CREXDATA

Critical Action Planning over Extreme-Scale Data

D1.3 Ethics Manual

Version 1.0

Documentation Information

Contract Number	**101092749**
Project Website	https://crexdata.eu/
Contractual Deadline	M6, 30.06.2023
Dissemination Level	PU-Public
Nature	Ethics
Author	Alexandros Nousias (NCSR)
Contributors	Jens Pottebaum (UPB) Arnau Montagud (BSC) Manolis Kaliorakis (MT)
Reviewer	Miguel Ponce-de-Leon (BSC)
Keywords	Ethics/risk assessment, design phase

CREXDATA has received funding from the European Union's Horizon Europe programme under grant agreement number 101092749.

Change Log

Version	Author	Date	Description Change
V0.1	Alexandros Nousias (NCSR)	22/05/2023	Creation
V0.2	Alexandros Nousias (NCSR)	25/05/2023	Added Section 2
V0.3	Alexandros Nousias (NCSR)	29/05/2023	Added Section 3
V0.4	Alexandros Nousias (NCSR)	31/05/2023	Document Completed and Submitted for Internal Review
V0.5	Miguel Ponce-de-Leon (BSC)	12/06/2023	Version after Internal Review
V0.6	Alexandros Nousias (NCSR)	19/06/2023	Internal Review Comments Incorporated
V1.0	Antonios Deligiannakis (TUC)	29/06/2023	Final Version

Contents

Executive Summary

As the Trustworthy AI domain gradually matures, the focus shifts towards value-based design methodologies that strike a balance between economic growth and societal sustainability. AI systems are considered socio-technical systems, which imply risks and negative impacts at the human and societal level [1]. Public concerns around AI systems need to be addressed and trust to be founded, subject to values, as contextualized in the given space and time. Assessment models that encompass a) human rights and b) ethical and societal issues seem to be necessary in the emerging AI system alignment process. Despite their current complexity, their ambiguity and the resistance they may drive to both technical stuff and the humanities, their inclusion as a component to the AI value chain seems fundamental.

The present document describes the process and methodologies to be followed throughout the CREXDATA lifecycle regarding its impact on health, safety and fundamental rights with a focus on the current design phase. It evaluates the relevant risks for the CREXDATA use cases and provides a manual on how to set the appropriate ethical profile and to identify at a later stage additional measures and safeguards.

The ethics assessment process and methodology includes the following steps as per each use case:

- AI System overview and conceptualisation.
- Socio-ethical and techno-ethical concerns and generated risks thereof.
- High level application of the EU Assessment List for Trustworthy AI.
- Risk classification subject to the Proposal for an AI Regulation.

1 Introduction

1.1 Project Information

CREXDATA is an EU funded project with a focus on developing a Prediction-as-a Service (PaaS) system for real time critical situation management. At a first instance, CREXDATA is a project developed and put into operation for the sole purpose of scientific research and development, where specific technologies from WP3 to WP5 are integrated into technical systems and operational procedures in pilot sites, as further described in D2.1, subject to the data flows, as described in D1.2. As such, at first instance it is out of the scope of the proposed Regulation on AI. However, in view of the adoption of CREXDATA's results in market applications and its overall dissemination and exploitation plan, as per WP6, it requires full compliance with the EU legal and ethical frameworks as shaped to date to:

a. ensure scientific and operational alignment with the EU values and human rights sets retrospectively,
b. identify and mitigate wider socio-technical concerns, if any, and
c. identify risk levels at hand.

CREXDATA solutions will be evaluated on three challenging use cases related to (1) maritime domain for forecasting hazardous situations at sea employing real time sensor data and earth observation data, (2) weather emergency management which is delivering exact terrain information and capturing phenomena in a given, fully protected designated environment, (3) health crisis management to limit pandemic outbreaks and come up with non-pharmaceutical means of patient treatment.

The expected impact of CREXDATA includes:

- Interpretable, verifiable and scalable ML-based proactive analytics and decision-making for humans-in-the-loop and autonomous systems alike
- Robust, resilient solutions in critical sectors of science and industry
- Accurate and timely forecasting in vertical sectors (maritime, weather, life sciences and health)
- Novel FAIR datasets for scientific research
- Novel resources and approaches for verifiable, interpretable, scalable and knowledge-aware machine learning.

1.2 Document Scope

This deliverable describes the process and methodology for the CREXDATA AI ethics assessment as per the use cases, which will be conducted in direct collaboration with all the involved work packages. This assessment is in line with the proposed Regulation on AI[1] and aims to ensure that in view of the adoption of CREXDATA's results in market applications and its overall dissemination and exploitation plan, as per WP6, the project is fully compliant with the EU legal and ethical frameworks as shaped to date, so as:

[1] Art.2.6 as per EU AI Act dated 25 November 2022 as adopted by the EU Council on 6 December 2022.

a. to ensure scientific and operational alignment with the EU values and human rights sets retrospectively,
b. to identify and mitigate wider socio-technical concerns, if any, and
c. to properly identify risk levels.

More specifically, this deliverable assesses whether any ethical concerns, related to human rights[2] and values as well as wider socio-ethical concerns could be raised in the context of the use cases. Following the above-mentioned ethical scrutiny, the deliverable details how the potentially raised issues will be addressed/mitigated, building on the work of the EU High Level Expert Group (HLEG) that has set the principles of trustworthy AI, which apply in three core dimensions, namely a) lawful, b) ethical, and c) technical robustness. Additionally, the deliverable follows an appropriate risk classification, subject to the Proposal for an AI Regulation. The present document refers to the use case-specific phases of the lifecycle of the CREXDATA AI system and the relevant areas of ethical and regulatory interest, from design through development, evaluation and operation. The objective is to anticipate, to the extent possible, the CREXDATA AI system's impact on the complex environments in which the use cases roll out, taking into account a) the identified risk levels and the following hard requirements and governance schema, that derive directly from EU regulation, and b) the relevant soft requirements and governance schema, which are undertaken to the CREXDATA contexts. At the present design phase, the focus lies on defining the problem to solve and conceptualizing it in its use cases. This conceptualization also requires identifying the relevant risks, benefits and metrics to measure success or failure.

1.3 Document Structure

This document is comprised of the following chapters:

Chapter 2 presents the CREXDATA ethics assessment process and methodology analysis at the design level, to demonstrate adherence to the relevant principles and norms. This methodology which is comprised by the following steps: a) the CREXDATA AI system overview as the necessary descriptive component of the ethics assessment and risk classification that is to follow, subject to the system's properties as defined, b) a general ethics assessment with the focus on the data, the model and the output at the design phase of the AI lifecycle, as well as relevant socio-technical concerns, c) application of the ALTAI principles, the most wider accepted EU ethical framework.

Chapter 3 presents a high-level alignment as per the use cases with the requirements of the ALTAI framework[3] and a relevant operationalization scheme as defined following the use cases conceptualisation.

Chapter 4 presents the logic behind the relevant risk classification subject to the Proposal for an AI Regulation and enters into a relevant risk classification subject to the Proposal for an AI Regulation, as per the use cases, so as to ensure legal compliance.

[2] Subject to the Charter and the European Convention on Human Rights (ECHR) its protocols and the European Social Charter.

[3] Assessment List for Trustworthy Artificial Intelligence

2 Ethics Assessment Process & Methodology

2.1 Process & Methodology

Trustworthy AI has three components which should be met throughout the system's entire lifecycle: (1) it should be **lawful**, complying with all applicable laws and regulations, (2) it should be **ethical**, ensuring adherence to ethical principles and values, and (3) it should be **robust**, both from a technical and social perspective since, even with good intentions, AI systems can cause unintentional harm. Each component in itself is necessary, but not sufficient for the achievement of Trustworthy AI and, therefore, all three must be addressed accordingly [2].

Aligned with the European Ethical Assessment in the context of the Horizon Europe Programme, a dedicated AI Ethical Assessment section has been integrated in the CREXDATA AI lifecycle as part of the ethical evaluation. This takes place at the design phase, so as to conceptually ensure respect towards the legal framework including a) AI legal requirements, namely the Proposal for an AI Regulation, the Proposal for AI Liability Directive and the Proposal for (revised) Product Liability Directive and General Product Safety Directive, and b) data legal requirements with the primary focus on the GDPR and due the course of time to data verticals, subject to the European Strategy for Data, regarding the Common European Data Spaces. CREXDATA opted to include in its ethics manual the EU Assessment List on Trustworthy AI (ALTAI), as introduced by the EU High Level Expert Group [2], taking into account that the proposed AI Regulation renders ALTAI from soft ethical requirements into hard law. ALTAI, despite its shortcomings in terms of complexity, lack of specificity, or even met resistance, remains the most commonly accepted EU ethical framework to date. On top of that, the present deliverable provides an additional layer of ethics assessment by examining concerns that may be raised directly due to CREXDATA's socio-technical instances, thus framing the wider socio-ethical and techno-ethical impact of the project in a holistic fashion.

CREXDATA understands AI assessment across the lifecycle of these AI systems. In particular, it will examine the following life cycle system phases: ***(1) Design-phase:*** AI system concept stage including research and design activities; ***(2) Development-phase***: AI system development phase (initial experimentation and validation); ***(3) Deployment-phase***: AI system operationalisation and deployment.

Following the submitted ethics self-assessment, where CREXDATA has conducted an a priori self-assessment as per the use cases, by detailing whether any ethical concerns, aligned with the Horizon Europe template may come at play, an additional socio-ethics assessment was circulated internally, as ethical imperatives are distinct to binding regulatory provisions but no less significant. The proposed AI ethics assessment methodology (quantitative and qualitative assessment), focuses on the design/conceptualization phase of the lifecycle of an AI system, and introduces a four-level approach aiming at:

- mapping the properties of the system as a whole and as per use case (System Overview),
- analysing the socio-technical implications of the AI system by focusing on relevant concerns following a risk-based approach (Socio-Technical Assessment),
- identifying the degree of compliance to the ALTAI principles, and
- following a risk-based classification subject to the Proposal for an AI Regulation (Risk Classification).

Regarding the logic behind ethics assessment at the design phase, Floridi et.al. assert that “*conceptualization in the design phase serves two goals. First, it prevents project misspecification, that is, a situation where the AI system is unreflective of the underlying problem. Second, it facilitates a feasibility assessment, which is a study of the system viability, limitations and trade-offs. Failure to meet any of these goals will result in an AI that malfunctions or unintentionally reinforces existing societal disparities*” [3]. CREXDATA shares the same view and facilitates both project misspecifications and a feasibility assessment via the described ethics assessment process and methodology that has been created in line with the EU legal and ethical imperatives.

2.2 CREXDATA AI System Overview

The proposed CREXDATA system aims at forecasting the occurrence of future events from early signs to support proactive and informed decision making. The system will be rolled out in three use cases, containing both simulation and real time data, subject to historical time series, as follows:

- Weather use case, namely in management of weather induced emergencies rolled out in two scenarios,
- Health crisis management use case, namely by providing efficient parameter exploration forecasting and effective interventions in the modelling of epidemics and drug treatment, and
- Maritime, in particular forecasting imminent vessel collision.

As such potential external stakeholders, namely customers, users, operators, are companies operating in the maritime domain (maritime data providers, port authorities, vessel pilots, maritime shipping companies), entities in civil protection, healthcare authorities and R&D and technology providers in AI/ML (SMEs) research organisations and the Academia. On top of that, there is a lot of CREXDATA’s impact potential at both the level of the individual and the level of the group that may be affected by the use of the employed systems, and the project is very much aware of this.

2.2.1 Identifying the Use Cases and Data Needs

The model design is subject to the given requirements set in the use cases and the set purposes, thus ensuring ‘*fit for purpose*’ contextual information quality. To that end, algorithms combining neural, statistical and symbolic methods for learning and reasoning will be employed and the ensuing models will be run on appropriate input data as per use case as follows:

- **Use case I: Weather emergency**. Local stationary and mobile sensor data (weather stations, aerial/ground vehicles), global data services (meteorological services, Copernicus EMS).
- **Use case II: Health Crisis**. Secondary use of anonymised phone-based daily mobility data socio-economic data, demographic and socioeconomic indicators, network of contacts among humans and time series of case reports for COVID 19 and other infection diseases, obtained from databases via relevant data processing agreements or freely available datasets.
- **Use case III: Maritime**. AIS (terrestrial and satellite) data, environmental data, navigational data from the sensors of the vessel.

The above input data are considered adequate and relevant for the use case concepts to ensure optimal data sourcing and conceptualization. The ethical focus lies on whether these input data do indeed accurately capture the problem at play and the tasks at hand. The project ensures that the predictive features do represent the underlying problem per use case, subject to the set task (micro level) and goal (macro level) and following best practices to ensure qualitative data (see Table 1 below). Similarly, the project ensures that the input data, on the basis of which the system produces its output, do not operate as proxies for other variables (i.e., COVID-19 infections as regional financial status proxy).

Table 1: Data review items for the design phase.

	Use Case I	**Use Case II**	**Use Case II**
Input Data Types	Local stationary and mobile sensor data (weather stations, aerial/ground vehicles), global data services (meteorological services, Copernicus EMS).	Secondary use of anonymised phone-based daily mobility data, demographic and socioeconomic indicators, network of contacts among humans and time-series of case reports for COVID-19 and other infectious diseases.	AIS (terrestrial and satellite) data, environmental data, navigational data from the sensors of the vessel
Point of Reference	Evolution of an emergency from forecasts/preparation through response to recovery	Identify changes in human mobility patterns that help to predict outbreaks of cases during an epidemic process.	AIS (terrestrial and satellite) data, environmental data, navigational data from the sensors of the vessel
Task	Use AI techniques to anticipate hazardous situations and their evolution and to efficiently deploy response resources	Extreme-scale model exploration will be combined with interactive learning approaches to explore the ample space of epidemiological parameters, as well as, to find optimal interventions.	Collisions between vessels at sea Rerouting vessels away from hazardous weather sea areas

	Use Case I	**Use Case II**	**Use Case II**
Goal	Improve situational awareness significantly so that informed decisions are taken by civil protection	Develop platform for simulating and optimizing intervention strategies in different scenarios to support decision-making.	Use AI techniques to predict the imminent collision of between two vessels in the short-term prediction horizon. In a second step, automatically provide a rerouting suggestion to avoid an imminent collision. Use AI techniques to monitor and predict the position of the vessel in relation to hazardous weather conditions and provide rerouting information in order for vessels to avoid sea areas with hazardous weather conditions
Data adequacy	Data is collected either from real sources (like weather data services) or realistic ones. To ensure the latter, realistic environments are setup in research infrastructures (indoor and outdoor). As far as possible, users use their real tools in, e.g., field trials.	Daily Origin-Destinations matrices are a critical element in understanding the spatial dynamics of an epidemic process. Additionally, the time series of reported cases is a good proxy of the real mobility patterns.	Detection and mitigation of forecasted collisions between vessels
Data relevance	Data sources are either provided for the specific purpose dealt with in the project (like data from Copernicus Emergency Management Services) or specifically identified	Data from the COVID-19 pandemic including cases and mobility from Spain used to test and validate models and simulations.	Data reflect real-world conditions/challenges required to successfully and efficiently predict and avoid hazardous maritime events

	Use Case I	Use Case II	Use Case II
	in interviews/workshops with end users (in Dortmund, Innsbruck/Tyrol and Helsinki).		

2.2.2 Technical Properties and Ethical Metrics

The system's output consists of trained neuro-symbolic models that allow for emitting reasoned and documented forecasts in the given contexts as per the use cases. Such forecasts are based on partial pattern matches (i.e., the pattern has not been fully matched yet when the forecast is issued). A forecast in this context is the likelihood that a full match will eventually occur at some point in the future, given an observed partial match, each one in the context of the specific use cases, thus satisfying both a) the purpose specification principle and b) the use limitation principle as originating from the GDPR and is considered a best practice. What is of great importance at this stage is to **identify error metrics** and **measure success** retrospectively [4]. To that end, relevant KPIs will be conceptualised accordingly as the project evolves.

Finally, a benchmarking analysis with existing systems at play is under way, so as to establish baseline metrics in this regard. Such benchmarking involves comparison with purely neural forecasters, trained on prefixes of the input to perform a sort of early classification of the input sequences and purely symbolic forecasters, using hand-crafted patterns only, in cases where it is possible to obtain such patterns using domain knowledge (i.e., without any learning).

Regardless of the '*fit for purpose*' design, the consortium partners are aware that the system could potentially be used in a plurality of contexts (see Section 4.3). CREXDATA is aware that an aspect of this sort could bring into the surface contextual discrepancies that require special ethical treatment regarding the system's output in terms of performance, risk classification, impact and accompanying socio-technical concerns. Such problematic is not applicable at this stage, but the project is aware that relevant transparency measures need to be adopted and communicated accordingly when due, namely at the deployment phase, so as to allow appropriate downstream uses under appropriate configurations thereof, with emphasis on data quality, appropriate data/model governance schema and further legal compliance.

2.3 CREXDATA AI Ethics Assessment

The CREXDATA AI system is a supportive tool that will allow organisations operating in each use case context like civil protection, health authorities or non-pharmaceutical treatments and maritime industry to reach informed decisions. These informed decisions derive from contextually set complex event forecasting as described in Section 2.1.1. The project's guiding values and ethical objectives are safety, inclusion, prevention of harm and human dignity. Organisational governance starts with a set of ethical values that steer the behaviour of developers and managers towards the good of society [5]. CREXDATA reflects nicely on that, as it understands the aforementioned values as a key dimension of its AI system among others like its purpose, as contextualized, its input/output data and its governance scheme.

Following the system's analysis (System Overview), the aim/goal of CREXDATA in the context of all three use cases is '*fit for purpose'* and for the public benefit and interest. Having defined a) the applicable value set, b) the ethical principle set, and c) the problem(s) to solve, having formulated the use cases with their specific tasks and having identified the relevant data needs, the present methodology, aligned with the emerging common practices, examines to the extent possible, concerns regarding the serving values. This examination is subject to the overall CREXDATA context, namely real-time critical situation management including flexible action planning and agile decision making over data of extreme scale and complexity and its sub-contexts as per the specific use cases thereof. Such a risk-based approach, regardless of the classification of the proposed EU AI Act, at the present design phase provides a high-level view in relation to a) the project's impact at the micro and macro socio level, as well as the environment and b) concerns regarding health, safety, fundamental rights and values that may be compromised. This is a fundamental preliminary step towards informed choices at the development phase, regarding training, validation and testing and related ethical requirements, as provided in the ALTAI framework.

2.3.1 Socio-Technical Concerns (High-Level)

In principle, no issues regarding health, safety, the environment and fundamental human rights are at stake at this phase, since the models are designed to be trained by data, generated in use case-specific scenarios and, thus, applied in protected environments. Any relevant risks refer to later phases. However, the high level-risk catalogue below will generate awareness regarding the system's potential trade-offs. Off note, many high-level risks are common for all three use cases, as they refer to typical legal and ethical issues as defined by the present ethical frameworks and the upcoming legal regulatory texts.

2.3.1.1 Use case I: Weather Emergencies

Some high-level potential risks associated with automated management in weather extremes can be summarized as follows:

- Inaccurate or incomplete data: In the event the data collected from weather sensors, satellite imagery, or other sources is inaccurate or incomplete, it can lead to incorrect decisions or inadequate response strategies.
- Technical failures: Automated systems are susceptible to technical failures, such as software glitches, hardware malfunctions, or communication breakdowns. If the system fails during a critical moment, it may impede the response efforts and delay necessary actions.
- False alarms or missed alerts: The automated system's algorithms and thresholds for triggering alerts or emergency responses may not always be perfect. False alarms can create panic or lead to unnecessary evacuations, while missed alerts can result in delayed or insufficient responses, potentially endangering lives and property.
- Lack of human judgment and adaptability: The systems at play operate based on predefined algorithms and rules. They may not possess the ability to assess complex and rapidly evolving situations or incorporate human judgment. This can limit their effectiveness in handling unique or unprecedented emergencies that require flexible decision-making.
- Cybersecurity risks: Automation introduces new cybersecurity vulnerabilities. If the automated management system is not adequately protected, it could be targeted by malicious actors who could manipulate the system, disrupt operations, or gain unauthorized access to sensitive data.

- Technological dependency and single-point failures: Relying heavily on automated systems can create a dependency on the technology as it incorporates a certainty bias. If the system encounters a failure or experiences downtime, it may hinder emergency response capabilities. Moreover, if there is a single point of failure in the system's infrastructure, it could have widespread consequences.
- Public trust and acceptance: The successful implementation of automated systems in emergency management requires public trust and acceptance. Some individuals may be sceptical of automated decision-making, preferring human involvement in critical situations. Building trust and ensuring transparent communication about the capabilities and limitations of automated systems is a crucial and core CREXDATA objective.

2.3.1.2 Use case II: Health Crisis Management

Some high-level potential risks associated with automated health crisis management, namely pandemic outbreaks and treatments can be summarized as follows:

- **In regard to pandemics**, on top of data quality and availability issues, further risks are:
  - Complexity of pandemics: Pandemics are complex, dynamic events, subject to factors like changes in human behavior and mobility due to awareness or the introduction of non-pharmaceutical interventions, the emergence of new virous variants and healthcare system capacity. It is rather uncertain whether automated models are capable of capturing the full complexity or accurately predicting the course of pandemics beyond a very short forecasting horizon.
  - Uncertainty and (inherent) unpredictability: Overreliance without acknowledging the uncertainties may lead to misguided decisions and inadequate response strategies. In connection to the previous identified risk for instance, is impossible to forecast the outcome of a vaccination campaign, or what is the best strategy.
  - Lack of context and nuance: There is a risk of not capturing the local context, cultural factors and specific characteristics of different regions or communities, thus lacking the ability to provide tailored and relevant forecasts.
  - Random ethical considerations: to amplify existing biases or inequalities subject to biased data or flawed assumptions. They can also have unintended consequences (i.e., stigmatizing certain groups or leading to resource allocations disparities.
- **In regard to (non) pharmaceutical treatments:**
  - Lack of personalized care if individual variations in health conditions, preferences or specific needs are not taken into account.
  - Inadequate assessment and diagnosis.
  - Limited human interaction and support, thus lack of human touch and empathy.
  - Data privacy and security concerns as the system collects sensitive health data, regardless the fact that these data are processed in an anonymized

format, as issue of unauthorized access to data, data breaches, misuse and the like may crop up.

2.3.1.3 Use case III: Maritime

Some high-level potential risks associated with automated management in the Maritime use case can be summarized as follows:

- Lack of real time data, namely weather conditions, navigational hazards, traffic patterns, port or cargo information.
- Inadequate consideration of factors at local or hyperlocal level including fishing zones, marine protected areas and the like.
- Limited situational awareness at the hyper local/time level.
- Legal and regulatory compliance as the system may not adhere to local or international laws, regulations and safety guidelines.

2.3.2 Techno-Ethical Concerns

2.3.2.1 Algorithm

No straightforward ethical issues specifically by the AI algorithms that will be used in the project do crop up whatsoever. However, there are, in principle, some broader (i.e., not project-specific) ethical concerns that could be raised, mainly due to the fact that CREXDATA relies on state-of-the-art (SoA) deep learning training algorithms. It is well known that the SoA in the field is currently incapable of shielding the output (i.e., the trained neural networks) against undesired behaviour that could indeed be harmful. In this respect, the ethical concerns that may be raised at the algorithmic level are those that apply to any approach that uses deep learning in mission-critical applications and are "rolled-over" to the ethical concerns raised at the "model" and the "output" levels, as outlined below.

2.3.2.2 Data & Model

CREXDATA's learning-based and rule-based techniques use trained neural networks that make sense of perception-level data. It is known that such models are susceptible to magnifying undesired characteristics that may be present in the data they are trained on, such as bias, or malicious noise, into their output. Moreover, they can be manipulated to do so on purpose and there is currently no technique that can conclusively rule-out such behavior in the general case. Yet, the ethical concerns that stem from this fact are milder in CREXDATA, due to the following reasons:

- The data that are used in the project's use-cases are generated by carefully designed simulations and specific field data in the form of time series i.e. for the pandemics. As such, they do not contain malicious noise. Additionally, the nature of the applications that CREXDATA addresses rules-out the presence of social discriminating bias in the data, which could otherwise be reflected in the output, thus violating basic human rights and values, should the trained model be deployed.
- A fundamental pillar of high ethical interest in the CREXDATA approach is formal verification for neural networks. The purpose of such techniques is to mathematically analyse a particular trained neural model and either prove that it is indeed robust to (potentially adversarial) perturbations in the input, or provide a counterexample (i.e., a specific example for which the verification fails). A network that is formally verified as robust can be considered shielded from "attacks" that could exploit a certain

perturbation pattern in the input, in order to manipulate the network into some harmful behaviour. On the other hand, counterexamples from failed verification attempts can be used to further train the network, thus increasing its robustness, until it passes a verification test.

The project is aware that these aforementioned points do not suffice to guarantee that the model will always behave as expected. First, even with simulated and real-time data, it is not possible to exclude cases of critical situations that have not been sufficiently analysed, thus being erroneously represented in the data, or even completely absent. This might lead a model trained on such data to unexpected behaviour. Regarding formal verification, it is infeasible to analyse all possible ways that make a model behave in an unexpected fashion.

It is thus advised in CREXDATA to follow processes for thorough model validation, testing and verification, as well as careful use-case requirements elicitation and data generation techniques, in close collaboration with the use-case domain experts.

3 Assessment List for Trustworthy AI

ALTAI sets a framework for achi`eving Trustworthy AI focusing on fostering and securing ethical and robust AI [2]. Below we present an ALTAI requirement analysis subject to relevant ethical concerns that may come into play. The objective is to raise awareness in regard to such risks, in order to operationalize them properly with the CREXDATA concept as described.

3.1 Human Agency and Oversight (R1)

3.1.1 Human Agency and Autonomy

In principle, there is little risk that the technology developed in the project might undermine human agency and autonomy, since it is not designed for direct, personalized interaction with individuals, but rather for delivering domain-specific insights to specialized decision-making personnel (e.g., civil protection workers, vessel pilots, or drug researchers). Insights extracted by the outcomes of the use cases potentially assist the professionals in designing or adapting certain strategies and approaches over time. However, it is such specialized personnel that is assumed to be mediating between low-level, algorithmic predictions and high-level decisions. Importantly, such personnel are more empowered to do so thanks to the transparency of the techniques developed in the project.

3.1.2 Oversight

CREXDATA allows for human oversight in the development and deployment of its technology via dedicated explainability techniques and the inherent interpretability of the developed rule-based models, which aim at making the trained models and the issued forecasts as transparent as possible, allowing for human intervention. Additionally, there is little risk related to the effects of lack of oversight, since the AI techniques that will be developed in these use cases are not designed for autonomously acting upon their predictions. Rather, the goal is to deliver timely forecasts for critical situations, which human decision makers are to assess, in order to take proactive measures, if necessary.

3.2 Technical Robustness and Safety (R2)

3.2.1. Resilience to Attack and Security

One of the main pillars in CREXDATA's research agenda involves techniques for formally verifying the robustness of its forecasting models against (potentially adversarial) data perturbations, thus opening the inherent black boxes via measures of relevance, logic rules, exemplars, prototypes or counter exemplars.

3.2.2. Accuracy

For AI systems, it is useful to think about any detriment to individuals that could follow from bias or inaccuracy in the algorithms and data sets being used [6]. What is of value at this stage is to identify the system’s tradeoffs due to inaccurate data and output thereof. At first instance, in the CREXDATA context false positive mistakes (false alarms) are relatively cheap (provided that there is not a flood of them), since the user can check the prediction (it is explainable, traceable). False negative mistakes (actual critical situations that are missed) are more important and need to be mitigated. CREXDATA understands that in theory and if misused, its system output may infer information that could pose risks for individuals and groups (i.e., health status in a given region that may affect the credit score of its residents, personal information of any sort, etc.), thus data provenance records should be maintained in order for the project to be able to track how it generated the inference and address it accordingly. Overall, however, statistical accuracy is in itself not useful and usually needs to be broken into different measures [6] like provenance mechanisms.

3.2.3. Reliability Fall-Back Plans and Reproducibility

In all use cases in the project, human decision makers are the sole consumers of the AI system's predictions. Regarding reproducibility, the CREXDATA consortium is committed to best practices related to reproducible research and plans to make code, experimental and evaluation processes fully reproducible.

3.3 Privacy and Governance (R3)

No direct privacy issues are at play at this stage. However, an organizational governance scheme needs to be designed, including: a) internal processes; b) personal data lifecycle monitoring especially in regard to adherence to the GDPR principles, mainly the data minimization and purpose limitation and c) mitigation of events where privacy rights and freedoms are under risk due to lack of awareness and relevant data protection safeguards as early as possible and to the maximum extend.

3.4 Transparency (R4)

3.4.1 Traceability

An advantage of methods that rely on logic and formal methods (as in the neuro-symbolic techniques that will be developed in the project) is that they allow to trace the predictions output by the system. Therefore, since in CREXDATA the high-level forecasting patterns will be interpretable, the produced forecasts will be also traceable.

3.4.2 Explainability

In principle, for interpretable models, traceability and explainability coincide, so we refer to the above. For the black-box (neural) part of the model, dedicated XAI techniques will be used, capable of highlighting the important factors that contribute to low-level predictions.

3.4.3 Communication

This sub requirement is mainly applicable at the deployment phase where the CREXDATA system needs to be communicated as an AI System followed by its technical specifications, instructions, risks, reasonably foreseeable uses and misuses subject to the obligations subject to the AI Liability Directive [3], the Product General Directive [4], and the Product Liability Directive [5] retrospectively.

3.5 Diversity and Non-Discrimination (R5)

3.5.1. Avoidance of Unfair Bias

Technical biases due to system limitations or data correlations may crop up. Algorithmic bias in the system's output is a possibility. All use cases entail relevant risks as defined (see Section 2.3).

3.5.2. Accessibility and Universal Design

End-users are specialized domain experts and, therefore, the project output applies to that level nicely, subject to the required technical expertise.

3.6 Societal and Environmental Well-Being (R6)

CREXDATA could be environmentally detrimental as per the risks defined. Financial implication & societal cohesion at a regional level may be substantially affected by the system’s output in the context of all use cases, to a different extent at each one.

3.7 Accountability (R7)

Audit trails regarding system’s accuracy will be rolled out, subject to clarity of operations and role/liability allocation.

4 THE RISK APPROACH UNDER THE AI ACT

4.1 Risk Classification in General

The proposed Regulation on AI is risk based by design. This means that the compliance measures as per AI system are subject to the level of risk according to the introduced risk classification mechanism and the applying set of binding rules thereof. The proposed Regulation on AI identifies three main AI system classes, subject to their impact on health, safety and fundamental rights, namely:

- prohibited systems,
- high risk systems and
- low risk systems.

To classify an AI System as above, a rather formalistic approach is introduced. Adhering CREXDATA to the risk level scheme as introduced by the proposed Regulation on AI we reach to the following classification scheme as per Section 4.2.

4.2 CREXDATA Risk Classification

4.2.1. Prohibited Systems

Article 5 identifies three main areas where systems need to be prohibited. These are AI systems that:

- deploy subliminal techniques beyond a person’s consciousness with the objective to, or the effect of, materially distorting a person’s behaviour in a manner that causes or is reasonably likely to cause that person or another person physical or psychological harm.
- exploit any of the vulnerabilities of a specific group of persons due to their age, disability or a specific social or economic situation, with the objective to or the effect of materially distorting the behaviour of a person pertaining to that group in a manner that causes or is reasonably likely to cause that person or another person physical or psychological harm.
- evaluate or classify natural persons over a certain period of time based on their social behaviour or known or predicted personal or personality characteristics leading to a number of detrimental treatments.
- employ ‘real-time’ remote biometric identification systems in publicly accessible spaces to be used by law enforcement authorities, or on their behalf, for the purpose of law enforcement, unless and in as far as such use is strictly necessary for specific objectives as defined.

No CREXDATA use case falls into any of the above categories, whatsoever.

4.2.2. High Risk Systems

In the context of CREXDATA we identify a set of drivers of potential high risk as described below:

Article 6(1)(2) identifies as high risk:

- AI systems that are themselves products covered by the Union harmonization legislation (as per Annex II of the Proposed Regulation), which refers to industrial domains like machinery, toys, lifts, equipment and protective systems intended for use in potentially explosive atmospheres, radio equipment, cableway installations, appliances burning gaseous fuels, medical devices and in vitro diagnostic medical devices.
- AI systems listed in the following areas at high level refer to:
  - Biometric identification and categorisation of natural persons
  - Management and operation of critical infrastructure
  - Education and vocational training
  - Employment, workers management and access to self-employment
  - Access to and enjoyment of essential private services and public services and benefits
  - Law enforcement
  - Migration, asylum and border control management.

No use cases are contextually compatible with such cases.

Further analysis is required however, subject to whether the system output:

- is purely accessory to the relevant action or a decision to be taken,
- is likely to lead to a significant harm to health and safety and adverse impact on fundamental rights subject to:
  - the intended purpose
  - the extent of usage of the AI system
  - the likelihood and severity of harm
  - the extent of harm already occurred
  - the extent to which harmful outcomes are not easily reversible
  - imbalance of power, knowledge, age or other socioeconomic circumstances between the system's user and the impacted person.

Following the use cases conceptualisation as described in Section 2.3, Use Case I could be classified as high risk, when considering the above. Although CREXDATA is aware of this potential high risk orientation of Use Case I AI system, such a stance may sound stretched at this stage, as weather emergency driven prediction model seems, at first instance, to operate as an accessory component on decisions regarding critical infrastructure maintenance and not as a standalone or absolutely necessary in terms of normal functionality.

4.2.3. General-Purpose AI Systems – A Field Scenario

On another note, the proposed Regulation on AI introduces the concept of 'general purpose AI'.

According to Article 3(1b): "'*general purpose AI system' means an AI system that - irrespective of how it is placed on the market or put into service, including as open source software - is intended by the provider to perform generally applicable functions such as image and speech recognition, audio and video generation, pattern detection (highlighted as per below), question answering, translation and others; a general purpose AI system may be used in a plurality of contexts and be integrated in a plurality of other AI system*".

Deep diving into the applicability of the proposed Regulation on AI to CREXDATA we come up with the below scenario:

1. The definition of ‘general purpose AI’ is subject to three core elements, namely:
    - ‘intended purpose’
    - ‘*generally applicable functions such as … pattern recognition…*’ (NOTE: we mention solely the pattern recognition function for the sake of the provided field scenario)
    - *AI system may be used in a plurality of contexts and be integrated in a plurality of other AI systems.*
2. At the high level, Complex Event Recognition and - at a later stage - Forecasting is a function that could be tagged as pattern recognition.
3. The goal of Use Case II is to identify optimised drug treatments for COVID 19 patients, as well as an optimal intervention to control outbreaks of cases or minimize its impact in the population. Such a function could be a basic value component and not purely accessory to a decision or action of an AI system listed in EU AI Act, (Article 6(3), Annex III), namely in employment and recruitment tools or a access to essential services like loans provision (i.e., credit scores). In the event the AI system of CREXDATA Use Case II – health crisis is:
    - integrated to another system in employment and recruitment tools or access to essential services, and
    - its output is not purely accessory to the relevant decision or action.

    In that case, it needs to be classified as a high-risk system, subject to Article 6(3). Subject to the above, it is highly recommended for Use Case II to be addressed as such and satisfy the retrospective requirements as set in Article 4b, which refer to a list of requirements for high-risk systems.

5 Conclusions and Perspectives

This document provides an overview of the CREXDATA ethics assessment methodology and process, subject to:

- the CREXDATA AI Systems' overall context, purpose, tasks and the technical elements thereof,
- the applicable regulation,
- wider socio-technical concerns and best ethics practices.

It identifies the CREXDATA AI system lifecycle in three core phases, namely a) design phase, b) development phase, c) deployment phase with the emphasis placed on the design phase and the focus on the system conceptualisation as per use case.

It provides a manual on how to set the appropriate ethical profile and to identify at a later stage relevant measures and additional safeguards to the extent necessary. Subject to its logic, the present deliverable, with its updates, will operate as an ethics manual throughout the CREXDATA lifecycle.

The CREXDATA partners will ensure that an appropriate ethics scrutiny will be followed throughout the project's lifecycle.

Additional information regarding the development and deployment of the project and its ethical implications will be documented in future T1.4 reports.

6 Acronyms and Abreviations

- ALTAI – Assessment List on Trustworthy AI
- HLEG – High Level Expert Group
- ICO – Information Commissioner’s Office

7 References

[1] IEEE (2019): Ethically Aligned Design v.2.0

[2] AI High Level Expert Group (2018): Ethics Guidelines for Trustworthy AI

[3] Floridi et al. (2023): The capAI procedure for conducting conformity assessments of AIS in line with the EU AI Act v.1.0

[4] IEEE (2010): IEEE Guide-Adoption of ISO/IEC TR 24748-1:2010 Systems and Software Engineering- -Life Cycle Management-Part 1: Guide for Life Cycle Management, in IEEE Std 24748-1-2011. 2011. p. 1-96.

[5] Floridi, L. et al. (2018): AI4People - an ethical framework for a good AI society" opportunities, risks, principles and recommendations. Minds and Machines, 28(4): p.689-707.

[6] ICO AI Guidance, March 2023