Robust Learning and Reasoning
for Complex Event Forecasting

Project Acronym:	EVENFLOW
Grant Agreement number:	101070430 (HORIZON-CL4-2021-HUMAN-01-01 – Research and Innovation Action)
Project Full Title:	Robust Learning and Reasoning for Complex Event Forecasting

DELIVERABLE

D1.3 – Ethics Manual on Trustworthy Neuro-symbolic Learning for Complex Event Forecasting

Dissemination level:	PU - Public, fully open
Type of deliverable:	R - Document, report
Contractual date of delivery:	31 March 2023
Deliverable leader:	NCSR
Status - version, date:	Final – v1.0, 2023-03-31
Keywords:	AI Ethics, Safe & Robust AI, Socio-Technical Assessment, Risk Assessment

Funded by the
European Union

This document is part of a project that is funded by the European Union under the Horizon Europe agreement No 101070430. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the Commission. The document is the property of the EVENFLOW project and should not be distributed or reproduced without prior approval. Find us at www.evenflow-project.eu.

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 to be socio-technical systems, which imply risks and negative impacts at the human and societal level. 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 deliverable describes the process and methodologies to be followed throughout the EVENFLOW lifecycle regarding its impact on health, safety and fundamental rights with the focus on the current design phase. It evaluates the relevant risks for the EVENFLOW 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.

Deliverable leader:	Alexandros Nousias (NCSR)
Contributors:	Nikos Katzouris (NCSR)
Reviewers:	Athanasios Poulakidas (INTRA), Alessio Lomuscio (ICL)
Approved by:	Athanasios Poulakidas, Dimitrios Liparas (INTRA)

Document History			
Version	**Date**	**Contributor(s)**	**Description**
0.1	2023-02-05	Alexandros Nousias	Initial ToC
0.2	2023-03-01	Alexandros Nousias	First draft version
0.3	2023-03-20	Alexandros Nousias, Nikos Katzouris	Complete draft version for internal review
0.4	2023-03-29	Athanasios Poulakidas, Alessio Lomuscio, Alexandros Nousias, Nikos Katzouris	Updates following internal review
0.5	2023-03-30	Alexandros Nousias, Nikos Katzouris	Final proofreading
1.0	2023-03-31	Athanasios Poulakidas, Dimitrios Liparas	QA and final version for submission

Table of Contents

List of Tables

Definitions, Acronyms and Abbreviations

Acronym/ Abbreviation	Title
ALTAI	Assessment List on Trustworthy AI
EC	European Commission
HLEG	High Level Expert Group
ICO	Information Commissioner's Office
SoA	State of the Art

1 Introduction

1.1 Project Information

EVENFLOW is developing hybrid learning techniques for complex event forecasting, which combine deep learning with logic-based learning and reasoning into neuro-symbolic forecasting models. The envisioned methods combine (i) neural representation learning techniques, capable of constructing event-based features from streams of perception-level data with (ii) powerful symbolic learning and reasoning tools, that utilize such features to synthesize high-level, interpretable patterns of critical situations to be forecast.

Crucial in the EVENFLOW approach is the online nature of the learning methods, which makes them applicable to evolving data flows and allows to utilize rich domain knowledge that is becoming available progressively. To deal with the brittleness of neural predictors and the high volume/velocity of temporal data flows, the EVENFLOW techniques rely on novel, formal verification techniques for machine learning, in addition to a suite of scalability algorithms for federated training and incremental model construction. The learnt forecasters will be interpretable and scalable, allowing for fully explainable insights, delivered in a timely fashion and enabling proactive decision making.

EVENFLOW is evaluated on three challenging use cases related to (1) oncological forecasting in precision medicine, (2) safe and efficient behaviour of autonomous transportation robots in smart factories and (3) reliable life cycle assessment of critical infrastructure.

Expected impact:

- New scientific horizons in integrating machine learning and machine reasoning, neural, statistical and symbolic AI
- Breakthroughs in verification, interpretability and scalability of neuro-symbolic learning systems
- 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 (healthcare, Industry 4.0, critical infrastructure monitoring)
- Novel FAIR datasets for scientific research
- Novel resources and approaches for verifiable, interpretable, scalable and knowledge-aware machine learning

Table 1: The EVENFLOW consortium.

Number[1]	Name	Country	Short name
1 (CO)	NETCOMPANY-INTRASOFT	Belgium	**INTRA**
1.1 (AE)	NETCOMPANY-INTRASOFT SA	Luxemburg	**INTRA-LU**

[1] CO: Coordinator. AE: Affiliated Entity. AP: Associated Partner.

Number[1]	Name	Country	Short name
2	NATIONAL CENTER FOR SCIENTIFIC RESEARCH "DEMOKRITOS"	Greece	**NCSR**
3	ATHINA-EREVNITIKO KENTRO KAINOTOMIAS STIS TECHNOLOGIES TIS PLIROFORIAS, TON EPIKOINONION KAI TIS GNOSIS	Greece	**ARC**
4	BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION	Spain	**BSC**
5	DEUTSCHES FORSCHUNGSZENTRUM FUR KUNSTLICHE INTELLIGENZ GMBH	Germany	**DFKI**
6	EKSO SRL	Italy	**EKSO**
7 (AP)	IMPERIAL COLLEGE OF SCIENCE TECHNOLOGY AND MEDICINE	United Kingdom	**ICL**

1.2 Document Scope

This deliverable describes the process and methodology for the EVENFLOW 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[2] and aims to ensure that in view of the adoption of EVENFLOW's results in market applications and its overall dissemination and exploitation plan, as per WP2, the project is fully compliant with the EU legal and ethical frameworks as shaped to date, so as:

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[3] 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 [REF-12]. The present deliverable refers to the use case-specific phases of the lifecycle of the EVENFLOW AI system and the relevant areas of ethical and regulatory interest, from design through development, evaluation and operation, so as to anticipate, to the extent possible, its impact on the complex environments in which they operate, taking into account

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

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

the identified risk levels and the following hard requirements and governance schema, that derive directly from EU regulation and relevant soft requirements and governance schema, which are more flexible to the EVENFLOW 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 1 presents an introduction to the project and the document.

Chapter 2 presents the EVENFLOW 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 EVENFLOW 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 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 Ethics Assessment Process and Methodology

2.1 Process and Methodology

Trustworthy AI has three components which should be met throughout the system's entire life cycle: (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 [REF-02].

Aligned with the European Ethical Assessment in the context of the Horizon Europe Programme [REF-07], a dedicated AI Ethical Assessment section has been integrated in the EVENFLOW 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 [REF-08], the Proposal for (revised) Product Liability Directive [REF-09] and General Product Safety Directive [REF-10], 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 [REF-11]. EVENFLOW opted to include in its ethics manual the EU Assessment List on Trustworthy AI (ALTAI) as introduced by the EU High Level Expert Group [REF-02], 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 EVENFLOW's socio-technical instances, thus framing the wider socio-ethical and techno-ethical impact of the project in a holistic fashion.

EVENFLOW understands AI assessment across the life cycle 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 EVENFLOW 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*" [REF-03]. EVENFLOW 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 by NCSR-D in line with the EU legal and ethical imperatives.

2.2 EVENFLOW AI System Overview

The proposed EVENFLOW system aims at forecasting the occurrence of future events from early signs, in order to support proactive and informed decision making. The system will be rolled out in three simulation use cases, a) industry 4.0, in particular examining robot's failure to achieve set goals, b) personalized medicine, namely in forecasting deterioration/relapse events in tumour revolution and c) infrastructure lifecycle assessment, namely in forecasting malfunctions in water pipe networks. As such potential external stakeholders, namely customers, users, operators, are companies operating in the Industry 4.0 (manufacturing, smart factories, AGV control), in Healthcare (personalized medicine) and technology providers in AI/ML (SMEs) research organisations and the Academia. On top of that, there is a lot of EVENFLOW's impact potential at both the level of individual (especially as per use case II and the level of group (mainly use case III but also use case I) in quite broad scale (i.e. neighbourhood or region) that may be affected by the use of the systems are patients 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 neuro-symbolic models will be run on appropriate input data as per use case as follows:

- **Use case I: Industry 4.0**. Sensor readings and position signals of AGV robots moving around smart factory floor in controlled simulations;
- **Use case II: Healthcare - Personalised medicine**. Virtual patient's gene expression profiles generated by a Variational Autoencoder (VAE) trained on publicly available, anonymized omics data related to breast cancer progression;
- **Use case III: Infrastructure lifecycle assessment**. Video feeds on-board cameras over the course of the set simulations and accelerometer readings from in-pipe sensors in water networks.

The above input data are considered adequate and relevant for the use case concepts, so as 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 2 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. water pipeline networks as regional financial status proxy).

Table 2: Data review items for the design phase.

	Use Case I	**Use Case II**	**Use Case II**
Input Data Types	On-vehicle sensor measurements, position signals, images.	Virtual patient's gene expression profiles generated by a Variational Autoencoder (VAE).	Accelerometer readings from in-pipe sensors in water networks.
Point of Reference	Typical AGV operating conditions.	Complex molecular interactions driving cancer progression.	Typical conditions of water flow in water pipe networks.
Task	Use AI-based event forecasting techniques to facilitate AGV navigation in smart factory floors.	Use AI techniques to forecast tumour evolution from early signs.	Use AI techniques to identify malfunctions and leakages in water pipe networks.
Goal	Optimized route planning and minimization of undesirable and unexpected situations in the AGV domain.	Improved oncological forecasting.	Improved predictive maintenance and infrastructure life-cycle assessment in water pipe networks.
Data adequacy	Data reflect real-world conditions/challenges.	Trained on real omics data that capture the complex molecular interactions, driving cancer progression.	Data reflect real-world conditions/challenges.
Data relevance	Data solely for research in robotics related to autonomous robot movement.	Data are subject and limited to biological indicators related to tumor	Data solely for leakage prediction

	Use Case I	Use Case II	Use Case II
		progression in breast cancer and relevant oncological research	

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's of great importance at this stage is to **identify error metrics** and **measure success** retrospectively [REF-04]. 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, 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 4.2.3). EVENFLOW 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 EVENFLOW AI Ethics Assessment

The EVENFLOW AI system is a supportive tool that will allow companies operating in the use case contexts like manufacturing, smart factories, AVG control, healthcare to reach informed decisions. These informed decisions derive from contextually set complex event forecasting as described in Section 2.2. 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 [REF-05]. EVENFLOW 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 EVENFLOW aim/goal in the context of all three use cases is *'fit for purpose'* and for the public benefit and interest. Having defined the applicable value set, the ethical principle set and 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 subject to the overall EVENFLOW context, namely Complex Event Forecasting and its sub-contexts, i.e. AVG mobility, cancer prediction and water leakage, following the scenarios set in 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 synthetic or simulated data, generated in use case-specific simulation environments. Any relevant risks refer to later phases. However, the high level- risk list below, will generate awareness regarding the system's potential trade-offs. More specifically:

2.3.1.1 Use case I: Industry 4.0

1. Safety risks, since AGVs may compromise the workers or other individuals' safety in the factory. For example, AGVs may collide with other machines, objects, or people, causing injuries or even fatalities.
2. Privacy violations in terms of processing location or biometric data, video recordings etc.
3. Environmental harm in terms of AGV production and disposal.

2.3.1.2 Use case II: Predictive Medicine

1. Misdiagnosis, or delayed diagnosis with serious consequences to the patient,
2. Inappropriate treatment recommendations that may lead to ineffective or harmful treatments.
3. False hope or unnecessary worry, as inaccurate or misleading predictions could lead patients to believe that their condition is better or worse than it actually is, leading to emotional distress and potentially inappropriate decision-making.
4. Privacy violations, as sensitive data will be processed and in the event of inappropriate security and governance schema may be subject to unlawful or unethical secondary uses.
5. Environmental harm, as they may suggest massive chemotherapies with subsequent environmental effects.

2.3.1.3 Use case III: Infrastructure Lifecycle Assessment

System's mistreatment that may lead to:

1. Failure to detect potential contaminants or pathogens in the water supply that will lead to public health risks.
2. Leaks or malfunctions in the water pipe network can lead to environmental harm.
3. Social inequality since the given model may not be accessible to all communities thus leading to unequal access to clean water.
4. Cybersecurity risks as these models may access sensitive data about the water pipelines network and surrounding areas.

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 EVENFLOW 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 and Model

EVENFLOW's neuro-symbolic 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 behaviour in the general case. Yet, the ethical concerns that stem from this fact are milder in EVENFLOW, due to the following reasons:

- The data that are used in the project's use-cases are either synthetic or generated by carefully designed simulations. As such, they do not contain malicious noise. Additionally, the nature of the applications that EVENFLOW 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 the EVENFLOW 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 counter-example (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 points (a) and (b) above do not suffice to guarantee that the model will always behave as expected. First, even with synthetic and simulated 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 EVENFLOW 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)

ALTAI sets a framework for achieving Trustworthy AI focusing on fostering and securing ethical and robust AI [REF-02]. 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 EVENFLOW 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. oncology researchers, robot engineers, or water pipe network maintainers). Although, due to its name, the personalized medicine use-case might initially seem to be an exception to that, it is actually not: this use-case does not involve any interaction with patients, nor does it aim at e.g. recommending treatments, or courses of action over the progression of a disease. Rather, "personalized" in the context of the use-case refers to the end-goal of seeking to model and explain tumour progression in terms of the biological, gene-level particularities of individual patients. Insights extracted by the outcomes of this use case could potentially assist medical doctors in designing or adapting a patient's treatment 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

EVENFLOW allows for human oversight in the development and deployment of its technology via dedicated explainability techniques and the inherent interpretability of the developed neuro-symbolic models, which aim at making the trained models and the issued forecasts as transparent as possible, allowing for human intervention. Additionally, for two out of the three use cases in the project (Personalized Medicine and Infrastructure Lifecycle Assessment), 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. The situation is different for the third use case, Industry 4.0, where the transportation robot controller that will be developed does in fact act upon its predictions (by properly steering the vehicle). Such lack of oversight that naturally comes with autonomous behaviour does indeed have risks related to the safety of human workers and equipment. Thorough testing is the key tool to mitigate such risks.

3.2 Technical Robustness and Safety (R2)

3.2.1 Resilience to Attack and Security

One of the main pillars in EVENFLOW's research agenda involves techniques for formally verifying the robustness of neuro-symbolic forecasting models against (potentially adversarial) data perturbations.

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 [REF-06]. What's of value at this stage is to identify the system's trade-offs due to inaccurate data and output thereof. At first instance in the EVENFLOW context false positive mistakes (false alarms) are relatively cheap (provided that there isn't 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. EVENFLOW understands that in theory and if misused, its system output may infer information that could pose risks for individuals and groups (i.e. societal 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 [REF-06] like provenance mechanisms.

3.2.3 Reliability Fall-Back Plans and Reproducibility

Reliability in EVENFLOW is a potential issue for the Industry 4.0 use case, where the transportation robots can act autonomously upon their decisions/predictions. In the other two use cases in the project, human decision makers are the sole consumers of the AI system's predictions. Regarding the Industry 4.0 use case, the EVENFLOW consortium is fully aware of the potential reliability risks involved in the use of autonomous AI systems and aims at taking all necessary mitigation measures, including thorough testing in carefully designed simulation environments. Regarding reproducibility, the EVENFLOW consortium is committed to best practices related to reproducible research and plans to make all code, experimental and evaluation processes fully reproducible.

3.3 Privacy and Governance (R3)

No direct privacy issues 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 EVENFLOW the high-level forecasting patterns will be interpretable, the produced forecasts by EVENFLOW will be 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 EVENFLOW 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 [REF-08], the Product General Directive [REF-10], and the Product Liability Directive [REF-09] retrospectively.

3.5 Diversity and Non-Discrimination (R5)

3.5.1 Avoidance of Unfair Bias

There are no issues of bias in the project. 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; therefore the project output applies to that level nicely, subject to the required technical expertise.

3.6 Societal and Environmental Well-Being (R6)

EVENFLOW 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 mainly in the context of use case III and to a lesser extent use case I.

3.7 Accountability (R7)

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

4 Assessment List for Trustworthy AI (ALTAI)

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 EVENFLOW 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 EVENFLOW 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 o 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 so as 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 EVENFLOW use case falls into any of the above categories, whatsoever.

4.2.2 High-Risk Systems

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

Article 6(1)(2) identifies as high risk, 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. **No use cases are contextually compatible with such cases.**

Some attention should be paid to Use Case I, which might raise risk issues, in case we perceive AGV navigation as critical infrastructure. Following the above, further analysis is required, subject to whether the system output:

- is purely accessory 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 III could be classified as high risk, when considering the above. Although EVENFLOW is aware of this potential high-risk orientation of use case III AI system, such a stance may sound stretched at this stage, as leakage prediction model seems, at first instance, to operate as an accessory component on decisions regarding water pipeline networks and not as a standalone or necessary in terms of 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, 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 systems*'.

Deep diving in the applicability of the proposed Regulation on AI to EVENFLOW 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…'*

- *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 Forecasting is a function that could be tagged as pattern recognition.

3. The goal of use case II is to predict tumor evolution from early signs. 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 access to essential services like loans provision (i.e. credit scores). In the event the AI system of EVENFLOW use case II – predictive medicine 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 Conclusion

This deliverable provides an overview of the EVENFLOW ethics assessment methodology and process, subject to:

- the EVENFLOW 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 EVENFLOW 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 EVENFLOW lifecycle.

The EVENFLOW 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 T1.5 reports.

6 References

[REF-01]	IEEE, Ethically Aligned Design v.2.0 [2019]
[REF-02]	AI High Level Expert Group, Ethics Guidelines for Trustworthy AI [2018]
[REF-03]	Floridi et.al, The capAI procedure for conducting conformity assessments of AIS in line with the EU AI Act v.1.0 [2023]
[REF-04]	IEEE, 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.
[REF-05]	Floridi, L., et.al., AI4People - an ethical framework for a good AI society" opportunities, risks, principles and recommendations. Minds and Machines, [2018]. 28(4): p.689-707
[REF-06]	ICO AI Guidance, March 2023
[REF-07]	EC, EU Grants: How to complete your ethics self-assessment, https://ec.europa.eu/info/funding-tenders/opportunities/docs/2021-2027/common/guidance/how-to-complete-your-ethics-self-assessment_en.pdf. Retrieved 2023-03-31.
[REF-08]	EC, Liability Rules for Artificial Intelligence, https://commission.europa.eu/business-economy-euro/doing-business-eu/contract-rules/digital-contracts/liability-rules-artificial-intelligence_en. Retrieved 2023-03-31.
[REF-09]	EC, Revision of the Product Liability Directive, https://single-market-economy.ec.europa.eu/single-market/goods/free-movement-sectors/liability-defective-products_en#Revision. Retrieved 2023-03-31.
[REF-10]	EC, General Product Safety Directive, https://commission.europa.eu/content/general-product-safety-directive_en. Retrieved 2023-03-31.
[REF-11]	Common European Data Spaces. http://dataspaces.info/common-european-data-spaces/#page-content. Retrieved 2023-03-31.
[REF-12]	Council of the European Union, Proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts - General approach, https://data.consilium.europa.eu/doc/document/ST-14954-2022-INIT/en/pdf. Retrieved 2023-03-31.