.2.2 Socio-ethical criticisms of trustworthy AI	16
2.2.3 AI-PROGNOSIS, ethics and society.....	17
3 Existing guidelines, standards, and regulations.....	17
3.1 ALTAI requirements.....	19
3.1.1 Overview of the seven HLEG-AI requirements within the ALTAI tool.....	21
3.2 AI Act.....	27
3.3 Regulatory compliance and risk management for AI in medical devices	29
3.4 ML technologies in medical devices.....	31
3.5 Responsible AI pillars in the industry	31
4 The AI-PROGNOSIS framework for AI development and evaluation.....	32
4.1 Design and specification	34
4.1.1 Embedding personal autonomy and exercising oversight.....	34
4.1.2 Explainability tailored to each user	34
4.1.3 Identify fairness goals and protected attributes	34
4.1.4 Equal access to stakeholders.....	34
4.2 Data preparation.....	35
4.2.1 Outlier detection and data sanitisation.....	35
4.2.2 Data pseudonymisation or anonymisation	36

4.2.3 Data quality assessment and bias identification	36
4.3 Development and internal validation	36
4.3.1 Training for generalisation	36
4.3.2 Adversarial training and regularisation	36
4.3.3 Performance benchmarking	37
4.3.4 Explainable model design	37
4.3.5 Uncertainty benchmarking.....	38
4.3.6 Assessment and mitigation of bias	38
4.3.7 Production model - performance and privacy risk trade-off.....	39
4.4 UX/UI and deployment.....	39
4.4.1 User feedback.....	39
4.4.2 Adversarial testing.....	39
4.4.3 Disclaimers	40
4.4.4 Input data validation	40
4.4.5 Data minimisation.....	40
4.4.6 Right to be forgotten.....	40
4.4.7 Explanation of models and their outputs.....	41
4.4.8 Educational content on AI	41
4.4.9 UI/UX best practices and performance optimisation	41
4.5 External validation	42
4.5.1 Diversity, non-discrimination and fairness in clinical studies	42
4.5.2 Observability for robust clinical data validation	42
4.5.3 End-user assessment of model output explanations in clinical studies	43
4.6 Overarching management and workflow.....	43
4.6.1 Model Operations.....	43
4.6.2 Model and dataset reporting.....	44
4.6.3 Open models and datasets	44
4.6.4 Security and GDPR compliance	45
4.6.5 Sustainability and societal impact.....	45
5 Conclusion – key takeaways	46
References	48

List of figures

Figure 1 The seven trustworthiness requirements published by the HLEG-AI (High-Level Expert Group on Artificial Intelligence, 2019).	20
Figure 2 An example of the spider diagram of the ALTAI web-based self-assessment tool.	21

List of tables

Table 1 Description of important frameworks and guidelines for trustworthy AI practices. ..	17
Table 2 Components of the human agency and oversight (HAO) requirement of the ALTAI.	22
Table 3 Components of the technical robustness and safety (TRS) requirement of the ALTAI.	23
Table 4 Components of the privacy and data governance (PDG) requirement of the ALTAI.	24
Table 5 Components of the TPR requirement of the ALTAI.	24
Table 6 Components of the diversity, non-discrimination, and fairness (DnDF) requirement of the ALTAI.	25
Table 7 Components of the societal and environmental wellbeing (SEW) requirement of the ALTAI.	26
Table 8 Components of the accountability (ACC) requirement of the ALTAI.	27
Table 9 ALTAI requirements and summary of relevant components associated with key stages in the development lifecycle of AI models within AI-PROGNOSIS.	32
Table 10 AI-PROGNOSIS components aligned with ALTAI Requirements for each step in the AI component lifecycle.	46

List of abbreviations

ACC	Accountability
AI	Artificial Intelligence
ALTAI	Assessment List for Trustworthy Artificial Intelligence
AUC	Area Under the receiver operating characteristic Curve
C-index	Concordance index
DnDF	Diversity, non-Discrimination and Fairness
EOSC	European Open Science Cloud
EU	European Union
FAIR	Findability, Accessibility, Interoperability, and Reusability
FDA	Food and Drug Administration
GDPR	General Data Protection Regulation
HAO	Human Agency and Oversight
HLEG-AI	High-Level Expert Group on Artificial Intelligence
HTTPs	Hypertext Transfer Protocol Secure
MDR	Medical Device Regulation
ML	Machine Learning
MLOps	Machine Learning Operations
PDG	Privacy and Data Governance
RMSE	Root Mean Squared Error
SEW	Societal and Environmental Wellbeing
SSL	Secure Sockets Layer
TPR	Transparency
TLS	Transport Layer Security
TRS	Technical Robustness and Safety
UI	User Interface
UX	User Experience
XAI	Explainable Artificial Intelligence

Executive summary

This deliverable outlines the fundamental framework for trustworthy Artificial Intelligence (AI) development in the AI-PROGNOSIS project. Initially, it defines the dimensions of trustworthy AI, drawing from an extensive review of existing literature. The document then details the essential components for trustworthy AI development as per the existing guidelines. Additionally, this deliverable presents the adherence of the AI-PROGNOSIS project to ethical and regulatory standards, highlighting the socio-ethical contributions of AI in medical devices and emphasising the importance of trust and compliance in AI development. In this foundational version of the deliverable, we present components to be followed throughout the entire lifecycle of the AI-PROGNOSIS project. These components, aligned with existing guidelines, aim to implement trustworthy and reliable practices. Any updates about these components will be presented in the final version of the deliverable.

1 Introduction

1.1 Document scope

One of the main objectives of AI-PROGNOSIS is to develop advanced predictive AI models. These models are intended to build a risk assessment and prognosis tool for Parkinson's disease (PD). These models will utilise a variety of data sources, including patient records and databases, to predict factors such as the time to a higher disability transition and individual responses to medications. Thus, this deliverable intends to shed light on the regulatory standards and utilise them for developing a framework and a guideline for building trustworthy AI models. Following this framework will increase trust in the development and utilisation of the AI systems. Moreover, this will maximise the positive impacts of the AI system as well as minimising its negative impact on individuals.

AI is increasingly being integrated into various parts of human life. People have begun to incorporate AI into their daily activities and expose their personal data. However, it is important to note that while current AI systems excel in performance, our primary goal should be the development of AI systems that are not only high-performing but also reliable and trustworthy. Thus, the primary objective of this deliverable is to provide the necessary steps for the development of trustworthy AI systems. First, the goal is to create a comprehensive framework that defines standards, procedures, tools, and metrics to be employed across the different phases of the AI system's lifecycle. This framework serves as a guide for aligning the project's AI components with the principles of trustworthy AI. To ensure the effectiveness of the framework, it is built based on The Assessment List for Trustworthy Artificial Intelligence (ALTAI) (High-Level Expert Group on Artificial Intelligence, 2020) requirements. Moreover, the ALTAI requirements are compared with established AI management systems, including the AI Act and the European Union (EU) Medical devices – Regulation (MDR). This analysis helps identify overlaps and deviations, ensuring broad coverage of trustworthy AI principles. Moreover, these requirements are shaped according to the EU ethical and legal standards.

1.2 Document structure

The document begins with Section 1 as an introduction to provide an understanding of the AI-PROGNOSIS project and setting the stage for the framework and its importance. Following this, Section 2 includes a detailed section on the dimensions of trustworthy AI, incorporating insights from existing literature. Also, in that section, the socio-ethical aspects of trustworthy AI are highlighted emphasising about their effect on the development of the AI-PROGNOSIS project. Section 3 summarises the trustworthy requirements of existing guidelines and standards. Finally, Section 4 which is the core of the document delves into the specific components of AI development, aligning them with recognised guidelines and discussing methods to ensure the implementation of these components throughout the AI-PROGNOSIS project's lifecycle. Section 5 concludes the document and summarises the steps to be followed in the project for developing trustworthy AI tools.

2 Trustworthy AI foundation

Different research attempted to lay a definition of trust between humans and AI (Glikson & Woolley, 2020) (Jacovi et al., 2021) (Gillath et al., 2021). The main core ideas of these definitions are that trust in an AI system is the willingness of humans to engage and rely on the AI system, with a belief that the system will act in the human's best interest; thus, raising

a sense of confidence in the positive outcomes of the AI system based on the understanding of its reliability and trustworthy intentions. The important dimensions or aspects of the AI system that emphasises its trustworthiness include **robustness, generalisation, explainability, transparency, reproducibility, fairness, privacy preservation, and accountability** (Li et al., 2023). To establish the foundation of our AI trustworthiness framework, we must carefully define these dimensions. It is worth mentioning that these trustworthiness characteristics are correlated and they influence each other. Also, it is worth noting that ensuring that all the trustworthy dimensions are applied to a certain AI system is less likely to be valid. Trade-offs between these dimensions take place in most cases. Thus, there is a significant need to reach a certain balance between these dimensions. This balance changes based on the problem we are addressing, and based on the outcome of the AI system we intend to develop.

2.1 Defining trustworthy AI dimensions

In this section, definitions of the dimensions of AI trustworthiness will be presented. Moreover, relationships between certain dimensions will be highlighted. These definitions heavily rely on the ISO/IEC TS 5723:2022 (ISO/IEC, 2022) source that provides a definition of trustworthiness for systems and their associated services, along with a selected set of their characteristics.

2.1.1 Robustness

Robustness is the ability of a system to maintain its level of performance under a variety of circumstances. In general, robustness refers to an algorithm's capacity to sustain its performance even when exposed to diverse perturbations, adversarial inputs, erroneous data, or unseen data (Li et al., 2023). According to a systemic review that inspected key definitions of robustness and robust AI done by Tocchetti et al. (2023), two main robustness branches have been identified: robustness to adversarial attacks or perturbations, and robustness to natural perturbations. Adversarial attacks may include: a) decision-time attack that perturbs input samples during prediction to mislead the model, b) training-time attack (poisoning attack) that injects carefully designed samples into training data to alter the system's response to specific patterns, c) feature space attacks that are directly generated as input features of the model, d) problem space attacks that modify input entities to indirectly produce attack-related features, and e) model stealing (exploratory attack), which attempts to steal knowledge about models to generate adversarial samples without directly changing model behaviour (Li et al., 2023). While natural perturbation may be in the form of commonly witnessed natural noise, it is a condition more likely to occur in the real world compared to adversarial perturbations (Tocchetti et al., 2023).

2.1.2 Generalisation

Generalisation refers to the ability to derive knowledge from a limited training dataset to make accurate predictions when presented with new, previously unseen data. It encompasses the model's capacity to classify objects it was never initially trained on. The foundation of generalisation theory is the balance between underfitting and overfitting, with the goal of achieving a generalised model. Underfitting is also referred to as over-generalisation, which means that the Machine Learning (ML) model is trained in an extremely simple way, while if the model is extremely complex, it will not generalise well from observed data to unseen data (Ghojogh & Crowley, 2019). The notion of generalisation in AI closely intersects with other dimensions of AI trustworthiness, particularly its robustness. In ML, developing a robust model against shifts in data distribution negatively affects the generalisation of the model. Therefore, the robustness and generalisation dimensions have some overlapping aspects (Li et al., 2023).

2.1.3 Explainability

Explainability can be viewed as an active characteristic of a model, denoting any action or procedure taken by a model with the intent of clarifying or detailing its internal functions (Barredo Arrieta et al., 2020). Explainability is closely linked to the internal logic within a ML system. The greater the level of explainability in a model, the more understanding humans can gain in terms of the internal processes that occur during model training or decision-making. Understanding how an AI model reaches its decisions is an important factor in AI research and in increasing the trustworthiness of the AI system. From a scientific research perspective, having the knowledge about the essential mechanisms of data, parameters, procedures, and outcomes within an AI system, contributes significantly to the overall trustworthiness of AI systems.

In the field of ML, to offer an interpretable description of a model's behaviour, several actions can be taken. For example, sending all the model's parameter data or providing example predictions. Another approach involves summarising the components of the developed model in a tree-like explanation or giving information about the most influential features affecting the model's predictions. Each of these actions represents a potential means of explaining a complex model to the end user. The selection of these actions depends on both the application of the model and the intended end user. Model explainability focuses on developing inherently interpretable models and/or utilising *post-hoc* explainability methods. These techniques aim to explain the internal functions of ML systems, making them comprehensible to human understanding. *Post-hoc* explanation methods can either provide a global explanation of the entire black-box model or a local explanation that explains the individual predictions of the model and what are the factors impacting these predictions (Mahya & Fürnkranz, 2023). On the other hand, there are models that are considered to be explainable by themselves. For example, models like Linear/Logistic Regression, Decision Trees, K-Nearest Neighbours, and Bayesian Models have significant levels of inherent model explainability and require no additional post-hoc analysis (Barredo Arrieta et al., 2020).

2.1.4 Transparency and accountability

Transparency is the open, comprehensive, accessible, clear, and understandable presentation of information (ISO/IEC, 2022). Transparency has long been a recognised requirement in software engineering, requiring the disclosure of information about a system. In the AI industry, this essential principle covers the entire lifecycle of an AI system, allowing stakeholders to verify that appropriate design principles are met. To achieve transparency in the lifecycle of an AI system, different important information must be disclosed including design objectives, data sources, hardware specifications, configurations, operational conditions, anticipated usage patterns, and system performance metrics (Li et al., 2023). Furthermore, a frequently raised query regarding AI regulation is to know the ways of taking advantage of what AI systems have to offer while ensuring that those who develop and use AI are held accountable. Accountability is a property that ensures that actions of an entity can be traced uniquely to the entity (ISO/IEC, 2022). In this context, accountability means the ability to determine if a decision adhered to both procedural and substantive criteria, and the capability to assign responsibility when these standards are not upheld (Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, 2017). The main prerequisite of accountability is transparency. In other words, transparency serves as the primary mechanism enabling accountability in an AI system. It is worth noting that when taking actions to make AI systems more transparent and accountable, it is important to evaluate how these actions affect the organisation implementing these measures. Therefore, a balanced approach should be followed where transparency and accountability are improved without the disclosure of unnecessary resources or risking the exposure of valuable proprietary information (National Institute of Standards and Technology, 2023).

2.1.5 Reproducibility

Within the AI community, there is a significant concern about reproducibility among researchers and developers. Thus, reproducibility is a fundamental element of trustworthy AI. It signifies the ability for an independent researcher to reproduce either identical or reasonably similar results using the same data, code and settings that are followed by the original model developers. In addition to enabling the effective validation and verification of research in terms of its reliability, accuracy, and effectiveness of the results, reproducibility allows the community to implement latest approaches to conduct follow-up research. It also helps to ensure that the reported outcomes are not due to chance or specific circumstances, and it allows other researchers to build upon, validate, or improve upon existing AI work. It is important to mention that in the context of developing ML models, in particular Deep Learning models, reproducing results has become more challenging (Cockburn et al., 2020).

2.1.6 Fairness

Fairness is a common concern for AI practitioners. It represents a sociotechnical challenge that is important in avoiding the amplification of social biases within AI systems. By definition, fairness in AI systems is incorporating concepts of equality and equity, aiming to address harmful bias and discrimination (National Institute of Standards and Technology, 2023). Achieving fairness in AI often involves approaches to mitigate bias rooted in protected or sensitive variables. These variables define aspects of data that are socio-culturally unsafe for the application of the AI system. Literature on fairness in AI emphasises two primary aspects: a) technical aspects, focusing on bias and fairness within the ML systems, and b) social, legal, and ethical theories, related to the discrimination in ML (Caton & Haas, 2023). Technical approaches to mitigate unfairness typically occur before, during, or after modelling (Binns, 2018). Pre-processing techniques recognise data biases (i.e., the distributions of specific sensitive or protected variables are imbalanced) and seek to rectify imbalances in distributions. In-processing methods recognise the tendency for the ML model to become biased by dominant features during the modelling phase. Post-processing strategies recognise whether the actual output of the ML model is potential unfair toward protected variables or subgroups within those variables. Therefore, prioritising fairness in AI demands a multifaceted approach that includes technical adjustments, as well as an understanding of broader social implications.

2.1.7 Privacy

Privacy is the freedom from intrusion into the private life or affairs of an individual (ISO/IEC, 2022). Privacy has a huge impact on AI trustworthiness, as it involves safeguarding data capable of identifying individuals or households, encompassing information like names, ages, genders, facial images, fingerprints, and more (Li et al., 2023). Privacy values, such as anonymity, confidentiality, and control, should inform decisions throughout the AI system's lifecycle, including design, development, and deployment. However, addressing privacy-related risks can affect other AI trustworthiness aspects like fairness and transparency, often requiring trade-offs among these characteristics. Furthermore, privacy-enhancing techniques, such as data sparsity, that contribute positively to enhancing the privacy of an AI system, can also result in a significant loss in its accuracy (National Institute of Standards and Technology, 2023).

2.2 Socio-ethical dimension

Over the last decade, trustworthiness has emerged as one of the core pillars of ethical and responsible development in AI. In its prominent role, it has nonetheless received ample criticism in the literature on ethical and broader social impacts of AI. In the following subsection (Section 2.2.1), we will highlight the socio-ethical aspects of trustworthy AI as defined

and outlined above (Section 2.1); we will then relay key criticisms of the overarching definition and focus on trustworthiness (Section 2.2.2); Finally, we outline how the highlighted perspectives may affect the development of the AI-PROGNOSIS project, and projects like it in the future (Section 2.2.3).

2.2.1 Socio-ethical aspects of trustworthy AI

As outlined in Section. 2.1, *Robustness* typically refers to an algorithm's capacity to sustain its performance even when exposed to diverse perturbations, adversarial inputs, erroneous data, or unseen data (ISO/IEC, 2022). While there are technical definitions of what such robustness may amount to, the lived experience of what robust - and, conversely, non-robust - AI implementation entails is not easily covered by such definitions, but requires further probing. Crucially, natural perturbations involving unseen and/or erroneous data may give rise to false positives and false negatives in diagnosis, inaccurate prognosis, and inadequate or directly harmful disease management protocol deployment. The implications of non-robust applications are therefore likely to harm users, as well as society at large. However, it should be noted that models considered robust may also be harmful, depending on how they are applied and used. Similarly, the *Generalisability* of the model(s) is, while a prerequisite for fair and equitable application, not a guarantee against harm to users. First, high generalisability of a model does not entail 100% success rate, meaning that over-reliance on the model(s) will inevitably risk harming outliers (Safdar et al., 2020). Furthermore, the robustness and generalisability of a model may prove harmful to end-users and society at large despite working as intended: while these features, i.e., robustness and generalisability, may support trustworthiness in terms of accuracy, they do not necessarily support trustworthiness with regard to key biomedical principles of ethics (e.g., beneficence, non-maleficence, and autonomy (Beauchamp T.L., 2001) or to facilitate trustworthy relationships between healthcare professionals, caregivers, and patients (Dalton-Brown, 2020)). This highlights the limits of robustness and generalisability as drivers of trustworthiness on a socio-ethical level, leading to the need for a broader understanding of trustworthiness as a comprehensive concept.

Explainability and *Transparency* pose further ethical and social challenges, particularly in contrast with Robustness and Generalisability. Now, it is a well-documented problem for AI development to balance the accuracy of a model against its explainability (London, 2019). While the severity of this conflict will chiefly depend upon model design and delivery (Hamon et al., 2020) (Rudin, 2019) (Lipton, 2018), the broader issue will likely remain in applications aimed at persons with little or no literacy in algorithmic model design. In order to deploy AI-powered applications, one needs to ask to whom those applications, or their results, are meant to be explainable and transparent. In AI-PROGNOSIS, and in projects like it, the challenge will be to allow for the data processing, output, and potential action plan suggestions to be transparent and explainable to a range of stakeholders, including health care professionals, professional and informal caregivers, patients, as well as persons whose risk of developing Parkinson's Disease is uncertain (Julie Gerlings, Millie Søndergaard Jensen, 2022) (Ehsan et al., 2021). The level and details of explainability shall be adapted based on the stakeholder.

How the principles of Explainability and Transparency are implemented will directly impact *Accountability* both within the system itself, and between the system and other agents. If a decision is (partially or wholly) made based on information presented by a model, or a set of models, where a clear rationale of that decision cannot be provided, accountability suffers, as it relies on agents acknowledging why a certain decision was made rather than another. In AI-PROGNOSIS, and other projects like it, the picture is complicated by the fact that there is not one, but multiple models and applications at play throughout the ecosystem, making oversight

and appropriate accountability assignment more challenging: accountability in treatment management planning relies on not only its own transparency and explainability, but also on that of the prognosis of disease progression, which in turn relies on that of diagnosis. In this way, opacity (i.e., lack of explainability and transparency) is potentially contagious throughout the ecosystem as a whole, and care must be taken to not only ensure that each model and application in itself supports transparency and explainability, but to also implement measures to ensure that the complete ecosystem fosters it.

With regard to the principle of *Fairness*, much of AI development is currently occupied with minimising or erasing bias (Lin et al., 2021) (Trisha Mahoney, Kush R. Varshney, 2020). Even though that may be part of the implications on fairness, there is a broader set of problems of equality and equity at place, including issues surrounding access and capacity. On an institutional level, for instance, there are clinics, hospitals, and indeed whole healthcare systems which are unlikely to be able to afford (or in other ways be prohibited from acquiring) and/or deploy an AI-powered ecosystem, such as AI-PROGNOSIS. Other institutions, while in principle able to acquire the system (as a whole, or parts of it), may lack the necessary capacity to train people or otherwise operate the system. Similar issues arise on an individual level, where persons – be they doctors or patients – may lack access to, or be incapable of operating, the relevant applications. When making design choices for systems such as those developed for AI-PROGNOSIS, care ought to be taken to allow fairness in dissemination and usability on a broader level.

Privacy, while often relatively clearly defined in steering guidance on AI development (ISO/IEC, 2022), will nonetheless be understood and practiced in a multitude of ways within and between any given society(ies). Indeed, how privacy – and the safeguarding of it – should be understood and achieved in AI development, is a heavily debated topic (Murdoch, 2021) (Y. Zhang et al., 2021). As noted above (Section 2.1.7), in research and development the value of privacy may stand at odds with aims of transparency and accuracy within any given model. This issue becomes larger and more complicated as models are implemented: (1) in the field, where medical records, insurance, and other systems may be at play, and/or (2) together with other models which rely on the same principles of transparency, privacy, and accuracy. The challenge will be to balance what kind of data, how much of it, and in what ways will be share, from one stage of intervention (e.g., onset risk assessment) to the next (diagnosis, prognosis, treatment management). A related issue pertains to explainability, in the sense that it may prove difficult to disclose to users what type of data will be used, and in what ways, if it is not known exactly what that data are, and how they will be used. Similarly intertwined, privacy and data control have a direct impact on accountability, in that whoever controls a certain amount and type of data can only be held accountable for any privacy breach on the grounds that they are aware of what that data are, or indeed of that they are controlling them at all.

2.2.2 Socio-ethical criticisms of trustworthy AI

In reaction to the contemporary focus on developing AI to be Trustworthy, criticisms have emerged in the literature against this particular aim. The main criticisms can be divided into two categories.

The first category of criticism conveys the view that the concept of Trustworthiness is problematic, because the term “trust” and/or or the purported constituent dimensions – robustness, generalisation, explainability, transparency, reproducibility, fairness, privacy preservation, and accountability – are inadequately defined, or defined in conflicting ways across guidelines (Reinhardt, 2023) (AI, 2023) (Freiman, 2023) (Ryan, 2020).

The second category of critique highlights problems relating to the focus on Trustworthiness overall, and calls for a different and/or broader range of values to be accounted for and pursued (Rathkopf & Heinrichs, 2023) (Kerasidou et al., 2022). On a societal level, it is advisable to account for a multitude of understandings of key terms in Trustworthy AI (“trust” and “trustworthy”, but also “robust”, “transparent” and so on), but also to look beyond those key terms, and account for other values which may or may not be tied to the application AI models per se, but rather to that of interaction with a health-promoting system. Such values include (but not exclusively) the values of autonomy, relationship building and retention, sense of identity, and safety.

2.2.3 AI-PROGNOSIS, ethics and society

The AI-PROGNOSIS project will be guided by a revised version of the Trustworthy development and evaluation framework, defined by four (4) key features.

- (1) *Addressing relevant issues as identified by partners through assessment using the ALTAI framework.* This will include the implementation of tools to measure adherence to key principles in each of the models and applications under development.
- (2) *Allowing for the addition of complementary values to the framework* – separate from “trustworthiness” and/or its constituent items – so to allow the pursuit of value-sensitive design and implementation across a range of domains, particularly in the domains of diagnosis, prognosis, and management of neurodegenerative disease, accounting for a broader range of ethical and social issues of pertinence to key stakeholders.
- (3) *Anticipating and reacting to relevant policy landscape,* to dynamically align with important guidelines and frameworks in this domain (see Section 3, **Table 1**).
- (4) *Implementing mechanisms for oversight* which guarantee the adherence to (1-3) above not only on a model-to-model or application-to-application basis, but also allowing an overarching perspective on the potential ethical and social impacts of the project and its outcomes.

3 Existing guidelines, standards, and regulations

Several frameworks and guidelines, as well as proposals, which highlight the principles for trustworthy AI systems and propose actions and requirements to raise trust between humans and AI systems have been developed and published by researchers, the industry, and policymakers in the recent past. **Table 1** summarises key aspects of a non-exhaustive list of important frameworks and guidelines related to trustworthy AI. This table is adopted from (Thiebes et al., 2021) and modified to include other relevant and important frameworks.

Table 1 Description of important frameworks and guidelines for trustworthy AI practices.

Framework/guidelines	Issued by (<i>in</i>)	Description
Asilomar AI Principles (Future of Life Institute, 2017)	Future of Life Institute (2017)	Describes 23 principles of beneficial AI. The principles are organised into three categories: research issues, ethics and values, and long-term issues.
Montreal Declaration of Responsible AI (Montreal Declaration) (University of Montreal, 2017)	Université de Montréal (2017)	Provides ten ethical principles that promote the fundamental interests of people and groups and, based on these, eight recommendations for the development of responsible AI.

UK AI Code (UK House of Lords, 2017)	UK House of Lords (2017)	Defines five overarching principles for an ethical AI code, intended to position the UK as a future leader in AI.
AI4People (Floridi et al., 2018)	Floridi et al. (2018)	A synthesis of six pertinent frameworks and guidelines, which resulted in five foundational principles for ethical AI. Based on the principles, a set of 20 action points in the four categories assessment, development, incentivisation, and support is proposed.
Ethics Guidelines for Trustworthy AI (EU TAI Guidelines) (High-Level Expert Group on Artificial Intelligence, 2019)	European Commission Independent High-Level Expert Group on Artificial Intelligence (2019)	Defines four principles of trustworthy AI and based on these derives seven key requirements for achieving trustworthy AI. Further provides an assessment list for the operationalisation of the seven key requirements.
OECD Principles on AI (The Organisation for Economic Cooperation and Development (OECD), 2019)	OECD (2019)	Recommends “five complementary values-based principles for the responsible stewardship of trustworthy AI” (OECD 2019). In addition to the OECD member states, other countries (e.g., Argentina, Brazil, Colombia, Costa Rica, Peru, and Romania) have signed up to follow the OECD principles.
Governance Principles for the New Generation Artificial Intelligence (Chinese AI Principles) (National Governance Committee for the New Generation Artificial Intelligence, 2019)	Chinese National Governance Committee for the New Generation Artificial Intelligence (2019)	Provides a framework and action guidelines for the governance of AI, based on eight principles for the development of responsible AI.
White House AI Principles (Vought, 2020)	White House’s Office of Science and Technology Policy (Vought 2020)	Defines ten principles for stewardship of AI applications and the development of trustworthy AI. These principles are to be considered by US agencies during the development of regulatory and non-regulatory actions on AI.
Proposed Regulatory Framework for Modifications to Artificial Intelligence/Machine Learning (AI/ML)-Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback (United States Food & Drug Administration, 2019)	U.S. Food and Drug Administration (FDA) (2019)	Describes the FDA’s foundation for a potential approach to premarket review for artificial intelligence and ML-driven software modifications

The Assessment List for Trustworthy Artificial Intelligence (ALTAI) (High-Level Expert Group on Artificial Intelligence, 2020)	European Commission Independent High-Level Expert Group on Artificial Intelligence (2020)	Provides an initial approach for the evaluation of trustworthy AI based on seven high level requirements of trustworthy AI
Artificial Intelligence Act proposal (European Parliament (Council of the European Union), 2021)	European Parliament, Council of the EU (2021)	The proposal presents a balanced and proportionate horizontal regulatory approach to AI that is limited to the minimum necessary requirements to address the risks and problems linked to AI, without unduly constraining or hindering technological development or otherwise disproportionately increasing the cost of placing AI solutions on the market.
Artificial Intelligence Risk Management Framework (AI RMF 1.0) (National Institute of Standards and Technology, 2023)	National Institute of Standards and Technology (2023)	Provides outcomes and actions that enable dialogue, understanding, and activities to manage AI risks and responsibly develop trustworthy AI systems. The Core is composed of four functions: GOVERN, MAP, MEASURE, and MANAGE

In this deliverable, the basis of the procedural trustworthy AI framework will be established based on the above-mentioned ALTAI tool, which is a cornerstone of the upcoming EU AI Act. In the following sections, a thorough review will be presented for the ALTAI requirements. Additionally, a comprehensive overview of two other significant regulations, namely the AI Act and the EU-MDR that will be allied along with the ALTAI requirements, is presented, forming a well-structured AI trustworthy framework.

3.1 ALTAI requirements

During the past years, the European Commission created the "High-Level Expert Group on Artificial Intelligence" (HLEG-AI) that published, in 2019, the "Ethics Guidelines for Trustworthy AI" (High-Level Expert Group on Artificial Intelligence, 2019). These guidelines set out a framework that presents the EU's strategy for achieving trustworthy AI systems. It is stated that a trustworthy AI system has three main components:

- (1) It should be **lawful**, complying with all applicable laws and regulations. Several legally binding rules at both European and national level already apply or are relevant to the development, deployment and use of AI systems. Besides horizontally applicable rules, other domain-specific rules exist that apply to AI applications (such as the MDR in the healthcare sector and of course in the AI-PROGNOSIS case). The law provides rights as well as obligations. That means that these guidelines aim to offer guidance on fostering and securing the other two components (ethical and robust AI).
- (2) It should be **ethical**, ensuring adherence to ethical principles and values. Ethical norms and social cheques require adequate adoption and the alignment with the legal cheques as well. The ethical principles that should be taken into consideration are:
 - Respect for human autonomy
 - Prevention of harm
 - Fairness

- Explicability

- (3) It should be **robust** both from a technical and social perspective. AI systems tend to be characterised as “egoistic” due to their extraordinary potential but even more so because of the excessive confidence people have in their abilities. The improvement of many different aspects of human’s life should be faced with a parallel awareness of the possible unintentional harm that those systems may cause. Robustness, security, and reliability should be ensured as well as safeguards should be foreseen to prevent any unintended adverse impacts. These requirements must be ensured both from a technical as well as a social perspective.

In addition, the HLEG-AI translated these components into seven concrete requirements that an AI system should take into consideration to achieve trustworthiness. These requirements are listed below and should be continuously evaluated throughout the whole AI life cycle as presented in **Figure 1**.

1. Human agency and oversight (HAO)
2. Technical robustness and safety (TRS)
3. Privacy and data governance (PDG)
4. Transparency (TPR)
5. Diversity, non-discrimination, and fairness (DnDF)
6. Societal and environmental wellbeing (SEW)
7. Accountability (ACC)

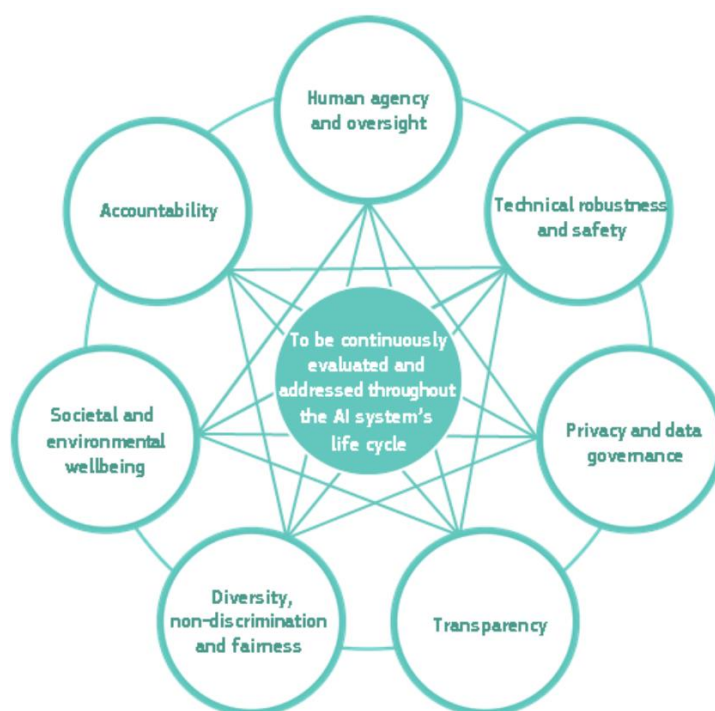


Figure 1 The seven trustworthiness requirements published by the HLEG-AI (High-Level Expert Group on Artificial Intelligence, 2019).

In the second half of 2019, the guidelines published by the HLEG-AI entered through a piloting phase, in which feedback was given by technical and non-technical stakeholders. Following this consultation period, in 2020, the HLEG-AI extended their guidelines and provided a self-assessment tool for the evaluation of the trustworthiness of an AI system. They published a document that contains the ALTAI (High-Level Expert Group on Artificial Intelligence, 2020) for self-assessment. These supplementary guidelines set out the process to be followed in

order to successfully proceed with an assessment on the development and operational characteristics of AI systems, with specific references to the procedures to be followed for any of the aforementioned requirements.

Furthermore, to show the functionality of this assessment list, the Vice-Chair of the AI HLEG and partners from the Insight Centre for Data Analytics at University College Cork implemented the ALTAI into a prototype web-based tool. This tool allows organisations to evaluate the trustworthiness of their AI system by answering a checklist online. This checklist is in the form of prompts and questions related to the seven guidelines published by the HLEG-AI and is aimed at providing practical assistance to AI developers and deployers by allowing them to easily evaluate the trustworthiness of their system. The output of the ALTAI web-based tool is a spider-diagram, as shown in **Figure 2**, assigns weights for each of the seven guidelines based on the organisation's responses during the self-assessment, as well as a list of the recommended steps the organisation can follow to improve their score.

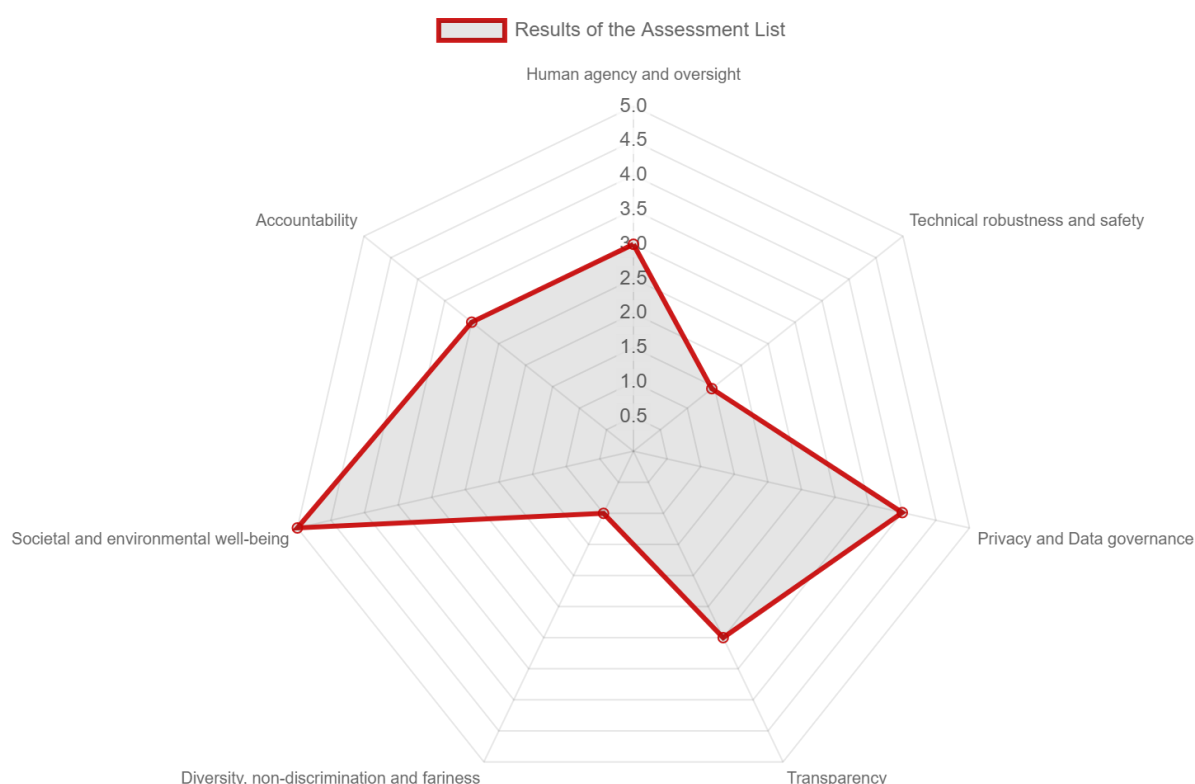


Figure 2 An example of the spider diagram of the ALTAI web-based self-assessment tool.

It should be noted that the weights produced by the spider diagram reflect the number of recommendations. The HLEG, however, has not yet published how every response to the ALTAI questions influences the calculation of the overall weights presented in the spider diagram (Rajamäki et al., 2023).

3.1.1 Overview of the seven HLEG-AI requirements within the ALTAI tool

In this section, a comprehensive coverage of the sub-components addressed by ALTAI is presented, with each sub-component linked to the seven main requirements outlined earlier by HLEG-AI in Section 3.1 above.

3.1.1.1 Human agency and oversight (HAO)

The components of the HAO section aim to evaluate how thoroughly the subject organisation has utilised actions to address the impact of AI systems on human behaviour in various

aspects including human decision-making processes, change of human perception and expectation when AI systems emulate human behaviour, and the impact on human emotions, trust, and autonomy. Additionally, the current requirement also refers to the human oversight factor which is also foreseen in the AI Act proposal and refers to the ability of the human to intervene in every decision cycle of the system. Therefore, some HAO components help organisations evaluate the level of consideration of oversight measures using governance mechanisms like human-in-the-loop, human-on-the-loop, or human-in-command approaches.

Respect to the fundamental rights of the human beings during the operation process of the AI systems is of outmost importance. The developers and providers of the AI systems shall ensure that there will be no violation of the fundamental rights of their users and that the related provisions will be respected throughout the life cycle of the system. Respect to the human dignity is particularly ensured by providing all the necessary information to the user for him/her to be able to comprehend and interact with the AI system to a satisfactory degree, and by having full awareness that they are interacting with such systems, in order to be able to reach autonomous decisions regarding the AI system. The user shall be in a position to be able to reach to a decision following his/her intentions and thought process and not end up relying solely to the automatic processing operations of the AI system. The HAO components are listed in **Table 2**.

Table 2 Components of the human agency and oversight (HAO) requirement of the ALTAI.

HOA components
Incorporate a process where end-users and/or subjects are adequately made aware that an AI-system influenced the decision, content, advice, or outcome.
Ensure that the end-users or subjects are adequately informed that they are interacting with an AI system.
Put in place procedures to avoid that end users over-rely on the AI system.
Put in place any procedure to avoid that the system inadvertently affects human autonomy.
Take measures to deal with the possible negative consequences for end-users or subjects in case they develop attachment. In particular, provide means for the user to have control of the interactions.
Take measures to minimise the risk of addiction by involving experts from other disciplines such as psychology and social work.
Take measures to mitigate the risk of manipulation, including providing clear information about ownership and aims of the system, avoiding unjustified surveillance, and preserving autonomy and mental health of users.
Give specific training to humans (human-in-the-loop, human-on-the-loop, human-in-command) on how to exercise oversight.
Establish detection and response mechanisms in case the AI system generates undesirable adverse effects for the end-user or subject.
Deploy a “stop button” or procedure to safely abort an operation when needed.
Take oversight and control measures to reflect the self-learning or autonomous nature of the AI system

3.1.1.2 Technical robustness and safety (TRS)

The TRS requirement is prioritised by the HLEG-AI in terms of the number of components which reflects its importance for building trustworthy AI systems. This requirement deals with four main subsections that are the security, safety, accuracy; and reliability, fall-back plans and reproducibility. Components of the TRS requirement focus on utilising strategies to mitigate risks of intentional and unintentional harm caused by the AI system. These actions

aim to increase the technical robustness of the system when exposed to adversarial changes, and to improve the dependability of the AI system by ensuring that it is delivering services and it is behaving reliably and as intended. The TRS components are listed in **Table 3**.

Table 3 Components of the technical robustness and safety (TRS) requirement of the ALTAI.

TRS components
Assess potential forms of attacks to which the AI system could be vulnerable.
Put in place measures to ensure the integrity, robustness and overall security of the AI system against potential attacks over its lifecycle.
Red-team/pen test the system
Inform users as soon as possible if some new threats are detected.
Define risk, risk metrics and risk levels of the AI system in each specific use case.
Identify the possible threats to the AI system (design faults, technical faults, environmental threats) and the possible resulting consequences.
Assess the risk of possible malicious use, misuse or inappropriate use of the AI system.
Assess the dependency of critical system's decisions on its stable and reliable behaviour.
Plan fault tolerance via, e.g., a duplicated system or another parallel system (AI-based or "conventional").
Develop a mechanism to evaluate when the AI system has been changed enough to merit a new review of its technical robustness and safety.
Put in place measures to ensure that the data (including training data) used to develop the AI system is up to date, of high quality, complete and representative of the environment the system will be deployed in.
Put in place a series of steps to monitor and document the AI system's accuracy.
Consider whether the AI system's operation can invalidate the data or assumptions it was trained on, and how this might lead to adversarial effects (e.g., biased estimators, echo chambers etc.)
Put in place processes to ensure that the level of accuracy of the AI system to be expected by end-users and/or subjects is properly communicated.
Put in place a well-defined process to monitor if the AI system is meeting the goals of the intended applications.
Test whether specific contexts or conditions need to be taken into account to ensure reproducibility.
Put in place verification and validation methods and documentation (e.g., logging) to evaluate and ensure different aspects of the system's reliability and reproducibility.
Clearly document and operationalise processes for the testing and verification of the reliability and reproducibility of the AI system.
Define tested failsafe fallback plans to address AI system errors of whatever origin and put governance procedures in place to trigger them.
Put in place a proper procedure for handling the cases where the AI system yields results with a low confidence score.
Consider potential negative consequences from the AI system learning novel or unusual methods to score well on its objective function.

3.1.1.3 Privacy and data governance (PDG)

The components of the PDG refer to one of the fundamental rights, privacy. Respect for privacy and data protection, quality, integrity, and access to data should be ensured during the development and, most crucially, the operation of the AI system. Compliance with the provisions of the General Data Protection Regulation and the applicable complementary national privacy-related legislation shall be ensured and the developer of an AI system shall comply with all the responsibilities related to the assurance of privacy and personal data integrity. The PDG components encourage organisations to implement actions that ensure the integration of the right to privacy and data protection in the design process of AI systems. The PDG components are listed in **Table 4**.

Table 4 Components of the privacy and data governance (PDG) requirement of the ALTAI.

PDG components
Take measures to consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection.
Consider establishing mechanisms that allow flagging issues related to privacy or data protection concerning the AI system.
When relevant, implement the right to withdraw consent, the right to object and the right to be forgotten in the AI system.
Consider the privacy and data protection implications of data collected, generated or processed over the course of the AI system's lifecycle.
Consider the privacy and data protection implications of the AI system's non-personal training-data or other processed non-personal data.
Whenever possible and relevant, align the AI system with relevant standards (e.g., ISO, IEEE) or widely adopted protocols for (daily) data management and governance.

3.1.1.4 Transparency (TPR)

The TPR requirement and its components mainly deal with the general idea of the explainability dimension within trustworthy AI systems, as defined in Section 2.1.3. The first element of this requirement is traceability, in which organisations are encouraged to consider a proper documentation strategy about methods used to design and develop the algorithmic system, the methods used to test and validate it, as well as the outcomes of the algorithmic system. The second element is explainability, that is tackled by ensuring that the decisions of the AI system are well understood by those directly and indirectly affected by it. The third element is communication, focusing on the importance of informing users about the limitation of AI-based decisions. The TPR components are listed in **Table 5**.

Table 5 Components of the TPR requirement of the ALTAI.

TPR components
Consider adopting measures to continuously assess the quality of the input data to the AI system.
Consider adopting adequate logging practices in place to record the decision(s) or recommendation(s) of the AI system
Consider explaining the decision adopted or suggested by the AI system to its end users.
Consider continuously surveying the users to ask them whether they understand the decision(s) of the AI system.
In case of interactive AI system, consider communicating to users that they are interacting with a machine.

Establish mechanisms to inform users about the purpose, criteria and limitations of the decision(s) generated by the AI system

3.1.1.5 Diversity, non-discrimination and fairness (DnDF)

As TRS, the DnDF requirement contains a significant number of components that are set by the HLEG-AI that are aimed at fostering fairness and inclusivity within AI systems. The components of the DnDF requirement focus on the importance of utilising strategies to mitigate bias in input data and AI algorithm design. Moreover, DnDF components aim to prioritise user inclusivity by ensuring that AI products or services are accessible to individuals of all ages, genders, abilities, or characteristics. The DnDF components are listed in **Table 6**.

Table 6 Components of the diversity, non-discrimination, and fairness (DnDF) requirement of the ALTAI.

DnDF components
Consider establishing a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design.
Consider diversity and representativeness of end-users and/or subjects in the data.
Test for specific target groups or problematic use cases.
Research and use publicly available technical tools, that are state-of-the-art, to improve your understanding of the data, model and performance.
Assess and put in place processes to test and monitor for potential biases during the entire lifecycle of the AI system (e.g., biases due to possible limitations stemming from the composition of the used data sets (lack of diversity, non-representativeness)).
Consider diversity and representativeness of end-users and or subjects in the data.
Put in place educational and awareness initiatives to help AI designers and AI developers be more aware of the possible bias they can inject in designing and developing the AI system.
Depending on the use case, ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system.
You should establish clear steps and ways of communicating on how and to whom such issues can be raised.
Identify the subjects that could potentially be (in)directly affected by the AI system, in addition to the (end)-users.
Your definition of fairness should be commonly used and should be implemented in any phase of the process of setting up the AI system.
Consider other definitions of fairness before choosing one.
Consult with the impacted communities about the correct definition of fairness, such as representatives of elderly persons or persons with disabilities.
Ensure a quantitative analysis or metrics to measure and test the applied definition of fairness.
Establish mechanisms to ensure fairness in your AI system.
You should ensure that the AI system corresponds to the variety of preferences and abilities in society.
You should assess whether the AI system's user interface is usable by those with special needs or disabilities or those at risk of exclusion.
You should ensure that Universal Design principles are taken into account during every step of the planning and development process, if applicable.

DnDF components
You should take the impact of the AI system on the potential end-users and/or subjects into account.
You should assess whether the team involved in building the AI system engaged with the possible target end-users and/or subjects.
You should assess whether there could be groups who might be disproportionately affected by the outcomes of the system.
You should assess the risk of the possible unfairness of the system onto the end-user's or subject's communities.
You should consider a mechanism to include the participation of the widest range of possible stakeholders in the AI system's design and development.

3.1.1.6 Societal and environmental wellbeing (SEW)

HLEG-AI considered SEW as a requirement for building a trustworthy AI system. SEW is composed of three elements. First, the environmental well-being element, which focuses on the potential impacts of the AI system on the environment. Second, the "Impact on Work and Skills" element, that involves components that focus on the influence and usage of the AI system within work environments. Furthermore, the SEW requirement covers the "Impact on Society at Large or Democracy" element that considers the effects of the AI system on institutions, democracy, and the overall fabric of society. The SEW components are listed in **Table 7**.

Table 7 Components of the societal and environmental wellbeing (SEW) requirement of the ALTAI.

SEW components
Consider the potential positive and negative impacts of your AI system on the environment and establish mechanisms to evaluate this impact.
Define measures to reduce the environmental impact of your AI system's lifecycle and participate in competitions for the development of AI solutions that tackle this problem.
Inform and consult with the impacted workers and their representatives but also involve other stakeholders. Implement communication, education, and training at operational and management level.
Take measures to ensure that the work impacts of the AI system are well understood on the basis of an analysis of the work processes and the whole socio-technical system.
Take measures to counteract de-skilling by means of continuous training, especially in areas sensitive in terms of safety and security.
Provide training opportunities and materials for re- and up-skilling measures.
Assess the societal impact of the AI system's use beyond the (end-)user and subject, such as potentially indirectly affected stakeholders or society at large.
Take actions to minimise potential societal harm of the AI system.
Take measures that ensure that the AI system does not negatively impact democracy.

3.1.1.7 Accountability (ACC)

The ACC requirement set by HLEG-AI contains components to establish mechanisms that facilitate the system's auditability, focusing on reporting and minimising negative effects, as well as components for training and education to help develop accountability practices and establishing processes for third parties or workers to report such vulnerabilities, especially in applications affecting fundamental rights and safety-critical areas. Furthermore, the components highlight the risk management element of the requirement by encouraging

organisations to consider reporting on actions or decisions contributing to outcomes and to manage the consequences, especially for those directly or indirectly affected by the AI system. The ACC components are listed in **Table 8**.

Table 8 Components of the accountability (ACC) requirement of the ALTAI.

ACC components
Establish mechanisms that facilitate the AI system's auditability (e.g., traceability of the development process, the sourcing of training data and the logging of the AI system's processes, outcomes, positive and negative impact).
Ensure that the AI system can be audited by independent third parties.
Foresee any kind of external guidance or third-party auditing processes to oversee ethical concerns and accountability measures.
Organise risk training for developers and deployers to inform them about the potential legal framework applicable to the AI system.
Establishing an AI ethics review board or a similar mechanism to discuss the overall accountability and ethics practices, including potential unclear grey areas.
Establish a process to discuss and continuously monitor and assess the AI system's adherence to this Assessment List for Trustworthy AI.
Establish a process for third parties (e.g., suppliers, end-users, subjects, distributors/vendors or workers) to report potential vulnerabilities, risks or biases in the AI system.
Ensure that redress-by-design mechanisms are put in place for applications that can adversely affect individuals.

3.2 AI Act

In terms of the creation of an AI-related regulation, there is currently a draft proposal of the AI Act (European Parliament (Council of the framework European Union), 2021) that is expected to be finalised in 2024. The AI Act aims at the implementation of specific measures and principles, highlighting the roadmap to a trustworthy AI by proposing at the same time a robust legal framework for the use and embodiment of AI in a more accurate, homocentric, and social-friendly way. Although a two-year period will be given to the AI developers and users to comply to the provisions of the regulation, in the context of AI-PROGNOSIS, the use of the AI technology and its development/function during the following years, will comply by design to the AI related legislation. Taking into consideration the forementioned factors, the proposed AI Act provides an insight into the to-be-adopted legislation and the measures that need to be implemented in order to ensure compliance with it.

The Commission puts forward the proposed regulatory framework on AI with the following specific objectives:

- Ensure that AI systems placed/used on the EU market are safe and respect existing law on fundamental rights and EU values.
- Ensure legal certainty to facilitate investment and innovation in AI.
- Enhance governance and effective enforcement of existing law and safety requirements applicable to AI systems.
- Facilitate the development of a single market for lawful, safe and trustworthy AI applications and prevent market fragmentation.

Moreover, in Article 5 of the proposed Regulation, the prohibition of the following specific AI practices is foreseen:

- I. The placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm.
- II. The placing on the market, putting into service or use of an AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group in a manner that causes or is likely to cause that person or another person physical or psychological harm.
- III. The placing on the market, putting into service or use of AI systems by public authorities or on their behalf for the evaluation or classification of the trustworthiness of natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics, with the social score leading to either or both of the following:
 - i. Detrimental or unfavourable treatment of certain natural persons or whole groups thereof in social contexts which are unrelated to the contexts in which the data was originally generated or collected and
 - ii. Detrimental or unfavourable treatment of certain natural persons or whole groups thereof that is unjustified or disproportionate to their social behaviour or its gravity.
- IV. The use of 'real-time' remote biometric identification systems in publicly accessible spaces for the purpose of law enforcement, unless and in as far as such use is strictly necessary for one of the following objectives:
 - i. the targeted search for specific potential victims of crime, including missing children.
 - ii. the prevention of a specific, substantial, and imminent threat to the life or physical safety of natural persons or of a terrorist attack and
 - iii. the detection, localisation, identification, or prosecution of a perpetrator or suspect of a criminal offence referred to in Article 2(2) of Council Framework Decision 2002/584/JHA 62 and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least three years, as determined by the law of that Member State.

In addition, in article 6 and those that follow, the classification of an AI system is regulated, and certain requirements and procedures are envisaged for the trustworthiness of a high-risk AI system to be ensured. The establishment of a risk management system and data governance and management practices is envisaged, the maintenance of technical documentation and operation logs is foreseen, and the active involvement of humans during the use of the AI systems is noted, through the provision of the necessary information to the users and the possibility of humans to oversee its operation. The principles of accuracy, robustness and respect to security are also highlighted, with references to the measures to be implemented in order to ensure compliance with them.

Following this, obligations to the providers, the users, and other parties related to the use or distribution of the AI technology, such as importers, manufacturers etc. are set out. In article 16 of the proposal, it is envisaged that providers of high-risk AI systems shall:

- i. ensure that their high-risk AI systems are compliant with the requirements set out in the articles of the proposal, to which we refer in the previous paragraph,

- ii. have a quality management system in place,
- iii. draw-up the technical documentation of the high-risk AI system,
- iv. when under their control, keep the logs automatically generated by their high-risk AI systems,
- v. ensure that the high-risk AI system undergoes the relevant conformity assessment procedure, prior to its placing on the market or putting into service,
- vi. comply with the registration obligations referred to in Article 51 of the proposal,
- vii. take the necessary corrective actions, if the high-risk AI system is not in conformity with the requirements set out in Chapter 2 of the proposal,
- viii. inform the national competent authorities of the Member States in which they made the AI system available or put it into service and, where applicable, the notified body of the non-compliance and of any corrective actions taken,
- ix. affix the CE marking to their high-risk AI systems to indicate the conformity with this Regulation in accordance with Article 49 of the proposal.
- x. upon request of a national competent authority, demonstrate the conformity of the high-risk AI system with the requirements set out in Chapter 2 of the proposal.

The ability to provide any information needed regarding any factor of the operation of AI systems is considered of outmost importance. The forementioned documentation with several other provisions such as the duty of information of Article 22 of the proposal in case of a known risk of the AI system, consist of the basic and obligatory procedures for the provider and the manufacturer of the AI system to provide all the necessary information both to the regulatory authorities and the users of the system.

Summarising, although the AI Act has not yet been published, we expect its adoption in 2024. But until then, we can only assume on the eventually implemented requirements regarding the use of AI and the overall procedures, that should be adopted. Those should not merely be indicated but required to be observed in view of assistance and promotion of the human activity in favour of privacy preservation.

3.3 Regulatory compliance and risk management for AI in medical devices

High-risk AI systems related to products, such as medical devices, are described by the New Legislative Framework (NLF). For the AI-PROGNOSIS medical devices, the requirements for AI systems set out in this proposal will be checked as part of the existing conformity assessment procedures under the relevant “Medical devices - Regulation (EU) 2017/745 (MDR)” (European Parliament (Council of the European Union), 2017). If a medical device that operates on AI or contains AI elements falls within risk class IIa or higher (the classification of medical devices is based on their intended purpose and their inherent risks, as set out in Annex VIII of the MDR) and, as a result, a notified body is to be involved in the conformity assessment procedures, this medical device is a high-risk AI system within the meaning of the AI ACT.

The rules concerning how the independent notified bodies that assess the conformity of medium and high-risk medical devices before they are placed on the market, have become stricter and are designated, organised and monitored.

As far as the proposal for the AI Act is concerned, given that a medical device is considered as high-risk AI system, there are some specific requirements with which compliance should be ensured. Among others:

- A risk management system shall be established, documented, and maintained which shall consist of a continuous iterative process run throughout the entire lifecycle of the high-risk AI system, requiring regular systematic updating.
- Data and data governance principles should also be considered, as well as the continuously updated technical documentation that should be drawn up before that system is placed on the market or put into service.
- Another aspect that is set as a prerequisite for the trustworthiness of the high-risk AI systems, is the record keeping of logs during the operation of the AI system. These logging capabilities shall conform to recognised standards or common specifications. These capabilities shall ensure a level of traceability of the AI system's operation throughout its lifecycle that is appropriate to the intended purpose of the system.
- Transparency, accuracy, robustness, and provision of information for users is also needed so that AI systems are concise, complete, correct and easily accessible to users.
- Human supervision during the period in which the AI system is in use aiming at preventing or at least minimising the risks to health, safety and fundamental rights that may emerge when a high-risk AI system is used in accordance with its intended purpose or under conditions of reasonably foreseeable misuse. It is important to note that human supervision should be ensured before the AI system is placed on the market or put into service to prevent incidents like automation bias. All principles and requirements should be translated into obligations for providers and users of high-risk AI systems, to ensure their compliance with them and to be able to contact the relevant notified bodies whenever necessary. The forementioned applies in a less extent to importers and distributors.

Compared to the MDR, the requirements of the proposal for the AI Act introduce additional requirements, so it is quite expected that this will result as a considerable burden for developers of such technologies in the conformity assessment procedure. There are already numerous requirements and obligations in the regulation for the medical devices, which include among others, requirements for risk management systems and documentation. The overlap of requirements though between the MDR and the new AI regulation in some places is to be resolved by subjecting the safety risks of AI systems to the requirements of the draft, while the safety of the product as a whole will be assessed under the MDR.

More specifically in article 43 of the proposal of the AI ACT, various conformity assessment procedures for high-risk AI systems are regulated. Providers of high-risk AI systems must prove that they meet the requirements according to the regulation. The provider must therefore follow the relevant conformity assessment procedures under MDR. The Notified Bodies designated under the MDR shall also be entitled to check the conformity of high-risk AI systems with the requirements of the draft Regulation and thus be able to constitute a “notified body” within the meaning of the draft Regulation. As a result, it is still necessary to carry out only one single conformity assessment procedure for AI medical devices in accordance with the requirements of the MDR, which must additionally ensure compliance with the requirements of the draft regulation. However, notified bodies must have sufficient internal competence to effectively assess the tasks performed by external bodies on their behalf. In this respect, all parties involved must ensure at an early stage that, in addition to the requirements of the MDR for medical devices – and in particular the software that falls under them - those of the future regulation that exceed these are observed. It must be checked whether the AI used meets the safety requirements, while internal processes and products are adapted accordingly.

Comparing the MDR and the AI ACT, it is evitable that there are some key differences between the two regulations and that there are certain topics that the proposed AI Act strives to address. The classification of the MDR to different level of risk according to the intended purpose of the medical device is regulated differently in the AI Act, which generally considers devices regulated in the MDR to be “high-risk”. Also, the conformity assessment remains a prerequisite but should additionally be carried out before, during and after the development of the high-risk AI system. Moreover, human oversight as a risk management measure on high-risk medical devices is provided in the AI Act to ensure that the quality of data is relevant, representative, free of errors and complete.

3.4 ML technologies in medical devices

ML AI systems are those that imitate humans by not acting independently of human reasoning, but instead utilise previously validated clinical protocols to diagnose medical conditions or deliver therapy. They do not think for themselves in the sense of understanding, making judgements, or solving problems, rather they are static rules-based systems, programmed to produce specific output based on the values of received inputs. While these systems can be very sophisticated, the rules they employ are static - they are not created or modified by the systems. There are however other types of AI that utilise large data sets and complex statistical methodologies to discover new relationships between inputs, actions, and outcomes. These data-driven or ML systems are not explicitly programmed to provide pre-determined outputs, but are heuristic, with the ability to learn and make judgements.

In sort, ML AI systems, unlike simple rules-based systems, are cognitive in some sense and can modify their outputs accordingly.

There are also some harmonised standards without specific reference to ML. The MDR allow proof of conformity to be provided with the aid of harmonised standards and common specifications such as:

- ISO 13485:2016 (ISO, 2016)
- IEC 62304 (IEC, 2006)
- IEC 62366-1 (IEC, 2015)
- ISO 14971 (ISO, 2019)
- IEC 82304 (IEC, 2016)

These standards include requirements that are also relevant for medical devices with ML.

To conclude, since the AI Act remains to be published, it can be predicted that there will be certain unwanted consequences as a result of its implementation, and more specifically, it can be assumed that there will be an overlap of requirements between the AI Act and the MDR. The MDR already requires, inter alia, cybersecurity, risk management, post-market surveillance, a notification system, technical documentation, a quality management system. Manufacturers will soon have to demonstrate compliance with both regulations as well as face the inconsistencies between them.

3.5 Responsible AI pillars in the industry

It is important to mention the AI pillars of leading companies in the industry that utilise AI, including IBM¹, Google², and Meta³. These companies present their responsible AI principles

¹ IBM. IBM Artificial Intelligence Pillars - IBM Policy

² Google. Google Responsible AI Practices – Google AI

³ Meta. Responsible AI - AI at Meta

with distinct emphases, but share common ethical themes as those in the EU's ALTAI framework. IBM focuses on explainability, fairness, robustness, transparency, and privacy, underlining the importance of trust, transparency, and ethical principles at the core of AI development. Google emphasises a human-centred design approach, consideration of adverse feedback, diverse user engagement, and rigorous testing, highlighting practical steps towards fairness, safety, and usability. Meta outlines pillars including privacy and security, fairness and inclusion, robustness and safety, transparency and control, and accountability and governance, stressing the need for systems that are equitable, secure, and accountable. These pillars collectively underscore a commitment to developing AI that is ethical, secure, fair, and transparent, mirroring ALTAI's focus on trustworthiness within its seven concrete requirements explained in Section 3.1.

4 The AI-PROGNOSIS framework for AI development and evaluation

Building on the concrete ALTAI requirements presented in Section 3 and the general definitions of AI trustworthy dimensions addressed in Section 2, this section establishes the framework (guidelines, methods, and tools) to be adopted by research and development teams to create the AI components of the project, ensuring trustworthiness. The main idea is to put in practice the concepts of HAO, TRS, PDG, TRP, DnDf, SEW, and ACC of the HLEG-AI for each key stage of the AI lifecycle in the project: 1) Design and specification, 2) Data preparation, 3) Development and validation, 4) user experience (UX) / user interface (UI) and deployment, 5) External validation, and 6) Overarching management and workflow. It is worth noting that, based on the suggestion of the HLEG, the ALTAI is intended for flexible use, and organisations can draw on elements relevant to their particular AI system. Therefore, the ALTAI components were reviewed in terms of their relevance to the products AI-PROGNOSIS aims to develop and the relevant components are summarised in **Table 9**, associated with key stages in the development lifecycle of AI models within the project.

Table 9 ALTAI requirements and summary of relevant components associated with key stages in the development lifecycle of AI models within AI-PROGNOSIS.

AI system lifecycle	ALTAI requirement	Summary of relevant ALTAI components
Design and specification	HAO	Implementation of protocols avoiding over-reliance by end users. Establishing processes to prevent inadvertent effect of the AI system on human autonomy. Enabling human oversight human-in-the-loop, human-on-the-loop, and human-in-command).
	TRS	Definition of risk metrics and levels. Establishing procedures for monitoring and documenting the system's accuracy, ensuring clear communication of expected accuracy to end-users or subjects.
	PDG	Consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection. Align the AI-system with relevant standards or protocol for data management and governance.
	TPR	Identification of user needs regarding explainability of the AI system's decision(s).

AI system lifecycle	ALTAI requirement	Summary of relevant ALTAI components
	DnDF	Utilisation of publicly available technical tools to enhance the understanding of data, model, and performance. Identification of potential biases throughout the AI system's lifecycle. Definition of group/individual fairness goals and of protected attributes. Establish mechanisms for flagging bias, discrimination, or low performance issues.
	ACC	Enabling and facilitating the AI system's auditability.
Data preparation	TRS	Ensuring the high quality of the data
	PDG	Application of data privacy mechanisms for data collected, generated or processed over the course of the AI system's lifecycle
	DnDF	Avoid, correct, and monitor unfair bias
Development and internal validation	TRS	Assessing potential vulnerabilities of the AI system. Implementing robust measures throughout the system's lifecycle to ensure its robustness and overall security. Ensuring the reproducibility of the developed AI models. Performance benchmarking, including reliability assessment through uncertainty and generalisation evaluation.
	PDG	Model assessment in terms of data leaking, missing data, and membership/attribute inference vulnerabilities.
	TPR	Explanation of the decision by the AI system to its end users (explainability benchmarking).
	DnDF	Assessing and mitigating bias throughout the model development phase.
UX/UI and deployment	HAO	Ensuring end-users or subjects are informed about the AI system's influence on decisions or outcomes. Establishing mechanisms to detect and address any adverse effects generated by the AI system on end-users or subjects (Adversarial testing in early deployment).
	TRS	Notification of users upon detecting new threats. Evaluation of the dependency of critical system decisions on its stable behaviour.
	PDG	Enabling end-users or subject to have the right to withdraw consent, object, and be forgotten in the AI system.
	TPR	Ensuring that the end-users understand the decision(s) of the AI system.
	DnDF	Assessing the system's user interface for usability by individuals with special needs, disabilities, or those at risk of exclusion is essential.
External validation	TRS	Performance evaluation, including reproducibility and reliability evaluation
	TPR	Evaluation of quality and adequacy of model output explanations by end-users in clinical studies
	DnDF	Diverse participant sample in clinical studies for external validation of performance and user acceptance evaluation.
Overarching management and workflow	TRS	Assessment of the risk of possible malicious use, misuse or inappropriate use of the AI system.
	PDG	Supporting AI governance strategies. Establishing mechanisms to flag issues related to privacy or data protection.

AI system lifecycle	ALTAI requirement	Summary of relevant ALTAI components
	SEW	Ensure the understanding and assessment of the impacts of the AI system on work, skills, and society at large as part of the general socio-ethical implications assessment.

Within the following subsections, specific components corresponding to each stage of AI development are presented. These components are the AI PROGNOSIS components and are designed to be integrated into the project. They contain actions aimed at building a trustworthy AI system that meets the outlined ALTAI requirements (summarised in **Table 9**) as well as of other pertinent trustworthy AI frameworks, such as the AI Act and the MDR. In the following, methods and steps described are linked to their associated ALTAI requirement(s) using light blue-shaded labels, e.g., **TPR**.

4.1 Design and specification

Some of the specifications that influence the design for creating a trustworthy AI system must be decided during the early co-creation and user research phase. These design considerations pertain to pillars of trustworthy AI such as human autonomy, explainability, fairness, biases, and equality. They will be embedded into the AI-enabled components as early as possible, when the ecosystem is still malleable, and will serve as a guide to adhere to for the next phases of development.

4.1.1 Embedding personal autonomy and exercising oversight

Trustworthy AI systems must ensure human autonomy and support human decision-making of the end user while being overseen by human agents in an end-to-end manner. In AI-PROGNOSIS these design specifications become a necessity due to the sensitive nature of healthcare. Personal autonomy and system oversight must be embedded into the system's specifications as early as possible.

4.1.2 Explainability tailored to each user

To guarantee transparent AI-enabled components and to ensure a user-centric system, the explanation of each model's output must be tailored to the background knowledge of each user group. AI-PROGNOSIS has a diverse range of stakeholders and users meaning special attention must be given to the means of tailoring the explainability experience by considering each user's type (person with PD, person at risk of PD, informal caregiver healthcare professional/enabler) individual needs and capabilities.

4.1.3 Identify fairness goals and protected attributes

Fairness heavily depends on the context of each user. The question "Fair to whom?" can lead to contradictory answers. To create a fair AI system, clear goals for what constitutes it fair to groups and individuals must be set. The goals must be explicit and agreed upon by all stakeholders by taking in account the perspective of the users. This requires the participation of a diverse group of people to the fairness goal identification phase. In addition to fairness goals, any protected attributes, such sex and race, must be also identified for avoiding possible biases and unfair AI enabled components.

4.1.4 Equal access to stakeholders

Ensuring equitable access and engagement of all stakeholders to the design process is vital for a high-quality result in co-creation phase. As a result, the principle of equality must be integrated into the specification and design phase.

4.1.4.1 Methods

Research groups that are involved in AI-enabled components will lead their respective specification and design phase. Then, they will receive crucial feedback from dedicated co-creation workshops and by primary user research that will:

- Evaluate how AI-enabled components impact human decision-making and autonomy **HOA**.
- Identify the end-users' requirements for AI components output that are explainable and interpretable **HOA / TPR**.
- Verify and, when necessary, direct research groups to maintain a Human-in-Command system design approach **HOA**.

During the evaluation for impact on human decision-making, a discussion topic will be any overreliance to the AI-enabled components by the end-users. In addition, by utilising results of primary user research, a tailored explainable experience to the knowledge base of every user type will be embedded into the AI-PROGNOSIS ecosystem. Finally, AI-PROGNOSIS will exercise oversight by following a Human-in-Command approach by trying to augment and/or enhance health care enabler/provider capabilities. This design approach will be affirmed and verified during co-creation workshops. Research groups will use the feedback produced by the workshops to stay on track.

A wide range of information sources will be used by research groups to set fairness goals and identify the protected attributes that can add bias to AI-PROGNOSIS ecosystem. The goals will be based on the feedback from the diverse and inclusive co-creation workshops. During these workshops the protected attributes will be identified based on the EU Fundamental Rights Agency guidelines for non-discrimination (FRA, 2019). In addition, primary user research and their definition of fairness will be considered. Finally, clinicians will provide their invaluable input on the variables that can add bias based on their previous experience **TPR**.

To ensure equal access to stakeholders and users, co-creation groups will involve people with diverse background, such as developers, expert patient groups and clinicians that are experts in their relevant domain. During the workshops, they will be included and empowered in the conversation. Committed to inclusivity, this assemblage will provide to the research groups their perspective and feedback, fostering collaboration while embracing the contributions of each individual **DnDF**. For fairness goals to reflect reality, these conditions for the co-creation workshop are crucial and mandatory.

4.2 Data preparation

4.2.1 Outlier detection and data sanitisation

Outlier detection may be applied by leveraging statistical methods such as the z-score, which measures the difference in standard deviations of an observation from the mean (a threshold, such as $z=3$, can be used to detect potential outliers) or the interquartile range (IQR), which identifies outliers based on the first and third quartiles. Any observation outside the $1.5 * IQR$ can be considered an outlier. Data cleansing and sanitisation may be implemented by the following:

- Removal of duplicate records
- Removal of noise from data by removing implausible ranges of values
- Removal of noisy data altogether
- Removal of records with missing values
- Imputation of missing values
- Standardisation of data formats such as dates, times, categories, units of measurements

- Fill of missing values using interpolation, clustering, and/or regression techniques
- Employment of ML sanitisation methods such as Isolation Forests, One-class support vector machine, auto-encoders.

The result of technical robustness of the described processes will be quantified by measuring and comparing one of or a number of the following metrics: model accuracy, L2 distance, F1 score, correlation (r), coefficient of determination (r^2), and significance (p) between models trained on the original and the cleaned versions of the data.

4.2.2 Data pseudonymisation or anonymisation

Data will be pseudonymised in order to preserve the privacy of the participants in the clinical studies. This process ensures that each subject will be represented by a study id, making it hard to be identified and/or mapped to their data.

4.2.3 Data quality assessment and bias identification

Differential privacy might be employed by leveraging Google's differential privacy libraries⁴. If applied, privacy threshold Epsilon (ϵ) and loss of privacy Delta (δ) parameters will be tuned accordingly to not degrade model predictions while ensuring the privacy of the data.

In house solutions or pre-existing libraries, such as AI Fairness 360⁵ or YData⁶ may be used to assess data quality and bias, focusing on identification and evaluation of origin, inclusivity, type of information and appropriateness, missing data, time frame and geographical coverage, as per the EU Fundamental Rights Agency. If applicable, bias mitigation algorithms may be employed, such as optimised pre-processing (Calmon et al., 2017). If needed, bias will be measured by employing metrics, such as the disparate impact (Feldman et al., 2015) and the equal opportunity (Hardt et al., 2016).

If necessary, synthetic data may be created to combat severe class imbalances using methods such as (Z. Zhang et al., 2021) (Tucker et al., 2020), that will be evaluated based on Fidelity (realism of synthetic samples), diversity (whether the variability of real data is reflected) and authenticity (number of copies of real samples generated) (R. J. Chen et al., 2021). *Synthetic data generation, as well as any approach that will transform/generate data, will be adopted after consultation with regulatory bodies in order not to impede regulatory approval of the predictive models in the future.*

4.3 Development and internal validation

4.3.1 Training for generalisation

These components aim at achieving a generalised model **TRS**. One approach to mitigate overfitting and enhance generalisation involves reducing model complexity. For example, adding a bottleneck layer to a neural network can be effective in this regard (Li et al., 2023). Furthermore, regularisation techniques (explicit or implicit regularisations) can be implemented, such as early stopping (Yao et al., 2007), batch normalisation (Ioffe & Szegedy, 2015), dropout (Srivastava et al., 2014), data augmentation, and weight decay (Krogh & Hertz, 1991). These techniques help improve model generalisation and aim to limit the model's complexity and guide learning toward a manageable hypothesis space.

4.3.2 Adversarial training and regularisation

During the development of the AI models in AI-PROGNOSIS, adversarial training, which is considered as a defensive method against adversarial attacks (Bai et al., 2021), will be taken

⁴ OpenMined. PyDP. <https://github.com/OpenMined/PyDP>

⁵ Trusted-AI. AIF360. <https://github.com/Trusted-AI/AIF360>

⁶ ydataai. ydata-quality. <https://github.com/ydataai/ydata-quality>

into consideration **TRS**. One approach begins with identifying potential adverse attacks, then improve the robustness of the model by augmenting the training data with adversarial samples to create a defence against them (Wang et al., 2019). For this purpose, tools will be adopted like Adversarial Robustness Toolbox⁷, that contains a large set of adversarial attacks and defences that are relevant to the models that will be developed, and it is not limited to deep learning. Although not updated recently, CleverHans⁸ will be considered to evaluate/protect models against adversarial threats of evasion, poisoning, extraction, and inference.

4.3.3 Performance benchmarking

In AI-PROGNOSIS we aim to exhaustively monitor and validate the AI models' performance **TRS**. Tools like the TensorFlow Model Analysis TFMA⁹ can be used to perform deep analysis of the AI model's performance and for performing model evaluation across different slices of data. Fit-for-purpose performance metrics will be used to assess the quality of AI models with respect to accuracy, calibration, and uncertainty. Depending on the different type of models implemented, different accuracy metrics will be adopted.

- For PD risk predictive modelling, key accuracy metrics for healthcare predictive models will be used, such as the concordance index (C-index), joined by classification metrics (i.e., classification accuracy, sensitivity, specificity, area under the receiver operating characteristics curve (AUC)).
- For PD progression modelling, two different approaches are considered, 1) prediction of time to disease milestone and 2) prediction of disease progression scores at certain future time points. In the former case, metrics such as the C-index, Kaplan-Meier curves, and log-rank tests will be used to assess accuracy, while in the latter, measures of error, such as the Root Mean Squared Error (RMSE), will be employed.
- Performance metrics for the PD medication response predictive model will be tailored to the nature of the model, regressor or classifier, that will depend on the nature of the model's target, i.e., categorical (categorised treatment effectiveness) or numerical (treatment effectiveness indicator), to be decided.

For calibration, the calibration intercept and slope should be estimated, complimented, if necessary, by the prediction interval coverage probability for regressors, or the expected calibration error or Brier score for classifiers. For uncertainty, the prediction interval or the prediction confidence score should be estimated.

On top of the aforementioned performance metrics, reliability will be assessed with uncertainty and calibration metrics, as well as performance metrics evaluated on external datasets (out of distribution data and various shifts) and the datasets collected in the validation clinical studies of the project (AI-PRA and AI-PMP studies).

4.3.4 Explainable model design

This component highlights one of the main dimensions of trustworthy AI which is the explainability of the AI system **TPR**. In the project, we will investigate approaches for the two aspects of ML explainability that are mentioned in Section 2.1.3, i.e., testing of inherently explainable models and rendering black box approaches interpretable through both local (explanation of individual outputs for end-users) and global (general model behaviour explanation for regulators and healthcare enablers) explanations. Other approaches can be employed (depending on the model) that combine both aspects in developing and embedding explainable linear models and prototype selection to deep learning models (Li et al., 2023).

⁷ Trusted-AI. *adversarial-robustness-toolbox*. <https://github.com/Trusted-AI/adversarial-robustness-toolbox>

⁸ Cleverhans-lab. *cleverhans*. <https://github.com/cleverhans-lab/cleverhans>

⁹ Tensorflow. https://github.com/tensorflow/tfx/blob/master/docs/tutorials/model_analysis/tfma_basic.ipynb

For this purpose, several actively maintained explainability tools will be explored, such as: (1) SHAP (SHapley Additive exPlanations)¹⁰ that provides a unified measure of feature importance, allowing the understanding of how individual features contribute to model prediction; (2) LIME (Local Interpretable Model-agnostic Explanations)¹¹ that helps explain the predictions of black-box models by generating locally faithful explanations; and (3) Captum¹² that can be used if deep learning models will be used. Furthermore, a variety of studies have used qualitative metrics to evaluate explainability with human participation. Representative approaches include the subjective human evaluation. The methods of evaluation include interviews, self-reports, questionnaires, and case studies that measure, e.g., user satisfaction, mental models, and trust (Hoffman et al., 2018). The goal in this context is to explore how key stakeholders perceive and comprehend model explanations in decision-making processes.

4.3.5 Uncertainty benchmarking

This component focuses on addressing uncertainty during the development of ML models **TRS**. A fail-safe system to be implemented to handle the ML model faults by means of recognising predictions that have low confidence score and should not be trusted and that should activate a healing procedure bringing the model to a safe state. There are two main types of uncertainty (Hüllermeier & Waegeman, 2021), i.e., reducible epistemic uncertainty that arises from the lack of knowledge about the perfect model (usually due to noisy and/or lack of enough training data) and irreducible aleatoric uncertainty, caused by the inherent randomness in the data generation process (e.g., label ambiguity). Predictive uncertainty is the aggregate of those two types. Depending on model type and stage of development (new or pre-trained model), we will estimate uncertainty either intrinsically, through models that inherently provide it (either epistemic or both) along with predictions (such as Bayesian approaches), or via extrinsic methods, extracting uncertainty *post-hoc* (such as Meta-Models). For classification models, uncertainty is expected to be communicated as a confidence score, while for regression models, it will be a prediction interval. Calibration metrics will be employed depending on model type, i.e., such as the Expected Calibration Error or Brier score (for classifiers) or the Prediction Interval Coverage Probability (for regressors). Tools such as Uncertainty Quantification 360¹³ will be used when applicable. It is a toolkit that provides a broad range of capabilities to streamline as well as foster the common practices of quantifying, evaluating, improving, and communicating uncertainty in the AI application development lifecycle. This tool contains a lot of assessments (*post-hoc* and intrinsic/during training) and metrics, and it is not limited to deep learning. It allows to perform a recalibration if the performance is poor. We will also adopt approaches, similar to previous studies (Becker et al., 2021), that use the “defer” to human option for handling the use cases with low confidence.

4.3.6 Assessment and mitigation of bias

As for the **DnDF** components of AI-PROGNOSIS, we will integrate bias mitigation techniques and incorporate feedback mechanisms to adapt to changing conditions and user needs. We intend to develop an AI system with minimum bias by conducting fair training and avoiding inequalities due to the discriminatory forms caused by biases. Fair training methods will be considered:

- (1) At the level of modifying training data before feeding into the model (Pre-processing Fairness) using implementations like re-sampling (balance representation of different groups), re-weighting (assign higher weights to underrepresented groups), and data

¹⁰ SHAP. <https://github.com/shap/shap>

¹¹ LIME. <https://github.com/marcotcr/lime>

¹² Captum. <https://github.com/pytorch/captum>

¹³ IBM. UQ360. <https://github.com/IBM/UQ360>

augmentation (generate synthetic data to increase representation) (Calmon et al., 2017) (Sun et al., 2022).

- (2) At the level of modifying learning algorithms or objective functions (in-processing training) using implementations like adversarial training, and adversarial debiasing (Wan et al., 2023) (B. H. Zhang et al., 2018).
- (3) At the level of adjusting the model's predictions after training (post-processing fairness) using implementations like re-ranking (adjust model predictions to ensure equalised odds) and calibration (calibrates model's predicted probabilities) (Petersen et al., 2021) (Putzel & Lee, 2022).

In this context, and during the development of ML models in the project, actively maintained libraries/tools that can help detect and mitigate unwanted bias in the developed models will be utilised such as the well documented AI Fairness 360, which provides a wide range of metrics, algorithms, and tutorials to help detect and mitigate bias in AI models, and FairML¹⁴, which offers various metrics for evaluating the fairness of ML models, such demographic parity, a fairness metric used to assess whether the positive outcome is distributed equally among different demographic groups.

4.3.7 Production model - performance and privacy risk trade-off

During the development of ML models in the project, if there is a risk that the model, even if accurate, might compromise privacy, leading to concerns about data leaks, re-identification of individuals, employing techniques like differential privacy can help minimise the exposure of sensitive data while still enabling model training or inference. The aim of this is to balance the trade-offs between performance and privacy risks.

4.4 UX/UI and deployment

4.4.1 User feedback

User feedback about the outputs of an AI model can help discover biases and incorrect predictions and understand the usefulness and adoption of an AI application. In order to assure that the feedback received is effective, it is crucial to decide which users need to provide it, an external domain expert or the end users.

Another factor that should be considered is the type of user feedback that is needed. Feedback can be explicit or implicit. Explicit feedback is the one that the user intentionally provides. It can be quantitative, where the user answers a simple question of correct/incorrect, gives a rating in a scale out of 5, or qualitative where the user provides written feedback freely, mostly through a text box. Explicit feedback is best if the aim is increasing the model performance. On the other hand, implicit user feedback refers to unintentional provided data based mostly on users' behavioural patterns (Hodgson, 2023).

General user feedback will be collected on the AI model's performance with options to verify, dispute, and correct the AI outputs **HAO**.

4.4.2 Adversarial testing

Adversarial ML involves incorporating a simulated adversary to assess and enhance the effectiveness of a ML system at various stages of its development and deployment lifecycle. This encompasses activities like training (such as data collection, model selection, and tuning), model testing (including vulnerability assessment and performance benchmarking), hardware implementation, system integration, and ongoing monitoring and updates (P.-Y. Chen & Hsieh, 2023). The goal of adversarial ML is to evaluate how well a model performs when it is

¹⁴ FairML. <https://github.com/adabayoj/fairml>

exposed to intentionally crafted adversarial examples—input data that has been specifically designed to mislead or deceive the model. Using tools, such as the Adversarial Robustness Toolbox, to test and evaluate the robustness of the ML models in an early development stage, can help eliminate adverse effects generated by the AI system on end-users **HAO / TRS**.

4.4.3 Disclaimers

A disclaimer is a statement or notice intended to limit the legal liability or responsibility of the person or entity providing the information. It is a way to inform others that the information presented may not be complete, accurate, or applicable in all situations, and that the provider is not assuming full responsibility for any consequences that may arise from relying on the information.

Disclaimers are commonly used in various contexts, such as websites, contracts, product labels, and written materials, to address issues like potential errors, omissions, or the specific conditions under which the information is valid. The goal is to protect the party providing the information from legal claims and to make users or consumers aware of the limitations associated with the provided content.

Within the project, disclaimers can be used to help increase transparency, mitigate risk and reduce liability and misinterpretation by the users. Those will include:

- Disclaimer on the expected model performance **TRS**.
- Disclaimer on AI function and required input of personal data **PDG**.
- Disclaimer regarding unavoidable bias **DnDF**.

4.4.4 Input data validation

Data validation is an important process when conducting research and it is crucial for ensuring the accuracy and completeness of the data. Inconsistent or incomplete data most of the times can be proven of little use. It is best that data are constructed so no time and resources are wasted in cleaning and transforming.

There are ways that data validity can be achieved within AI-PROGNOSIS:

Data type, completeness, and structure check: Before utilising the data, model providers should ensure that the data received have the desired type and structure and there are no missing values. If there are missing values, the data should be either populated or discarded altogether.

Standardised user input: The user should input data through standardised procedures, mostly through answering multiple choice questions and disabling submission of the input before it is complete. Fields with free text typing should be avoided since they are very prone to inconsistencies **TRS**.

4.4.5 Data minimisation

The principle of data minimisation means that the collection and use of personal data should be limited to the complete necessary to fulfil a specific task and only that. The retention of those data should be also limited to only the specific amount of time that is required for the completion of that task¹⁵ **PDG**.

4.4.6 Right to be forgotten

The right to be forgotten, also known as the right to erasure, allows individuals to request the deletion of their personal data when it is no longer necessary for the purpose for which it was collected, or when they withdraw their consent. Data controllers (organisations or entities

¹⁵ https://edps.europa.eu/data-protection/data-protection/glossary/d_en

collecting and processing personal data) are obligated to respond to erasure requests within a reasonable timeframe and to communicate the decision to the data subject **PDG**.

4.4.7 Explanation of models and their outputs

With AI being widely adopted in human lives, there is a need to understand the reasoning behind the decisions made by the AI models, especially if the outcomes of models that are treated as black boxes are affecting human lives (Goodman & Flaxman, 2017). Black box is an AI system that is so complex that it is very difficult, if not impossible, to explain the decision-making process and its outcomes, resulting sometimes in unwanted bias in the results (Björklund et al., 2023).

As already mentioned, a method to solve this problem is Explainable AI (XAI) which provides explanations of the model's outputs and presents them in an easy-to-understand way for humans. It is important to understand the decisions made by machines for humans, so fairness, equality, transparency and accountability is assured (Scantamburlo et al., 2019).

From a UI/UX perspective, the challenge lies in effectively communicating the complex information generated by XAI methods, such as LIME or SHAP, to users. Traditional data visualisation libraries might not suffice for presenting the intricate data produced by XAI algorithms. Therefore, in AI-PROGNOSIS, solutions, based on XAI libraries, will be investigated to create visual reports which can then be seamlessly integrated into mobile or web applications. Tailored, in terms of nature and form (i.e., text, graphs, statistics), in-app model output explanations and elements for communicating uncertainty will eventually be developed, with the aim of ensuring that the explanations provided are accessible and understandable to each user group, enhancing the overall experience and trust **TPR**.

4.4.8 Educational content on AI

Creating and providing tailored educational content about how the underlying AI can give users a better understanding how the technology works behind the scenes and make it more accessible and less intimidating for users who may not have a technical background **TPR**.

4.4.9 UI/UX best practices and performance optimisation

UI and UX best practices are guidelines and principles that help designers create effective, user-friendly, and aesthetically pleasing digital products. Adopting such practices is crucial in the overall success of a product, including collection of input data and communication of AI outputs and related disclaimers. Those include:

- Format consistency throughout the app so users do not get confused.
- Hierarchy and readability of the components so users can easily be guided through the content, by prioritising important elements using visual cues like font size, colour and spacing.
- Accessibility factors such as colour contrast and text size to make the app easier to use by people with disabilities.
- Performance optimisation by reducing loading times and resource usage. This can be achieved by limiting data transmissions to the minimum and closing any unnecessary connections.

Depending on the development platform, the relevant UI/UX principles (including those for accessibility) can be adopted as the design basis, i.e., Google's Material Design¹⁶ or Apple's

¹⁶ Google Material Design. <https://material.io/design>

Human Interface Guidelines¹⁷. Those principles will be validated during co-creation sessions and user test feedback **DnDF**.

4.5 External validation

4.5.1 Diversity, non-discrimination and fairness in clinical studies

In AI-PROGNOSIS, the mitigation of bias, although potentially complex and challenging, will be collaboratively addressed by both data scientists and clinical experts. Both groups will establish specific guidelines and strategies pertaining to both the ML training datasets and the datasets derived from clinical studies, which will be employed to evaluate the developed ML models.

Efforts will be made to include a diverse participant sample in the AI-PRA and AI-PMP clinical studies for external validation of performance and user acceptance evaluation of the AI tools. In this vein, the datasets derived from clinical studies must exhibit diversity and representativeness of the population, with careful consideration given to the selection of features. Demographic elements, such as age and gender, will be subject to examination, given the variation in symptoms of PD among individuals of different demographics. Additionally, the various stages of PD can serve as another criterion for the diversity of the datasets. Within AI-PROGNOSIS, co-creation workshops will be conducted to gather crucial feedback from clinical partners, aiding in the identification of privileged/unprivileged groups and relevant features. The clinical experts' feedback is also essential to the model's bias evaluation. Appropriate groups of datasets must be selected, tested, and verified against biases **DnDF**.

4.5.2 Observability for robust clinical data validation

Observability is a critical aspect of ML, especially when it comes to external validation using clinical data. It involves the capability to gain a comprehensive understanding of a model's performance through monitoring and alerting. This goes beyond simple debugging, as it plays a key role in providing insights into metrics that signal potential issues within a model operating in a real-world clinical setting.

In the context of clinical data, observability becomes crucial for detecting anomalies such as data drifts and concept drifts. Data drifting occurs when there is a shift in the statistical properties of input data, causing the model to make inferences from values that deviate from the original training datasets. Meanwhile, concept drifting involves a change in the relationship between features and the dependent variable over time. Both phenomena can result in model degradation and compromise the reliability of ML pipelines within clinical applications.

To address these challenges in AI-PROGNOSIS, different tools and services can be utilised. One such tool is Evidently AI¹⁸, which employs statistical algorithms to detect anomalies, providing a valuable means of identifying deviations in clinical data. It allows maintaining constant logging of model predictions, and an alert system needs to be in place to facilitate an efficient process of model re-evaluation and retraining (Shankar & Parameswaran, 2022). This emphasis on observability ensures the ongoing reliability and accuracy of ML models in the dynamic change of clinical data **TRS**.

¹⁷ Apple Human Interface Guidelines. <https://developer.apple.com/design/human-interface-guidelines>

¹⁸ Evidently AI. <https://www.evidentlyai.com/>

4.5.3 End-user assessment of model output explanations in clinical studies

In AI-PROGNOSIS, transparent documentation of the developed ML models will ensure that clinical practitioners and researchers will have access to the necessary information to assess the reliability and relevance of the models' outcome within the context of the project.

As already mentioned, different XAI outputs can be generated for the same model depending on the end user, whether it is a patient, a healthcare professional, or even an ML engineer who is required to test and validate their model. In the AI-PROGNOSIS project, co-creation workshops will be conducted to develop guidelines on how to effectively explain models, ensuring a shared understanding among stakeholders. Probing further, in the prospective AI-PRA and AI-PMP clinical studies for external validation of performance and user acceptance evaluation of the AI tools, evaluation of quality and adequacy of AI model output explanations will be requested by participants, including healthcare professionals **TRS**.

4.6 Overarching management and workflow

The elements that are described in this section are related to the general procedures needed to constitute a system trustworthy that are not strictly related to any lifecycle phase of the system but more to management of by-products and processes of AI-PROGNOSIS. Methods that will be used are described under each section and they are accompanied by their respective ALTAI requirements.

4.6.1 Model Operations

Machine Learning Operations (MLOps) is an emerging paradigm that combines DevOps with ML. It formulates a set of principles and best practices in ML development, monitoring, and deployment to be used in production. By following a set of procedures and concepts, such as having reproducible models, versioning (models, data and code), continuous training/evaluation and logging, ML workflow can be optimised. Some key aspects of MLOps that are related to trustworthy AI are:

- Model performance monitoring
- Output logging
- Data, model and code versioning
- MLOps platform selection to ensure security

The establishment of a model performance monitoring mechanism during the development lifecycle can enable model tracking and expedite the generation of model performance reports. By keeping a log-record of model inferences (output logging), model performance can be analysed, and can result in AI enabled components that can be audited. Moreover, reproducible ML models are crucial for a trustworthy AI system as they can uphold the verification and validity of the results and behaviour, help error detection, and facilitate identification and addressing of biases. For these key aspects to be accomplished, it is imperative to utilise a secure and robust platform to assure data integrity and confidentiality.

4.6.1.1 Methods

The MLFlow platform¹⁹ will be utilised to facilitate MLOps procedures. It is a highly regarded open-source platform that, among other features, offers model performance monitoring, logging, data and model versioning. The MLFlow project further has a system of vulnerability reporting to safeguard their application programming interface (API) or notify users of alternative means to secure their application.

¹⁹ MLFlow - <https://mlflow.org/>

Model performance logging during validation will be performed by utilising MLFlow logging. Cases of underperformance will be flagged and scrutinised by research groups **TRS**.

All the models produced in the AI-PROGNOSIS project will reside in a model registry along with information regarding their validation metrics. This type of model versioning will ensure reproducibility but also facilitate tracking the development process by keeping previous iterations of models with their respective validation metrics **TPR / ACC**.

Output logging will be performed in all models used in the production environment **TRS**.

Git version control system will be used to track all code produced in AI-PROGNOSIS project **TPR / ACC**.

4.6.2 Model and dataset reporting

For an ML model to be considered transparent, it is necessary to provide a detailed reporting during its lifecycle. In addition, reports that can be accessed by end users and third parties are elevating the accountability of the system. The essential reports pertain to:

- Datasets used
- Model development workflow
- Model validation

4.6.2.1 Methods

Data used for training AI-PROGNOSIS ML models will be sourced mostly from restricted datasets. In contrast, data used for validating AI-PROGNOSIS ML models will be products of the project. In absence of standardised means of documentation on datasets, dataset reports and documentation will be produced based on “datasheets for datasets” (Gebru et al., 2021) were possible with emphasis in datasets used in validation **TPR / ACC**.

Reports for each ML model developed and validated in AI-PROGNOSIS project will be generated based on the TRIPOD checklist²⁰ **TPR / ACC**.

4.6.3 Open models and datasets

Opening datasets and ML model architectures brings several tangible benefits such as:

- Transparency and responsible AI
- Validation and reproducibility
- Research advancement

In order to be aligned with ethical and trustworthy AI development practices, models and datasets should be published and made open. This, fosters trust, transparency and accountability. In addition, open models/validation datasets ensure the reproducibility of the results. Finally, innovation can be fostered as researchers with diverse background can use the results for collective advancement and knowledge. However, it is imperative to safeguard the system and ensure security, privacy, protection of intellectual property rights (IPR), and mitigate potential risks of future exploitation. This means that a careful balance between transparency and securing sensitive components of the system should be maintained.

4.6.3.1 Methods

All datasets produced by AI-PROGNOSIS project will be pseudonymised and made freely available after an embargo period. The Zenodo²¹ open science repository will be used. As a

²⁰ TRIPOD checklist - <https://www.tripod-statement.org/resources/>

²¹ Zenodo - <https://zenodo.org/>

result, data archiving and sharing will conform to the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles²² **TPR**.

ML models that are not protected by intellectual property rights will be made available only after ensuring that sensitive data cannot be extracted by model manipulation **TPR / ACC**.

4.6.4 Security and GDPR compliance

All the components of the system must be fully compliant with the relevant regulations and the General Data Protection Regulation (GDPR). This includes all project processes and products including AI components and associated tools. Furthermore, key components of the system must be scrutinised for security.

4.6.4.1 Methods

AI-PROGNOSIS has a dedicated data management team that supervise the data collection, exchange, storing and processing to be compliant with the GDPR. They will ensure full compliance of all processes and tools with the regulatory framework, including the AI components and associated tools **PDG**.

Ethics and data protection impact assessment (deliverables D1.3 and D1.4) will be conducted at strategic time points in the project (month 7 and month 23), before main research and development, as well as data collection phases. Moreover, a data management plan will be produced and maintained through the European Open Science Cloud (EOSC) Argos service²³ to facilitate the compliance of data management and curation with GDPR **PDG**.

Third party datasets that will be accessed by AI-PROGNOSIS consortium members will be handled according to their respective license. Moreover, for the exchange of data, partners will sign data transfer and data use agreements **PDG**.

A privacy-by-design approach will be used while designing the project's digital ecosystem. By being proactive, data de-identification (removal all personal information) will be performed (PDG). In addition, data will be disclosed only to partners that must process them **PDG**. Given that the concept of privacy is intertwined with system security, a security assessment of the components of the AI-PROGNOSIS ecosystem will be performed by the AI-PROGNOSIS consortium **TRS / PDG**. In addition, industry security best practices will be followed such as:

- hosting the Cloud infrastructure with a reputable provider
- having firewalls to protect the Cloud infrastructure
- use of VPN to access hosted resources
- use of Secure Sockets Layer (SSL)/ Transport Layer Security (TLS)/ Hypertext Transfer Protocol Secure (HTTPS) for encrypted communication where applicable
- use of Identity and Access Management agents where applicable

4.6.5 Sustainability and societal impact

Trustworthy AI systems must be evaluated for their sustainability and their environmental footprint. ML models may need significant computational resources for training, validation, deployment and maintenance. In addition, AI tools, such as those AI-PROGNOSIS is developing, could have socio-ethical implications, impacting work, skills, and society at large, as they are linked to healthcare and involve human-computer interaction.

4.6.5.1 Methods

To reduce the environmental impact of the AI-PROGNOSIS ecosystem, including its AI operations, the project has selected the Hetzner Cloud provider that is taking measures to

²² FAIR principles - <https://www.go-fair.org/fair-principles/>

²³ EOSC Argos - <https://argos.openaire.eu/splash/>

protect the environment and has adopted sustainable practices to power its data centres²⁴. **SEW.**

A report on socio-ethical implications of the AI-PROGNOSIS tools will be included in the next version of this deliverable (month 25), when the release of the minimum viable product versions of the tools will be nearing (month 28) **SEW.**

5 Conclusion – key takeaways

To conclude, this version of the deliverable shapes the fundamental trustworthy AI framework for the AI-PROGNOSIS project and it presents a clear understanding of trustworthy AI. The document highlights the socio-ethical considerations and their effects on the AI-PROGNOSIS project. It culminates in detail the development components, mapped across the project's AI lifecycle with alignment to the existing guidelines. **Table 10** summarises the AI-PROGNOSIS components to be followed to ensure that each stage of AI development in adheres to the principles of trustworthy AI.

This fundamental version of the deliverable, particularly the AI-PROGNOSIS components in Section 4, will be updated over the course of the project. Updates will be guided by the actual implementation experiences of various partners in their respective tasks, as well as any changes in relevant regulations. The aim is to further tailor these components specifically to AI-PROGNOSIS, ensuring that the framework remains relevant and effective in practical applications.

Table 10 AI-PROGNOSIS components aligned with ALTAI Requirements for each step in the AI component lifecycle.

AI step	AI-PROGNOSIS component	ALTAI requirement
Design and Specification	Embedding personal autonomy and exercising oversight	HAO
	Explainability tailored to each user	TPR
	Identify fairness goals and protected attributes	DnDF
	Equal access to stakeholders	DnDF
Data Preparation	Outlier detection and data sanitisation	TRS
	Data pseudonymisation or anonymisation	PDG
	Data quality assessment and bias identification (via data distribution metrics)	DnDF
Development and Validation	Training for generalisation (balance between under/ overfitting)	TRS
	Adversarial training and regularisation	TRS
	Performance benchmarking, including reliability assessment	TRS
	Selection of model to enter production based on trade-off between performance and privacy risks	PDG
	Explainable model design	TPR

²⁴ Hetzner Environmental Protection - <https://www.hetzner.com/unternehmen/umweltschutz>

AI step	AI-PROGNOSIS component	ALTAI requirement
	Uncertainty benchmarking	TPR
	Assessment and mitigation of bias	DnDF
UX/UI and Deployment	User feedback	HAO
	Adversarial testing	HAO, TRS
	Disclaimers	TRS, PDG, DnDF
	Input data validation	TRS
	Data minimisation	PDG
	Right to be forgotten	PDG
	Explanation of models and their outputs	TPR
	Educational content on AI	TPR
	UI/UX best practices and performance optimisation	DnDF
External validation	Diversity, non-discrimination and fairness in clinical studies	TPR, DnDF
	Observability for robust clinical data validation	TRS
	End-User assessment of model output explanations in clinical studies	TRS
Overarching Management and Workflow	Model Operations	TRS, TPR, ACC
	Model and dataset reporting	TPR, ACC
	Open models and datasets	TPR, ACC
	Security and GDPR compliance	PDG
	Sustainability and societal impact	SEW

References

- Al, P. (2023). (E)-Trust and Its Function: Why We Shouldn't Apply Trust and Trustworthiness to Human–AI Relations. *Journal of Applied Philosophy*, 40(1), 95–108. <https://doi.org/https://doi.org/10.1111/japp.12613>
- Bai, T., Luo, J., Zhao, J., Wen, B., & Wang, Q. (2021). Recent Advances in Adversarial Training for Adversarial Robustness. *IJCAI International Joint Conference on Artificial Intelligence*, 2, 4312–4321. <https://doi.org/10.24963/ijcai.2021/591>
- Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Benetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58(December 2019), 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- Beauchamp T.L., C. J. F. (2001). Principles of Biomedical Ethics, 5th edn. In *Oxford University Press, USA* (Issue 5). <https://doi.org/10.1136/jme.28.5.332-a>
- Becker, T., Vandecasteele, K., Chatzichristos, C., Van Paesschen, W., Valkenburg, D., Van Huffel, S., & De Vos, M. (2021). Classification with a Deferral Option and Low-Trust Filtering for Automated Seizure Detection. *Sensors (Basel, Switzerland)*, 21(4). <https://doi.org/10.3390/s21041046>
- Binns, R. (2018). Fairness in Machine Learning: Lessons from Political Philosophy. *Proceedings of Machine Learning Research*, 81(2016), 149–159.
- Björklund, A., Henelius, A., & Puolamäki, K. (2023). Explaining any black box model using real data. *Frontiers in Computer Science*, 5. <https://doi.org/10.3389/fcomp.2023.1143904>
- Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., & Varshney, K. R. (2017). Optimized Pre-Processing for Discrimination Prevention. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems* (Vol. 30). Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/9a49a25d845a483fae4be7e341368e36-Paper.pdf
- Caton, S., & Haas, C. (2023). Fairness in Machine Learning: A Survey. *ACM Computing Surveys*. <https://doi.org/10.1145/3616865>
- Chen, P.-Y., & Hsieh, C.-J. (2023). *Adversarial Robustness for Machine Learning* (P.-Y. Chen & C.-J. Hsieh (eds.)). Academic Press. <https://doi.org/https://doi.org/10.1016/B978-0-12-824020-5.00009-0>
- Chen, R. J., Lu, M. Y., Chen, T. Y., Williamson, D. F. K., & Mahmood, F. (2021). Synthetic data in machine learning for medicine and healthcare. *Nature Biomedical Engineering*, 5(6), 493–497. <https://doi.org/10.1038/s41551-021-00751-8>
- Cockburn, A., Dragicevic, P., Besançon, L., & Gutwin, C. (2020). Threats of a replication crisis in empirical computer science. *Communications of the ACM*, 63(8), 70–79. <https://doi.org/10.1145/3360311>
- Dalton-Brown, S. (2020). The Ethics of Medical AI and the Physician-Patient Relationship. *Cambridge Quarterly of Healthcare Ethics*, 29(1), 115–121. <https://doi.org/10.1017/S0963180119000847>
- Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2021). *The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations*. <http://arxiv.org/abs/2107.13509>

- European Parliament (Council of the European Union). (2017). *Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on medical devices, amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EE*. <https://doi.org/10.1177/2165079915576935>
- European Parliament (Council of the European Union). (2021). The EU Artificial Intelligence Act. *European Commission*. <https://doi.org/10.4324/9781003319436>
- Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and Removing Disparate Impact. *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 259–268. <https://doi.org/10.1145/2783258.2783311>
- Floridi, L., Cowls, J., Beltrametti, M., Chatila, R., Chazerand, P., Dignum, V., Luetge, C., Madelin, R., Pagallo, U., Rossi, F., Schafer, B., Valcke, P., & Vayena, E. (2018). AI4People—An Ethical Framework for a Good AI Society: Opportunities, Risks, Principles, and Recommendations. *Minds and Machines*, 28(4), 689–707. <https://doi.org/10.1007/s11023-018-9482-5>
- FRA. (2019). Data quality and artificial intelligence – mitigating bias and error to protect fundamental rights. *FRA Focus*, 18.
- Freiman, O. (2023). Making sense of the conceptual nonsense ‘trustworthy AI.’ *AI and Ethics*, 3(4), 1351–1360. <https://doi.org/10.1007/s43681-022-00241-w>
- Future of Life Institute. (2017). *Asilomar AI Principles*. <https://futureoflife.org/open-letter/ai-principles/>
- Ghojogh, B., & Crowley, M. (2019). *The Theory Behind Overfitting, Cross Validation, Regularization, Bagging, and Boosting: Tutorial*. 3. <http://arxiv.org/abs/1905.12787>
- Gillath, O., Ai, T., Branicky, M., Keshmiri, S., Davison, R., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, 115, 106607. <https://doi.org/10.1016/j.chb.2020.106607>
- Glikson, E., & Woolley, A. W. (2020). HUMAN TRUST IN ARTIFICIAL INTELLIGENCE: REVIEW OF EMPIRICAL RESEARCH. *Academy of Management Annals*, 14(2), 627–660. [https://leeds-faculty.colorado.edu/dahe7472/OB 2022/glickson 2021.pdf](https://leeds-faculty.colorado.edu/dahe7472/OB%2022/glickson%2021.pdf)
- Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision Making and a “Right to Explanation.” *AI Magazine*, 38, 50–57. <https://arxiv.org/abs/1606.08813>
- Hamon, R., Junklewitz, H., & Sanchez, I. (2020). Robustness and Explainability of Artificial Intelligence. In *Publications Office of the European Union*. <https://doi.org/10.2760/57493>
- Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. *Advances in Neural Information Processing Systems, Nips*, 3323–3331.
- High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. High-Level Expert Group on Artificial Intelligence. *European Commission*, 1–39. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419
- High-Level Expert Group on Artificial Intelligence. (2020). The Assessment list for trustworthy artificial intelligence (ALTAI) for self assessment. *European Commission*, 0–33. <https://ec.europa.eu/digital-single-market/en/news/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
- Hodgson, J. (2023). *User Feedback — The Missing Piece of the ML Monitoring Stack*.

- Towards Data Science. <https://towardsdatascience.com/user-feedback-the-missing-piece-of-your-ml-monitoring-stack-46b2bbf0b5e4>
- Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for Explainable AI: Challenges and Prospects*. 1–50. <http://arxiv.org/abs/1812.04608>
- Hüllermeier, E., & Waegeman, W. (2021). Aleatoric and epistemic uncertainty in machine learning: an introduction to concepts and methods. In *Machine Learning* (Vol. 110, Issue 3). Springer US. <https://doi.org/10.1007/s10994-021-05946-3>
- IEC. (2006). *Medical device software — Software life cycle processes*. IEC 62304:2006. <https://www.iso.org/standard/38421.html>
- IEC. (2015). *Medical devices — Part 1: Application of usability engineering to medical devices*. IEC 62366-1:2015. <https://www.iso.org/standard/63179.html>
- IEC. (2016). *Health software — Part 1: General requirements for product safety*. IEC 82304-1:2016. <https://www.iso.org/standard/59543.html>
- Ioffe, S., & Szegedy, C. (2015). Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift. In F. Bach & D. Blei (Eds.), *Proceedings of the 32nd International Conference on Machine Learning* (Vol. 37, pp. 448–456). PMLR. <https://proceedings.mlr.press/v37/ioffe15.html>
- ISO/IEC. (2022). *Trustworthiness – Vocabulary*. ISO/IEC TS 5723:2022. <https://www.iso.org/standard/81608.html>
- ISO. (2016). *Medical devices — Quality management systems — Requirements for regulatory purposes*. ISO 13485:2016. <https://www.iso.org/standard/59752.html>
- ISO. (2019). *Medical devices — Application of risk management to medical devices*. ISO 14971:2019. <https://www.iso.org/standard/72704.html>
- Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. *FACCT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Section 2*, 624–635. <https://doi.org/10.1145/3442188.3445923>
- Joshua A. Kroll, Joanna Huey, Solon Barocas, Edward W. Felten, Joel R. Reidenberg, D. G. R. & H. Y. (2017). Accountable Algorithms. *165 U. Pa. L. Rev.*, September, 633–639. https://scholarship.law.upenn.edu/penn_law_review/vol165/iss3/3
- Julie Gerlings, Millie Søndergaard Jensen, A. S. (2022). Explainable AI, but Explainable to Whom? An Exploratory Case Study of xAI in Healthcare. *Handbook of Artificial Intelligence in Healthcare: Vol 2: Practicalities and Prospects*. https://doi.org/https://doi.org/10.1007/978-3-030-83620-7_7
- Kerasidou, C. X., Kerasidou, A., Buscher, M., & Wilkinson, S. (2022). Before and beyond trust: reliance in medical AI. *Journal of Medical Ethics*, 48(11), 852–856. <https://doi.org/10.1136/medethics-2020-107095>
- Krogh, A., & Hertz, J. A. (1991). A Simple Weight Decay Can Improve Generalisation. *Proceedings of the 4th International Conference on Neural Information Processing Systems*, 950–957.
- Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From Principles to Practices. *ACM Computing Surveys*, 55(9), 1–46. <https://doi.org/10.1145/3555803>
- Lin, Y.-T., Hung, T.-W., & Huang, L. T.-L. (2021). Engineering Equity: How AI Can Help Reduce the Harm of Implicit Bias. *Philosophy & Technology*, 34(1), 65–90.

<https://doi.org/10.1007/s13347-020-00406-7>

Lipton, Z. C. (2018). The mythos of model interpretability. *Queue*, 16(3), 31–57. <https://doi.org/10.1145/3233231>

London, A. J. (2019). Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. *Hastings Center Report*, 49(1), 15–21. <https://doi.org/10.1002/hast.973>

Mahya, P., & Fürnkranz, J. (2023). An Empirical Comparison of Interpretable Models to Post-Hoc Explanations. *AI (Switzerland)*, 4(2), 426–436. <https://doi.org/10.3390/ai4020023>

Murdoch, B. (2021). Privacy and artificial intelligence: challenges for protecting health information in a new era. *BMC Medical Ethics*, 22(1), 122. <https://doi.org/10.1186/s12910-021-00687-3>

National Governance Committee for the New Generation Artificial Intelligence. (2019). *Governance Principles for the New Generation Artificial Intelligence--Developing Responsible Artificial Intelligence*. <https://www.chinadaily.com.cn/a/201906/17/WS5d07486ba3103dbf14328ab7.html>

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. <https://doi.org/10.2139/ssrn.4141546>

Petersen, F., Mukherjee, D., Sun, Y., & Yurochkin, M. (2021). Post-processing for Individual Fairness. *Advances in Neural Information Processing Systems*, 31(NeurIPS), 25944–25955.

Putzel, P., & Lee, S. (2022). Blackbox Postprocessing for Multiclass Fairness. *CEUR Workshop Proceedings*, 3087.

Rajamäki, J., Gioulekas, F., Rocha, P. A. L., Garcia, X. del T., Ofem, P., & Tyni, J. (2023). ALTAI Tool for Assessing AI-Based Technologies: Lessons Learned and Recommendations from SHAPES Pilots. *Healthcare (Switzerland)*, 11(10), 1–20. <https://doi.org/10.3390/healthcare11101454>

Rathkopf, C., & Heinrichs, B. (2023). Learning to Live with Strange Error: Beyond Trustworthiness in Artificial Intelligence Ethics. *Cambridge Quarterly of Healthcare Ethics*, 1–13. <https://doi.org/10.1017/s0963180122000688>

Reinhardt, K. (2023). Trust and trustworthiness in AI ethics. *AI and Ethics*, 3(3), 735–744. <https://doi.org/10.1007/s43681-022-00200-5>

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>

Ryan, M. (2020). In AI We Trust: Ethics, Artificial Intelligence, and Reliability. *Science and Engineering Ethics*, 26(5), 2749–2767. <https://doi.org/10.1007/s11948-020-00228-y>

Safdar, N. M., Banja, J. D., & Meltzer, C. C. (2020). Ethical considerations in artificial intelligence. *European Journal of Radiology*, 122(July 2019), 2019–2021. <https://doi.org/10.1016/j.ejrad.2019.108768>

Scantamburlo, T., Charlesworth, A., & Cristianini, N. (2019). *Machine Decisions and Human Consequences*. Oxford University Press.

Shankar, S., & Parameswaran, A. (2022). Towards Observability for Production Machine Learning Pipelines. <https://arxiv.org/pdf/2108.13557.pdf>

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning*

Research, 15, 1929–1958.

- Sun, Y., Fung, B. C. M., & Haghighat, F. (2022). The generalizability of pre-processing techniques on the accuracy and fairness of data-driven building models: A case study. *Energy and Buildings*, 268, 112204. <https://doi.org/https://doi.org/10.1016/j.enbuild.2022.112204>
- The Organization for Economic Cooperation and Development (OECD). (2019). *OECD AI Principles*. <https://oecd.ai/en/ai-principles>
- Thiebes, S., Lins, S., & Sunyaev, A. (2021). Trustworthy artificial intelligence. *Electronic Markets*, 31(2), 447–464. <https://doi.org/10.1007/s12525-020-00441-4>
- Tocchetti, A., Corti, L., Balayn, A., Yurrita, M., Lippmann, P., Brambilla, M., & Yang, J. (2023). A.I. Robustness: a Human-Centered Perspective on Technological Challenges and Opportunities. *Computing Surveys Manuscript*, 434, 1–5. <https://ejournal3.undip.ac.id/index.php/jamt/article/view/5101>
- Trisha Mahoney, Kush R. Varshney, M. H. (2020). *AI Fairness*. O'Reilly Media, Inc. <https://www.oreilly.com/library/view/ai-fairness/9781492077664/>
- Tucker, A., Wang, Z., Rotalinti, Y., & Myles, P. (2020). Generating high-fidelity synthetic patient data for assessing machine learning healthcare software. *NPJ Digital Medicine*, 3(1), 147. <https://doi.org/10.1038/s41746-020-00353-9>
- UK House of Lords. (2017). *Artificial Intelligence Committee AI in the UK: ready, willing and able?* <https://publications.parliament.uk/pa/ld201719/ldselect/ldai/100/10002.htm>
- United States Food & Drug Administration. (2019). Proposed Regulatory Framework for Modifications to Artificial Intelligence / Machine Learning (AI/ ML) -Based Software as a Medical Device (SaMD) - Discussion Paper and Request for Feedback. *U.S Food & Drug Administration*, 1–20. <https://www.fda.gov/media/122535/download>
- University of Montreal. (2017). *Montreal Declaration for a Responsible Development of AI*. <https://declarationmontreal-iaresponsable.com/>
- Vought, R. T. (2020). *Guidance for Regulation of Artificial Intelligence Applications*. <https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf>
- Wan, M., Zha, D., Liu, N., & Zou, N. (2023). In-Processing Modeling Techniques for Machine Learning Fairness: A Survey. *ACM Trans. Knowl. Discov. Data*, 17(3). <https://doi.org/10.1145/3551390>
- Wang, Y., Ma, X., Bailey, J., Yi, J., Zhou, B., & Gu, Q. (2019). On the Convergence and Robustness of Adversarial Training. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 6586–6595). PMLR. <https://proceedings.mlr.press/v97/wang19i.html>
- Yao, Y., Rosasco, L., & Caponnetto, A. (2007). On Early Stopping in Gradient Descent Learning. *Constructive Approximation*, 26(2), 289–315. <https://doi.org/10.1007/s00365-006-0663-2>
- Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). Mitigating Unwanted Biases with Adversarial Learning. *AIES 2018 - Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, 335–340. <https://doi.org/10.1145/3278721.3278779>
- Zhang, Y., Wu, M., Tian, G. Y., Zhang, G., & Lu, J. (2021). Ethics and privacy of artificial intelligence: Understandings from bibliometrics. *Knowledge-Based Systems*, 222, 106994. <https://doi.org/10.1016/j.knosys.2021.106994>
- Zhang, Z., Yan, C., Lasko, T. A., Sun, J., & Malin, B. A. (2021). SynTEG: a framework for

temporal structured electronic health data simulation. *Journal of the American Medical Informatics Association : JAMIA*, 28(3), 596–604. <https://doi.org/10.1093/jamia/ocaa262>