TANGO

D5.2 Design of the use cases

Submission date: 30/06/2025

Due date: 30/06/2025

DOCUMENT SUMMARY INFORMATION

Grant Agreement No	101120763	**Acronym**	TANGO
Full Title	It takes two to tango: a synergistic approach to human-machine decision-making		
Start Date	01/10/2023	**Duration**	48 months
Project type	Research and Innovation Action (RIA)	**Funding programme**	Horizon Europe
Deliverable	D5.2: Design of the use cases		
Work Package	WP5 – TANGO case studies		
Type	R	**Dissemination Level**	PU
Lead Beneficiary	FBK		
Authors	Simone Centellegher, Bruno Lepri		
Co-authors	Roberto Pellungrini, Novi Quadrianto, Marco Angheben, Daniele Regoli, Maria Piwonka, Micheal Bergner, Ben Wilson, Dubravko Culibrk, Artur Bogucki, Giovanni De Toni		
Reviewers	Annalisa Andaloro, Federica Giovanella		

Funded by the European Union

DISCLAIMER

COPYRIGHT

DOCUMENT HISTORY

Version	Date	Changes	Contributor(s)
V0.1	05/02/2025	Deliverable outline	Simone Centellegher, Bruno Lepri, Andrea Passerini, Annalisa Andaloro
V0.2	07/05/2025	First draft, use cases contributions	Roberto Pellungrini, Novi Quadrianto, Marco Angheben, Daniele Regoli, Maria Piwonka, Micheal Bergner, Ben Wilson, Dubravko Culibrk, Arthur Bogucki, Bruno Lepri, Giovanni De Toni, Simone Centellegher
V0.3	15/05/2025	Revision and second draft version	Simone Centellegher, Bruno Lepri
V0.4	30/05/2025	First complete draft	Simone Centellegher, Bruno Lepri
V0.5	18/06/2025	Internal review feedback	Annalisa Andaloro, Federica Giovanella
V1.0	26/06/2025	FInal version	Simone Centellegher, Bruno Lepri

PROJECT PARTNERS

Partner	Country	Short name
UNIVERSITÀ DEGLI STUDI DI TRENTO	IT	UNITN
TECHNISCHE UNIVERSITAT DARMSTADT	DE	TUDa
UNIVERSITA DI PISA	IT	UNIPI
CONSIGLIO NAZIONALE DELLE RICERCHE	IT	CNR
SCUOLA NORMALE SUPERIORE	IT	SNS
UNIVERSITE PARIS CITE	FR	UPC
FONDAZIONE BRUNO KESSLER	IT	FBK
CARR COMMUNICATIONS LIMITED	IE	CARR
ISTRAZIVACKO-RAZVOJNI INSTITUT ZA VESTACKU INTELIGENCIJU SRBIJE	RS	IVI
SURGICAL SCIENCE SWEDEN AB	SE	SuS
MARTIN LUTHER UNIVERSITY HALLE-WITTENBERG	DE	MLU
CENTRE FOR EUROPEAN POLICY STUDIES	BE	CEPS
BCAM - BASQUE CENTER FOR APPLIED MATHEMATICS	ES	BCAM
AFLIANT SRL (previously U-HOPPER SRL)	IT	AFL (prev UH)
INTESA SANPAOLO SPA	IT	ISP
EIT DIGITAL	BE	EITD
SHARE FONDACIJA	RS	SHARE
A11-INITIATIVE FOR ECONOMIC AND SOCIAL RIGHTS	RS	A11
MINISTARSTVO ZA BRIGU O PORODICI I DEMOGRAFIJU	RS	MFWD
AZIENDA PROVINCIALE PER I SERVIZI SANITARI	IT	APSS
SWANSEA UNIVERSITY	UK	UoS
THE UNIVERSITY OF WARWICK	UK	UoW
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	FR	CNRS
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE	FR	CNRS

LIST OF ACRONYMS

Acronym	Definition
DoA	Description of Action
EC	European Commission
HaDEA	European Health and Digital Executive Agency
WP	Work Package
AI	Artificial Intelligence
XAI	explainable AI
BMI	Body mass index
IFAC	Interpretable & Fair Abstaining Classifiers
PEAR	Preference Elicitation Algorithmic Recourse
RAG	Retrieval-Augmented Generation
AAAs	Abdominal Aortic Aneurysms
SPECTRE	SPECTral uncertainty set for Robust Estimation
HDSS	Hybrid Decision Support System

Executive Summary

This deliverable presents the design and development framework for four applied use cases that demonstrate how the TANGO platform enables trustworthy, human-centered AI to support decision-making in critical domains. The primary objective is to translate the theoretical, methodological, and technical foundations of the TANGO project into concrete, real-world applications that address pressing societal challenges. In addition, the deliverable clarifies the role of the use cases within the broader project structure, illustrating how they build upon the methodological work carried out in Work Packages 2 and 3, and how they interface with the technical architecture and functionalities of the TANGO platform, developed within Work Package 4 .

The use cases are designed to illustrate how AI, integrated with human oversight, can enhance decision quality, transparency, and fairness across diverse, high-stakes environments.

In the healthcare domain, AI-driven tools are being developed to assist pregnant women and healthcare professionals in making informed, personalized decisions during pregnancy and the postpartum period. Another use case focuses on integrating AI into surgical planning and simulation for abdominal aortic aneurysm treatment, offering intraoperative guidance and post-operative risk prediction. In the financial sector, AI is applied to credit lending decisions, supporting both loan officers and applicants with transparent recommendations and personalized recourse options. Finally, the project introduces an AI-supported policy design tool for the Serbian government, demonstrating how explainable AI and interactive systems can contribute to fair, data-driven social welfare policy development.

Each use case has been carefully designed to incorporate rigorous ethical, legal, and socio-economic considerations. The use of explainable AI, strong privacy protections, and mechanisms for human oversight ensures that these technologies respect fundamental rights while addressing real-world needs.

The next stages of development will be detailed in the forthcoming Deliverable D5.3, which will document progress on implementation, report early results, and provide an initial assessment of the technical, ethical, and societal impact of the use cases. This iterative process will enable the project to continuously refine and improve its approach to developing scalable, trustworthy AI solutions that effectively enhance human decision-making across sectors.

Table of Contents

LIST OF FIGURES

LIST OF TABLES

1 Introduction

1.1 Purpose of the document

This deliverable presents the design and initial development of four real-world use cases that demonstrate the capabilities of the TANGO project in supporting hybrid human-AI decision-making across a set of diverse and socially impactful domains. These use cases address critical areas such as perinatal health and wellbeing, surgical assistance, credit lending, and public policy design—each involving complex decision-making processes where the integration of artificial intelligence (AI) enhances transparency, efficiency, fairness, and overall effectiveness.

Each of the four use cases presented in this document demonstrates how the TANGO platform enables domain-specific applications of hybrid human-AI decision-making. These use cases span diverse societal sectors and tackle complex real-world problems:

- **Supporting Women in Perinatal Health and Wellbeing**: A personal assistant is being developed to support women in making informed decisions during pregnancy and the postpartum period, enhancing autonomy, trust, and access to reliable information.
- **Supporting Surgical Teams in Intraoperative Decision-Making**: A hybrid AI tool is being designed to assist surgical teams with pre-, intra- and post-operative decision support, improving safety, precision, and collaboration between human expertise and machine intelligence.
- **Supporting Loan Officers and Applicants in Credit Lending**: A decision support system is being implemented to aid bank officers in making fair and transparent lending decisions. The system also provides loan applicants with understandable feedback and actionable recommendations through counterfactual explanations.
- **Supporting Policy Makers in Shaping Social Welfare Policies**: A hybrid decision-support tool is being developed to assist policy makers in designing fair and transparent family welfare policies, with a particular focus on the scientific and research community.

The purpose of deliverable “D5.2 - Design of case studies”, in alignment with the project’s Description of Action (DoA), is to provide an overview of the four use cases, detailing how they have been designed and are evolving throughout the project. This document helps clarify their specific roles within the TANGO framework and their contribution to the project’s objectives.

"D5.2 - Design of case studies" is a public deliverable so it is also intended to be made available to the general public interested in the TANGO project as well as to other EU funded projects involving validation of methodologies and tools through the use of case studies.

1.2 Relation to other project work

The methodological foundation for these use cases is rooted in the research and development activities carried out in Work Package 2 (WP2) and Work Package 3 (WP3) of the TANGO project. WP2 focuses on algorithms for synergistic human-machine learning, while WP3 delivers advanced algorithmic solutions for hybrid decision making. These theoretical and technical contributions are embedded within the use cases and assessed through case studies deployments, providing practical validation of the TANGO platform's core principles.

At the core of this effort lies the TANGO software ecosystem (WP4), a flexible and modular infrastructure designed to support the creation, deployment, and integration of hybrid decision-support systems. As detailed in deliverable "D4.1-TANGO System Specifications", the TANGO platform exposes functionalities to both integrate algorithmic contributions from WP2 and WP3, and to make them accessible via API to the WP5 case studies through interactive workflows.

The TANGO ecosystem comprises three major components:

- **TANGO Library Core**: contains interfaces that TANGO contributors must implement to integrate their AI models into the ecosystem. Each model should include components for training, invocation, and pre/post-processing, ensuring seamless lifecycle management through TANGO's MLOps infrastructure.
- **TANGO Library Utils and Model Code**: The central repository for reusable software artifacts where developed libraries must be published under an open-source license (Apache v2). This repository also includes comprehensive documentation, and is made accessible via public Git repositories.
- **API Access and Deployment**: Deployed models can be registered and accessed via the TANGO REST API, enabling interactions such as predictions, classifications, or dialog-based recommendations. The platform supports both the default Databricks-based MLOps (Databricks and MLFlow) stack and custom alternatives through interface adapters, providing flexibility and scalability.

The use cases described in this document will leverage the TANGO platform to develop their respective hybrid AI applications, but also provide ground for validation of the TANGO software components through

practical case study applications. Together, these use cases serve not only as pilots for the TANGO methodology and infrastructure but also as examples of human-centric AI applications with tangible societal benefits. The insights derived from their implementation will inform future developments of the TANGO platform and contribute to the broader adoption of trustworthy, transparent, and effective hybrid decision-support systems.

1.3 Structure of the document

To ensure consistency and clarity, each use case is presented following a standardized structure:

- **Introduction**: Context and relevance of the use case to TANGO project objectives.
- **Use Case Description**: Detailed definition of the problem and objectives. It includes (i) a description of the data and/or data collection methods, offering an overview of the datasets and acquisition processes; (ii) the methodology implemented within the use case; and (iii) a brief overview of the integration into the TANGO platform, including the identification of the specific TANGO ecosystem components to be developed.
- **Ethics**: Analysis of the Ethical, Legal and Socio-Economic Considerations of each case study based on internal interviews with the involved participating beneficiaries.

2 T5.2: Supporting women in perinatal health and wellbeing

2.1 Introduction

The pregnancy and the peripartum period have the potential to improve development opportunities for families and children early and sustainably. There is a wide range of support services aimed at families with different needs. These services are not being sufficiently utilized. A digital navigator (the mobile app “Mein Pia Health Mama Care") can act as an effective guide for coordinating the needs of families with special support requirements during pregnancy and after birth. Here, specific needs can be determined using patient questionnaires and thus targeted knowledge and support services can be provided.

The development of artificial intelligence (AI) offers growing potential, also in healthcare applications. Within the TANGO project the application of AI in medical and social contexts of hybrid decision-making is investigated. The main focus is on the development of AI models, based on a dataset of a preliminary study (eNAV 1.0 of UKHD), that can provide comprehensible support in specific human decision-making situations. After the initial training phase, the algorithm will be validated on an independent dataset to assess its

generalizability and reliability. Human decision-makers' acceptance of AI-generated summaries and annotations of patient data is expected to depend strongly on their transparency and comprehensibility.
In the first decision-making scenario, medical specialists must determine whether or not to involve early support services (e.g., early childhood intervention, Help Network, Youth Welfare Office) by the health care specialists (doctor, midwife, nurses, psychologists). The decision for or against the intervention has so far depended on the personal assessment of the medical staff and was therefore susceptible to subjective judgment regarding the patient's living conditions. To screen for objective risk factors (e.g., drug use before or during pregnancy, social isolation, etc.) the GuStL-Screening tool was established. Yet this tool might be limited by the small number of items and by the assessment by health care specialists.

The second decision-making scenario concerns the pregnant individual's choice of delivery mode following a previous cesarean section. Specifically, the decision is whether to attempt a vaginal birth after cesarean (VBAC) or to undergo a primary cesarean section. The decision is primarily made by the pregnant woman according to her personal preferences. Specialists might give information of the individual risk factors or the chance of successful vaginal delivery after a previous caesarean.

Both scenarios involve individual decision-making and have the potential to explore applications of models for understandability and reliability (WP2), as well as methods for explainable individual decision-making (WP3). While the structure and quantity of the provided data limit the use of neural models, adapted methods for transparent models that incorporate symbolic explanations are expected to remain applicable.

For further data acquisition and validation of AI algorithms the bicentric clinical trial eNav 2.0 (MARTIN LUTHER UNIVERSITY HALLE-WITTENBERG) is aiming to recruit 500 pregnant women. Patients can enroll digitally and receive information about the study via QR code after downloading the digital navigator. Patient-reported outcomes are collected through questionnaires to assess knowledge, use of support services, and health risk factors, including body mass index (BMI), tobacco use, physical activity, and blood pressure at five defined time points (T1–T5). Psychosocial burdens such as depressive symptoms and quality of life are also evaluated using validated questionnaires. The app provides personalized information on pregnancy-related topics such as preventive healthcare, medical check-ups, and lifestyle adjustments, as well as contact information for local support services. The content is tailored to each user's individual medical background and risk profile.

In sum, the study aims to evaluate a family-centered digital navigator to support personalized, needs-based decision-making during pregnancy and the postpartum period. An AI algorithm supporting decision making in pregnancy and postpartum will be developed, validated and integrated in the TANGO ecosystem and finally

into the healthcare app. The project seeks to improve both physical and mental well-being in young families. The application of innovative AI systems in the needs-based personalization of health prevention aims to optimize and expand human decision-making with a positive impact on social justice and social progress. Core objectives include improving screening, prevention, and medical care to strengthen maternal and child health, while fostering effective collaboration between medical, social, and technical sectors.

2.2 Use Case description

On the basis of the preliminary dataset (prestudy eNav 1.0) the developed AI-model is aiming to support the following decision-making scenarios:

- **Decision-making scenario 1**: Decision of the medical staff based on the overall clinical impression and the "Good Start in Life" (GuSTL) screening tool regarding the need for social support services (e.g., early childhood intervention, Help Network, Youth Welfare Office).
- **Decision-making scenario 2**: Patient's decision for vaginal birth after a previous cesarean (VBAC - vaginal birth after cesarean) depending on the outcome of an uncomplicated vaginal birth versus a secondary cesarean.

Using the dataset from our current bicentric clinical study eNav 2.0 (University hospital Halle/Tübingen GER), the validation of the data is carried out taking the following outcomes into account. The evaluation is conducted through using a study-related questionnaire.

The expected output of the AI-model will include a list of weighted recommendations supporting medical staff for pregnant women under supervision of medical staff as part of a patient-centered informed consent discussion. During validation phase the model output will be assessed according to the following outcomes:

- Traceability of AI-processed indications;
- Decision quality (scenario of divergence between AI-processed indication and clinical overall impression or GuSTL screening);
- User satisfaction (health care professionals) with AI-processed indications.

Finally the software implementation and integration in the TANGO ecosystem as well as the digital navigator app is planned.

2.2.1 Data description/Data collection

Prestudy dataset for AI-model development

The dataset includes 685 patients in total recruited in Heidelberg (n=344), Jena (n=102) and Berlin (n=239). The baseline dataset shown in the table below with crucial patients-reported outcomes is collected at the beginning of app usage.

Sociodemographic Variables	Obstetric Variables	Anamnestic Variables
Age	Gravidity	BMI
Highest educational degree	Parity	Blood pressure
Marital status	Pregnancy complications	Pre-existing conditions
Employment status	Fertility treatment	Medication intake
Household income	Gestational age at birth (GA)	Previous treatments
Number of children (in household)	Mode of delivery	Family medical history
Migration background	Duration of labor	
	Birth weight	
	Height	
	APGAR score 1/5/10	
	Birth injuries	
	Malformations	
	Breastfeeding behavior	
	Sex	
	Consumption behavior	

Table 1: Basic data

	Timepoint	Data	Assessment form
T1	≤35. weeks of gestation (App registration)	Anamnesis form Awareness and use of psychosocial services Depression, anxiety Stress burden Prenatal bonding Traumatic experiences Physical activity, BMI	Study questionnaire EPDS PRAQ-R PSS-10 IES IPAQ-SF
T2	36+0 weeks of gestation	Awareness and use of psychosocial services Pregnancy-related anxiety Quality of life Prenatal bonding Partnership satisfaction Urinary incontinence Physical activity, BMI Risk behavior during pregnancy System Usability Scale	Study questionnaire PRAQ-R WHOQOL-BREF PFB F-Soz-U QUID IPAQ-SF HPQ-II SUS
Birth	Birth	Obstetric outcomes	eCRF

T3	4 weeks post partum	Depression, anxiety Stress burden Postpartum bonding Early childhood behavioral regulation BMI	EPDS PSS PBQ-16 CoviFam
T4	3 months pp	Awareness and use of psychosocial services Pregnancy-related anxiety Quality of life Postpartum bonding Partnership satisfaction Psychosocial burden Urinary incontinence Physical activity, BMI Risk behavior during pregnancy System Usability Scale	Study questionnaire WHOQOL-BREF, EBI PFB, F-Soz-U IES, CoviFam QUID, BSES-SF
T5	6 months pp	Depression, anxiety Stress burden Postpartum bonding Early childhood behavioral regulation Risk behavior during pregnancy Breastfeeding self-efficiency System Usability Scale Physical activity, BMI	EPDS, PSS-10 PBQ-16 CoviFam HPQ-II BSES-SF SUS IPAQ-SF

Table 2: Study schedule

Current clinical trial eNav 2.0

Recruitment: Through the low-threshold recruitment of pregnant women via QR codes on smartphones and informational materials shared in outpatient specialist practices and with midwives, the participation barrier is expected to decrease, ensuring a pluripotent participant spectrum.

2.2.2 Methodology

Currently the case study is focused on evaluating machine learning models that can use early stage patient data to predict the value or category of relevant variables at stage of the decision or later stages. In particular, the focus is on evaluating different classifier models (e.g. decisions trees, logistic regression, tabular transformer) suitable for tabular categorical inputs. In addition, the case study will investigate suitable explanation methods (local, rule-based, etc.) as well as explore TRACE as an interface.

2.2.3 Integration into the TANGO platform

As detailed in the D4.1 TANGO system specifications, the TANGO Ecosystem exposes functionalities to both integrate libraries and models contributions from WP2 and WP3 and to access the published AI method via API, giving WP5 use cases the interactive dialog needed to fulfill the specified task.

The contributions to the TANGO Library related to the use case we are currently discussing to integrate are the following:

- At this stage of the case study's development, we are still evaluating the most appropriate models to be used. In the coming months, we will identify the methodology best suited to our needs, with the aim of integrating or leveraging components from the TANGO platform.
- Based on the features of the models needed by the current use case, one or more interfaces among TangoModel, ExplainableTangoModel, TangoModelMetadata will be chosen to achieve the desired functionality. The integration of the methods are currently being discussed.

The libraries will become part of the TANGO Library by satisfying the following requirements:

- Being published in a public Git repository, accessible to the TANGO community and other potential users, ensuring transparency and ease of access; the library must be released under Apache v2 licence;
- Being delivered with comprehensive and detailed documentation, covering the library's purpose, functionality, installation instructions and usage examples.

The TANGO Library provides interfaces to enable integration of models into the TANGO Ecosystem, allowing contributors to define AI methods that fit the platform functional constraints for being exposed through API.

Based on the features of the models needed by the current use case, one or more interfaces among TangoModel, ExplainableTangoModel, TangoModelMetadata can be chosen to achieve the desired functionality. The integration of the methods are currently being discussed.

2.3 Ethics

2.3.1 Ethical Considerations

The use of AI in the T5.2 case study—an application supporting women during pregnancy and the postpartum period—raises critical ethical questions relating to autonomy, transparency, and privacy. The proposed digital navigator aims to provide needs-based, personalised decision support through an app, potentially enhancing patient autonomy by making relevant information and service referrals more accessible. Importantly, the development team emphasized during interviews that the AI will not infringe upon human decision-making: patients and clinicians retain full authority, particularly in sensitive scenarios such as vaginal birth after cesarean (VBAC). In all major decision pathways, the AI's role is explicitly supportive, intended to augment rather than determine outcomes.

Ethical guidance from the WHO and the European Commission underscores that autonomy must be preserved not only through access to information, but also by ensuring users understand the role AI plays in shaping recommendations (World Health Organization, 2021; European Commission, 2020). The team supports this by designing the system to provide explanations rather than directives, framing predictions in a way that reflects underlying data patterns.

Transparency is another key requirement. As the app aims to support decisions based on psychosocial data, health indicators, and behavioural variables, it must ensure that users and clinicians can verify how outputs are generated. The system will produce self-explaining outputs designed to be understandable by both patients and professionals. While discourse on healthcare AI emphasizes that trust depends on interpretable, user-facing explanations—whether local (e.g., tailored to a specific recommendation) or global (e.g., explaining model logic and limitations) (Cartolovni, 2022; Gerke, 2020).

Privacy and informed consent are paramount. The system collects and processes sensitive health data across multiple stages of the procedure, including psychosocial burdens such as depressive symptoms. The project has obtained ethics approval and includes data protection measures integrated into the app, including consent forms and transparent information about data handling. Users are informed at login about how their data will be used. The development team noted that if the dataset proves insufficient, they will reduce system expectations rather than seek additional personal data, demonstrating a privacy-first orientation.

2.3.2 Legal Considerations

Legally, the T5.2 system must navigate obligations under General Data Protection Regulation (GDPR), the AI Act, and potentially the Medical Devices Regulation (MDR), depending on how its recommendations influence clinical decisions. The team acknowledges that the system qualifies as high-risk under the AI Act and will therefore follow strict requirements for transparency, governance, and documentation. Their proactive integration of ethical and legal compliance measures, including data security, consent, and stakeholder awareness, provides a promising foundation in achieving Trustworthy AI level of compliance.

Accountability and liability are being addressed contractually through agreements with the consulting firm responsible for the app and data infrastructure. However, the project remains early-stage, and the team recognizes the need to further clarify these legal dimensions as development progresses. Informed consent processes have already been built into the app and trial design, and medical professionals are explicitly informed that the system uses AI and patient-reported data.

In the scenario at hand the concern was raised that anonymisation measures might not sufficiently prevent the identification or singling out of participants. Given the limited number of participants involved and their shared geographical origin—specifically, that they all reside within the same city—there is an elevated risk that individuals could be indirectly identified, even if explicit identifiers are removed. From a GDPR standpoint, this presents a significant risk, as the regulation emphasizes the principles of data minimisation, storage limitation, and integrity and confidentiality (Articles 5(1)(c), (e), and (f)). Particularly relevant here is Article 4(1), which defines personal data broadly as information capable of identifying individuals directly or indirectly. Recital 26 explicitly highlights the risk of singling out as undermining the effectiveness of anonymisation, which further underscores the urgency of addressing this risk.

Given these considerations, the risk assessment classifies this scenario as high risk, since standard anonymisation techniques such as pseudonymisation or simple removal of direct identifiers would likely not be adequate due to the contextual details that could lead to re-identification.

To mitigate this risk effectively, it is recommended that additional measures be implemented. For instance, reporting data only in aggregated form would prevent individual-level insights. Generalising geographical or demographic characteristics could reduce granularity and thereby lessen the risk of indirect identification. Incorporating techniques like adding noise or perturbing data points could further diminish the potential for re-identification. Lastly, ensuring strict access controls, where access to potentially identifiable data is restricted only to necessary personnel bound by confidentiality agreements, is crucial.

2.3.3 Socioeconomic Considerations

From a socio-economic perspective, the T5.2 targets well-documented gaps in perinatal care—especially in the uptake of early support services—and aims to reduce inequalities in maternal health outcomes by offering digital access to personalized, context-aware resources. The application was specifically designed to be inclusive and accessible for women with lower income and education levels, avoiding complex jargon and offering content tailored to risk profiles. Outreach measures include printed materials and targeted communication with midwives and gynaecologists.

Interview participants acknowledged limitations in accessibility, including the German-only interface, lack of accommodations for visually impaired users, and some reliance on reading skills. Nonetheless, the application provides a digital onramp to services that many women might otherwise not engage with, addressing inequity in access. In line with literature, the team suggested that the AI system may even reduce implicit bias, offering more objective identification of those in need than certain human gatekeepers.

Bias in training data remains an open concern. The development team is using existing ENAV study data and has not yet implemented formal bias mitigation techniques. However, they expressed openness to addressing bias in future stages. They also noted that expanding access to higher-quality and more diverse data—potentially through European Health Data Space initiatives—would enhance both fairness and technical performance.

The app may also alter the traditional patient-provider relationship. AI-supported identification of risk factors may increase patient receptivity to early support services and reduce stigma often associated with provider-initiated recommendations. This shift, where patients feel more included and in control, could strengthen trust and compliance, provided the AI is presented as a collaborative tool.

3 T5.3: Supporting surgical teams in making intraoperative decisions

3.1 Introduction

The case study focuses on the development and evaluation of an AI agent supporting the treatment of abdominal aortic aneurysms (AAAs), which are enlargements in the lower part of the aorta that can lead to aortic dissection and rupture. The agent will support decision making in pre-, intra- and post-operative phases. The pre- and intra-operative components of the agent will be deployed in a surgical simulation environment (ANGIO Mentor) used to train surgical trainees. Trainees will use established simulator cases on which the AI agent will provide pre- and intra-operative support. The post-operative trial will be conducted as a retrospective evaluation against records of past cases and their procedure outcomes.

The AI agent will support the decision making process in all three phases of the treatment:

1. **Pre-operative phase**. The AI agent will provide planning suggestions for the operation given the characteristics of the patients. These include providing planning measurements of vascular features, choosing the inserted endograft (also known as a prosthesis or stent), the imaging C-arm angulation to use during the operation to best visualise the procedure and which side (or leg) from which to introduce the catheter and endograft.
2. **Intra-operative phase**. The AI agent will provide guidance to the surgeon for a correct placement of the endograft so that its fixed ends are positioned within the optimal “landing zones”. This includes

highlighting renal and iliac branch arteries to avoid obstructions or suboptimal positioning which would lead to complications or increased likelihood of failure.

3. **Post-operative phase**. The AI agent will provide estimates about the potential for various leakage problems with the inserted endograft (known as endoleaks) during the follow-up phase, so as to ensure appropriate monitoring of patients according to risk.

Current AI solutions for AAA treatment support are limited in terms of number of morphological features being extracted, segmentation quality (e.g., calcifications and thrombi are typically not included) and amount of manual labour being required. As a minimum valuable product (MVP), we aim to completely replicate the planning data currently manually annotated on a planning sheet.

Alongside the measurements, we aim to provide specific suggestions for decisions such as choosing the endograft, the optimal C-arm angulation to use during the operation and the leg from which to introduce the catheter and endograft.

Moreover, existing solutions mostly focus on pre-operative planning support, and little to no support is provided for intra- and post-operative decision making. Current practice intra-operatively is to have manual annotation to mark the 'no-landing zones' (critical branching arteries) and then check placement by means of further contrast imaging. Reliable and timely annotation would reduce the need for additional contrast medium and radiological exposure - which benefits both patient and staff. Current practice post-operatively is based on clinical intuition about the likelihood of post-op complications and on responding reactively to complications that do arise. Reliable prediction of complications would allow focused attention and what can be disruptive and expensive follow-up on an appropriate subset of patients.

3.2 Use Case Description

The trial involves mostly three partners: UNITN will focus on the development of the AI agent; SuS will contribute to the integration of the agent in the functionalities of the surgical simulator; APSS will provide the required domain expertise as well as the surgical teams that will be involved in the evaluation of the agent. UoS will support drawing up trial protocols, establishing human-centred evaluation methods and developing interaction designs that optimise the human-agent combination.

Clinical staff available are 3 expert surgeons, 5 junior surgeons and 10-20 trainees from the Santa Chiara hospital centre (Trento, Italy), supplemented by surgical colleagues from other centres. We will use a mix of trainees and experts in the evaluation study as follows.

Participants for the pre-operative phase will need to be drawn from a user population which is not familiar with the selection of cases available to the trial, such as surgical trainees and surgical colleagues from other centres. This follows from the need to use retrospective case data available in the simulator.

For the intra-operative phase, case data will be limited to the same simulator cases as used in the pre-operative phase. And participants will need to be drawn from the same or a similar clinician population for the same reasons.

For the post-operative component, in which we hope to demonstrate an ability to predict endoleaks and other complications, we will need cases with a period of follow-up. This means conducting research on cases not in the simulator. The requirement for follow-up means we must retrospectively access past cases where both imaging and follow-up events are recorded. The algorithm will only access data that were available at the time of procedure, while evaluation will make use of data on subsequent events. There is no restriction on participation in this phase other than that we recruit surgeons and assign them cases to which they were not previously exposed.

3.2.1 Data description/Data collection

The dataset used within the simulator (pre- and intra-operative phases) comprises imaging data (computed-tomography angiograms) from which the agent derives detailed anatomical measurements (a total of around 45 features that are mostly lengths, diameters and angles). Each of these features has some influence on the choice of endograft stent, which leg to use for main catheter insertion and what image angles to leverage during the procedure for best visualisation. All other material derives from these image sets for the pre- and intra-operative phases.

For data to be used outside the simulator, APSS provides a dataset from several hundreds of patients affected by AAA, which include, for each patient:

- Angio computed-tomography (CT) scans before, right after, and one year after Endovascular Aneurysm Repair (EVAR).
- Visit reports one year after EVAR.
- Digital subtraction angiographies to confirm prosthesis placement.

Pre- and post-operative CT scans will be used to predict expected outcomes in terms of any complications (e.g., endoleaks, movement of the stent, compromise of post-operative circulation to the lower limbs, etc.). The accuracy of the predictive model will be assessed by comparing its predictions with follow-up visit

reports written by vascular surgeons one year post-operation, where they document any complications that occurred in the respective patients. Ethical approval for the use of the dataset has already been obtained, and a Data Protection Impact Assessment (DPIA) is currently being prepared to ensure the appropriate use and secure storage of the data.

3.2.2 Methodology

Extraction of anatomical features from CT imaging is accomplished by a computer vision segmentation process designed by UNITN. This development has been undertaken with the involvement of surgical staff at APSS and the support of technical experts at SuS. Specifically, we use a pre-trained neural network to identify key anatomical structures, which are then localized and segmented using a custom-designed algorithm. Based on this segmented data, it becomes possible to measure relevant morphological features. These measurements enable the suggestion of the most suitable prosthesis for the patient via another algorithm, along with the suggested leg in which to insert the prosthesis's main body, and the identification of the C-arm projection that offers the best view of the no-landing zone during the EVAR procedure.

Similarly, development of intra-operative support visualisations is being undertaken with the same contributions. The agent’s development will focus on emphasizing the no-landing zones to assist clinicians in inserting the endoprosthesis as safely and effectively as possible.

The plan for recruitment of trainees as participants for pre- and intra-operative trials is under development, but is done with the active involvement of the director of vascular surgery at APSS, so there is no expectation of issues of incentivisation or engagement.

Pre- and post-study semi-structured interviews are planned for the participants to obtain both quantitative and qualitative data to support objective planning and (simulator) procedure outcomes. The planning process will be evaluated with input from senior surgeons, while the simulator procedure outcomes are captured by the simulator itself. Research questions will reflect the need to evaluate both objective and subjective outcomes and will include clinical quality measures as well as time savings and cognitive burden reduction. We will also include estimates of any reduction to radiological exposure and volumes of contrast medium used intra-operatively.

No participants are expected for the post-operative phase, so no trials will take place. During this phase, a model-based forecasting tool will be developed and tested to predict whether a particular patient may experience post-operative complications in the future, aiding clinicians in patient monitoring.

3.2.3 Integration into the TANGO platform

As detailed in the D4.1 TANGO system specifications, the TANGO Ecosystem exposes functionalities to both integrate libraries and models contributions from WP2 and WP3 and to access the published AI method via API, giving WP5 use cases the interactive dialog needed to fulfill the specified task.

The segmentation system will be deployed to the TANGO platform and accessed from there for the trial. There are some local software dependencies which are not making use of intelligent systems. These are not expected to be deployed to the TANGO platform.

The contributions to the TANGO Library related to the T5.3 use case we are currently discussing to integrate are the following ones:

- LORE (Guidotti et al., 2019); as a need to explain the models' predictions to the surgeons in high stake scenarios
- Thrombotic-Vessel Segmentator; a TANGO Model comprised of a novel framework for segmenting blood vessels with thrombotic areas

The libraries will become part of the TANGO Library by satisfying the following requirements:

- Being published in a public Git repository, accessible to the TANGO community and other potential users, ensuring transparency and ease of access; the library must be released under Apache v2 licence;
- Being delivered with comprehensive and detailed documentation, covering the library's purpose, functionality, installation instructions and usage examples.

The TANGO Library provides interfaces to enable integration of models into the TANGO Ecosystem, allowing contributors to define AI methods that fit the platform functional constraints for being exposed through API.

Based on the features of the models needed by the current use case, one or more interfaces among TangoModel, ExplainableTangoModel, TangoModelMetadata can be chosen to achieve the desired functionality. The integration of the methods are currently being discussed.

3.3 Ethics

3.3.1 Ethical Considerations

The integration of AI into surgical practices for treating abdominal aortic aneurysms (AAA) raises significant ethical concerns, particularly regarding individual autonomy, transparency, and inclusiveness. Ensuring individual autonomy necessitates that human decision-making remains central, with AI providing supportive rather than authoritative input (WHO, 2021; European Commission, 2020). Interviews with surgeons involved in the T5.3 trial highlighted comfort levels varying from moderate to cautious reliance on AI predictions, emphasizing that clinical judgment remains paramount, particularly in complex cases.

Transparency remains critical, as both surgeons expressed the necessity for visual verification of AI measurements to trust the system fully. The ambiguity in AI functionalities further underscores this requirement (European Commission, 2020; Cartolovni, 2022; Elendu, 2023). Surgeons acknowledged difficulty in effectively communicating AI predictions to patients, especially given patients' varying levels of understanding, thus complicating informed consent (WHO, 2021; Gerke, 2020).

Equity and inclusiveness in AI deployment were recognized as potential benefits by reducing disparities between well-resourced and less-resourced healthcare settings. However, challenges such as poor-quality CT scans affecting AI outcomes highlight the importance of inclusive and representative data (Filippis, 2025).

3.3.2 Legal Considerations

Legally, liability, informed consent, and data protection remain critical concerns. Interviews indicated that surgeons strongly believe responsibility should always remain with human practitioners, especially if outcomes based on AI recommendations lead to adverse events. This stance highlights the urgent need for clear legal frameworks delineating responsibilities among AI developers, deployers, and healthcare providers (Corfmat, 2025; WHO, 2024).

Informed consent poses additional complexities, notably due to surgeons' own limited understanding of AI functionalities. Transparent communication to patients about the specific use and limitations of AI in surgical planning and decision-making was identified as an essential component for legal and ethical compliance (Gerke, 2020).

Data protection remains a key legal issue, particularly the secondary use of patient data for AI model development. It is imperative to realise that the risks associated with data privacy depend heavily on who

processes the data, which emphasizes the need for rigorous compliance mechanisms under GDPR and data legislation (Corfmat, 2025; WHO, 2024).

3.3.3 Socioeconomic Considerations

From a socio-economic perspective, AI integration promises operational efficiencies and reduced clinician stress, thus potentially improving labour conditions. Surgeons viewed AI positively, expecting benefits in surgical planning efficiency and accuracy. However, they cautioned against completely excluding surgeons from decision-making processes due to potential risks of deskilling and overreliance on AI (Cartolovni, 2022; Martin, 2024).

Commercialisation and power asymmetries were additional concerns, with surgeons emphasizing the increasing role of big tech companies in healthcare. This power shift underscores the necessity of strong regulatory oversight to protect public interests against commercial pressures, ensuring AI developments remain aligned with societal and ethical values (WHO, 2021; Filippis, 2025).

4 T5.4: Supporting loan officers and loan applicants in credit lending decision making

4.1 Introduction

Consider a prospective borrower seeking a loan from a local financial institution. The bank utilizes an AI system to assess loan applications and determine approvals or rejections. However, the applicant's request is denied, and the AI system provides no clear explanation for the decision or guidance on potential steps for reapplying. Can we truly consider this decision fair?

This hypothetical scenario illustrates the potential risks of deploying AI in high-stakes decision-making processes, where the absence of explainability and fairness may lead to unintended negative consequences. Without appropriate safeguards, AI-driven decision systems that lack transparency, interpretability and fairness may inadvertently replicate the biases they are intended to mitigate.

This use case focuses on developing and evaluating an interactive AI agent that simultaneously supports loan officers in making safe, fair, and well-informed lending decisions while empowering loan applicants with transparent information about their applications, as well as interactive counterfactual recommendations for actionable recourse.

Building on the theoretical work developed by FBK and UNITN in WP3 on algorithmic recourse and the research conducted by SNS and BCAM on explainability and algorithmic fairness, we apply these methodologies to bridge the gap between financial institutions and loan applicants. The AI agent enhances decision-making by supporting both bank officers and loan applicants throughout the credit lending process.

Specifically, the AI agent helps identify subgroups that may have been systematically treated differently in past decisions due to unconscious or conscious biases, alerting bank officers on those potential biases. Additionally, it assists loan applicants by providing clear, transparent explanations for declined applications, enabling them to understand the reasons behind the decision and explore possible avenues for improvement.

This trial is particularly relevant in light of the upcoming implementation of the Artificial Intelligence Act (AI Act), which establishes regulatory requirements for high-risk AI systems, including those used in credit lending. The AI Act mandates that such systems "shall be designed and developed in such a way to ensure that their operation is sufficiently transparent" (Article 13). Furthermore, it underscores the critical importance of human oversight in the deployment of high-risk AI systems (Article 14), reinforcing the necessity of explainable and accountable AI in financial decision-making.

4.2 Use Case description

The AI agent is designed to support the decision-making process in which a bank officer evaluates whether to approve or reject a loan application based on the loan applicant's characteristics. It provides transparent explanations for its recommendations, ensuring clarity in the decision-making process. Simultaneously, it assists loan applicants by helping them understand their application outcomes and guiding them toward actions that could improve their chances of securing a loan.

The AI system is built as a continuous interaction process between the AI agent and both the bank officer and the loan applicant.

- **Interaction with the Bank Officer**: The AI agent enhances the bank officer's decision-making by complementing their domain expertise with the pattern recognition capabilities of machine learning and the reasoning power of TANGO platform. Additionally, the system can detect patterns of systematic disparities in past lending decisions, identifying potential biases in the evaluation of data or decision-making process. When such biases affecting sensitive subgroups are detected, the system alerts the bank officer, promoting decision explainability and algorithmic fairness.

- **Interaction with the Loan Applicant**: If a loan application is rejected, the AI agent helps the applicant understand the reasons behind the decision. It provides clear, actionable insights on what changes or improvements could increase their chances of approval in the future. This process, known as algorithmic recourse, empowers applicants by offering them transparent, data-driven recommendations to enhance their financial standing.

4.2.1 Data description/Data collection

Intesa Sanpaolo (ISP), the largest Italian bank, has provided a comprehensive anonymized dataset containing approximately 235,000 personal loan applications collected between 2016 and 2019. This dataset includes over 50 financial attributes covering a wide range of personal, financial, and application-related information. Specifically, it contains:

- **Socio-demographic data**: Age ranges, income levels, number of family members, and other relevant personal details.
- **Bank-related services**: Information on whether applicants have additional financial products, such as insurance policies.
- **Risk assessment**: The applicant's estimated risk level based on internal banking criteria.
- **Sensitive attributes**: Gender and citizenship, which are critical for evaluating algorithmic fairness.
- **Loan application details**: The requested loan amount, loan duration, and the final decision (approval or rejection).

Beyond providing the dataset, ISP will contribute its domain and legal expertise to support the evaluation of the AI agent's algorithmic fairness and recourse mechanisms. A dedicated group of expert domains from the bank will actively participate in the assessment, ensuring that the AI agent's recommendations align with real-world banking practices and regulatory requirements.

4.2.2 Methodology

This section provides a detailed description of the algorithms and methodologies used by the AI agent to support decision-making for both loan applicants and bank officers.

Personalized Algorithmic Recourse

Algorithmic Recourse seeks to empower individuals affected by automated decision systems by prescribing actionable changes that transform an unfavorable outcome into a desirable one (Karimi et al., 2021). Formally, given a black-box decision function and an individual's profile such that the black-box yields an

undesirable prediction, given the user's profile, algorithmic recourse computes a sequence of feasible actions which, when applied by the user, produces a counterfactual status achieving the desired classification, while minimizing some notion of "effort" or cost associated with those actions. Early work in this field focused on counterfactual explanations, highlighting minimal feature changes (e.g., "increase income by $1000") but did not account for users' heterogeneous abilities or preferences in implementing them. Subsequent research has enriched these methods by incorporating causal constraints and optimisation techniques to ensure recourse is both actionable and low-cost. Developed in the TANGO WP3, personalised algorithmic recourse addresses the key limitation of one-size-fits-all recommendations by tailoring action costs to individual users (De Toni et al., 2024). In PEAR ("Preference Elicitation for Algorithmic Recourse"), the algorithm iteratively refines a user-specific cost function through choice-set queries, selecting only those candidate interventions that maximize expected information gain about the user's true preferences (De Toni et al., 2024). This human-in-the-loop approach yields several advantages: (1) reduced implementation cost, as PEAR converges to interventions that can be up to 50% cheaper than those from user-agnostic methods; (2) fewer interaction rounds, since Bayesian preference updates rapidly hone in on plausible cost estimates; and (3) robustness to feedback noise, owing to the principled assessment of uncertainty via the Expected Utility of Selection criterion. By making users first-class participants in recourse generation, personalized approaches like PEAR ensure that recommended actions are not only valid but also feasible and aligned with each individual's circumstances.

Algorithmic Fairness and Explainability

In the dataset of ISP some sensitive information of loan applicants is recorded, namely their gender and citizenship. Hence it is essential to ensure that an AI system trained to make automated loan decisions does not discriminate based on demographic groups as defined by these attributes.

To test for potential discrimination and to empower bank officers to appropriately override discriminatory decisions, we will make use of the selective classification framework behind Interpretable & Fair Abstaining Classifiers - IFAC (Lenders et al., 2024), which was developed within WP3 of the TANGO Project. IFAC utilizes explainable-by-design methods, to detect unfair behaviour in classification tasks: using association rule mining it can extract which specific subgroups within different gender and citizenship categories are at risk of receiving unfair treatment. Within these subgroups, IFAC can also determine which particular individual instances do not get the decision outcome they deserve. This is done by an individual fairness analysis, where the decision outcome of one instance is compared to highly similar ones, that only differ on the specified sensitive attributes. Being a selective classifier, IFAC can refrain from making predictions on these instances

that are deemed to be at risk of group- or individual discrimination. These rejected instances, along with an explanation behind their rejection, can then be passed on to a bank officer, who can further contextualize or potentially challenge the detected bias patterns.

Moreover, we will use SPECTRE (SPECTral uncertainty set for Robust Estimation), a new method designed to boost fairness without needing access to demographic data. SPECTRE was developed within WP2 of the TANGO project. Many fairness-enhancing methods assume something quite unrealistic: that we always know each person's group identity (like race, gender, or other sensitive attributes). But in real life, that information is often missing — and sometimes, we shouldn't even be collecting it. SPECTRE is based on a concept called robust optimization, which focuses on improving performance in worst-case scenarios. But here's the twist: instead of obsessing over extreme outliers (which often makes models overly pessimistic and less useful), SPECTRE takes an alternative route. It uses a technique involving spectral analysis and Fourier features to tweak just what's necessary. The result? A model that doesn't need group information but still manages to be fair and effective.

Collaborating with actual loan officers of ISP, will give us a unique opportunity to see how well the methodology behind IFAC and SPECTRE fits into their current workflow, and how useful the provided explanations behind rejected instances are to understand and override discriminatory behaviour in the AI system.

Finally, to support the decision making process on both parts, counterfactual rule explanation could be applied to produce alternative decision scenarios for both the applicant and the officer. Through a comprehensive visualization of the generated explanation it is possible to offer to both participants an understanding of the most relevant features for the decision and the logical explanation of said decision, with possible alternatives. To do this we can employ the LORE algorithm (Guidotti et al., 2019) as a backend for the TRACE visualization framework, enabling a full understanding of the local decision provided by the credit prediction model. TRACE creates a cohesive narrative of the model logic by integrating multiple levels of information, from high-level predictions to the granular importance of features. TRACE merges into a unified interface the contribution of multiple explorations: model prediction, feature importance, data distribution, logi clues and counter rules. Moreover, interpretable and fair mechanism for abstention (IFAC) can enable an understanding of the most uncertain cases, i.e., of those instances for which the credit prediction model is most unsure.

4.2.3 Integration into the TANGO platform

As detailed in the D4.1 TANGO system specifications, the TANGO Ecosystem exposes functionalities to both integrate libraries and models contributions from WP2 and WP3 and to access the published AI method via API, giving WP5 use cases the interactive dialog needed to fulfill the specified task.

The contributions to the TANGO Library related to the Credit Lending use case we are currently discussing to integrate are the following:

- Algorithmic recourse;
- SPECTRE (SPECTral uncertainty set for Robust Estimation);
- LORE (Local Rule-Based Explanation);
- TRACE;
- IFAC classifier.

The libraries will become part of the TANGO Library by satisfying the following requirements:

- Being published in a public Git repository, accessible to the TANGO community and other potential users, ensuring transparency and ease of access; the library must be released under Apache v2 licence;
- Being delivered with comprehensive and detailed documentation, covering the library's purpose, functionality, installation instructions and usage examples.

The TANGO Library provides interfaces to enable integration of models into the TANGO Ecosystem, allowing contributors to define AI methods that fit the platform functional constraints for being exposed through API.

Based on the features of the models needed by the current use case, one or more interfaces among TangoModel, ExplainableTangoModel, TangoModelMetadata can be chosen to achieve the desired functionality. The integration of the methods are currently being discussed.

4.3 Ethics

4.3.1 Ethical Considerations

Ethically, the T5.4 system is designed to improve the autonomy of both the loan applicant and the loan officer. The algorithmic recourse tool enables applicants to receive counterfactual explanations that reflect their preferences and focus on actionable variables, thus expanding their practical agency within the credit

application process. This approach marks a departure from traditional credit scoring, where reasons for rejection are opaque or entirely inaccessible.

From the loan officer's perspective, the system introduces an alert mechanism designed to mitigate automation bias. As noted during the interview, this is meant to reinforce the officer's critical engagement with algorithmic recommendations and preserve their professional judgment. However, developers acknowledged that if alerts are overly prescriptive—signalling only one fair option—this might reduce the loan officer's autonomy.

Furthermore, the interview highlighted the importance of balancing transparency with cognitive accessibility. It remains undecided how exactly the system will present information to officers, and this presents a key design challenge. Making the bias detection mechanism interpretable and embedding uncertainty metrics could help maintain informed oversight.

Systemically, the interview responses echoed concerns raised in the literature. Stakeholders were aware that while the project focuses on individual fairness, its real-world impact will depend on how the system generalizes across diverse populations and institutions. Currently, it does not address issues such as synthetic data generation, cyber-security, or full compliance with AI Act safety standards, as it is not intended for market deployment in its current stage.

4.3.2 Legal Considerations

Legally, the T5.4 system adheres to the principles enshrined in the GDPR and anticipates several obligations under the AI Act, especially Articles 13 and 14, which emphasize transparency and human oversight. Interviewees emphasized that current EU law prohibits fully automated decision-making in credit lending. As such, the AI tool is explicitly framed as a decision-support system.

Nonetheless, legal challenges remain. The developers acknowledged that no real customer informed consent was obtained, given that only anonymised data was used. For future commercial deployment, however, informed consent mechanisms will be necessary—especially if algorithmic recourse tools are offered to end users. Moreover, while sensitive data like gender and citizenship are used for fairness auditing, a robust justification under Article 10(5) of the AI Act would be required in a production setting.

Uncertainty about whether the AI Act's prohibition on social scoring would apply to the case study remains open. While the system uses financial variables typical of credit evaluation, the regulatory line distinguishing

legitimate scoring from prohibited social scoring remains blurred. This ambiguity highlights the need for detailed legal mapping.

The issue of accountability also remains central. Under GDPR, the loan officer remains responsible for the final decision, but as algorithmic recourse tools gain influence, clarity is needed on who bears responsibility for errors arising from their use. This raises a gap in both the literature and regulatory practice.

4.3.3 Socioeconomic Considerations

Socioeconomically, the T5.4 case study offers potential for improving financial inclusion by offering applicants a pathway to understand and potentially improve their creditworthiness. However, this benefit assumes the availability of truly actionable recommendations. The interview confirmed that the system avoids recommendations based on immutable characteristics, and this should enhance real-world usability across varied applicant profiles.

Still, as acknowledged by the stakeholders, systemic deployment of such systems can have implications for labour markets. While AI is framed as a supportive tool, it also reduces the human workload, especially in data input and initial assessment tasks. Over time, this may lead to staff reductions or role reconfigurations within financial institutions. Finally, while the technology doesn't aim to replace loan officers, it may inevitably impact job numbers.

In terms of adoption, commercial incentives remain a question. Developers recognize that while fairness is a primary goal, banks may be wary of systems that reduce profitability due to increased false positives or conservative bias detection. As one interviewee commented, prioritizing fairness over predictive accuracy could conflict with institutional incentives, particularly if it leads to more loan defaults.

This tension underscores the importance of aligning algorithmic fairness with institutional business models, potentially through policy incentives or regulatory requirements.

5 T5.5: Supporting policy makers in shaping social welfare policies

5.1 Introduction

This use case addresses the critical need to support government policy makers in Serbia in shaping fair, transparent, and data-driven social welfare policies. It focuses particularly on the Ministry for Family Welfare

and Demography (MFWD) and leverages artificial intelligence to assist with decisions around targeting vulnerable social groups, with a primary focus on researchers—especially women—affected by maternity leave policies.

The motivation stems from the lack of systematic and AI-supported policy tools in Serbian governance, where most welfare policies are based on rule-of-thumb decisions, often lacking transparency and responsiveness to real-time data. This case study aims at demonstrating the value of embedding Retrieval-Augmented Generation (RAG) AI models and eXplainable AI (XAI) into the public sector's decision-making processes.

This use case is deeply aligned with the theoretical and methodological components of WP2 (Trust and Mutual Understanding) and WP3 (Algorithms for Hybrid Decision Making), especially regarding explainability, human-in-the-loop systems, and decision support under uncertainty.

The core goal is to pilot and validate a human-AI collaborative platform that empowers policy makers to better understand the challenges faced by vulnerable groups and to design targeted, evidence-based welfare strategies.

5.2 Use Case description

This use case develops and pilots a policy support system based on the TANGO framework. The key stakeholders include:

- **Technical team (IVI)** – Responsible for system architecture, RAG integration, and data analytics.
- **Ministry of Family Welfare and Demography (MFWD)** – Public authority leading policy creation.
- **NGOs (A11, SHARE)** – Providing social insights and support in outreach and evaluation.
- **CEPS and BCAM** – Contributing to impact assessment and fairness evaluation.

The system uses a combination of **structured data from the E-Science portal** and **qualitative data from custom questionnaires** to detect disparities and recommend tailored policy interventions.

The AI tool is will be interactive and conversational: decision makers enter natural-language prompts and receive real-time feedback in the form of:

- Dashboards highlighting demographic and geographic insights.
- AI-generated narrative explanations.
- Suggestions for interventions, aligned with fairness criteria.

To evaluate and improve the solution a two phase approach is planned. Phase 1 will be a pilot with NGO policy experts (A11, SHARE), allowing us to improve the solution, before entering Phase 2. Phase 2 will consist of a field test with MFWD, once the political situation stabilizes.

Users will assess ease of use, perceived trust, and clarity of recommendations via structured surveys and interviews.

5.2.1 Data description/Data collection

The primary dataset is sourced from the **E-Science Portal**, which includes:

- Profiles of researchers in Serbia.
- Employment history, productivity metrics, demographic information.

Key fields include:

- Researcher identifiers (national ID, ORCID).
- Academic fields and titles.
- Gender, birth year, and institutional affiliation.
- Academic promotion histories.
- Citation counts from multiple bibliometric databases.
- Research outputs classified by national standards.

This will be augmented by a **custom questionnaire** targeted at women researchers, focusing on:

- Career obstacles during and after maternity leave.
- Funding access, work-life balance, mentorship, and bias.

Together, this forms a rich dataset for training and evaluating the AI system.

Due to political instability in Serbia (fall of government and ministerial reorganization), questionnaire distribution was delayed, and pilot testing has been adapted to involve NGO stakeholders in the first phase.

5.2.2 Methodology

The full system is envisioned as a multi-agent framework designed to coordinate specialized agents, each focused on a distinct set of responsibilities within the data-driven research analysis and policy support workflow. This modular architecture ensures scalability, flexibility, and efficiency by distributing tasks across agents with complementary capabilities.

The overall structure of the framework is illustrated in Figure 1., showing the flow from the user's input through orchestration, specialized agent collaboration, and final policy-oriented outputs.

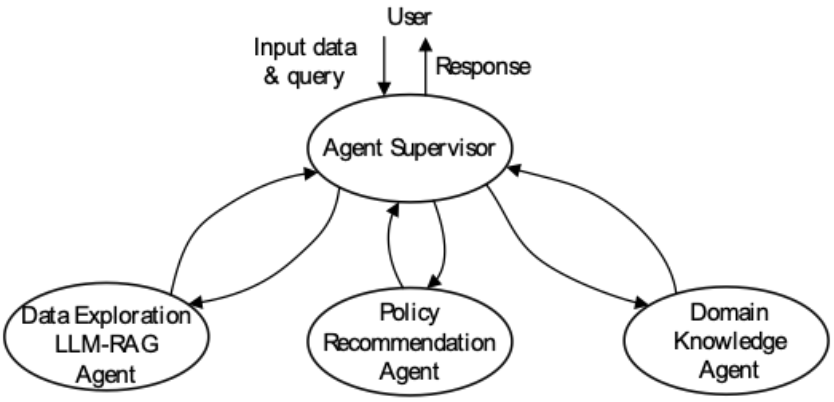

Figure 1: Block diagram of Multi-agent LLM-based system

In the framework, the Agent Supervisor receives input statements from the user and devises a plan to break down the problem into smaller tasks and then forwards the tasks to agents who are best suited for the specific tasks. The supervisor checks the validity of the task executed by the assigned agent. The reasoning process is inspired by plan-and-solve prompting. The Data Exploration LLM-RAG Agent executes structured data querying, statistical analysis and visualization based on user prompts and retrieved data. The Domain Knowledge Agent enhances the analytical process by integrating domain-specific expertise, models and policy frameworks to ensure contextually relevant outputs. The Policy Recommendation Agent synthesizes analytical results and domain knowledge into coherent, evidence-based policy options and strategic insights.

Through seamless interaction between these agents, the system is capable of addressing complex domain questions and translating analytical findings into actionable recommendations. This multi-agent setup fosters distributed intelligence, allowing specialized components to collaborate dynamically and deliver rich, multi-dimensional insights to users.

Example of an Agent: Data Exploration LLM-RAG Agent

The Data Exploration LLM-RAG Agent is designed to transform natural language queries into structured, executable analyses. It implements a full Retrieval-Augmented Generation (RAG) architecture, consisting of three integrated modules: retrieval, augmentation, and generation as is illustrated in Figure 2. These modules enable the system to guide large language models (LLMs) with contextual data, relevant analytical methods, and execution logic, delivering both statistical summaries and visual outputs.

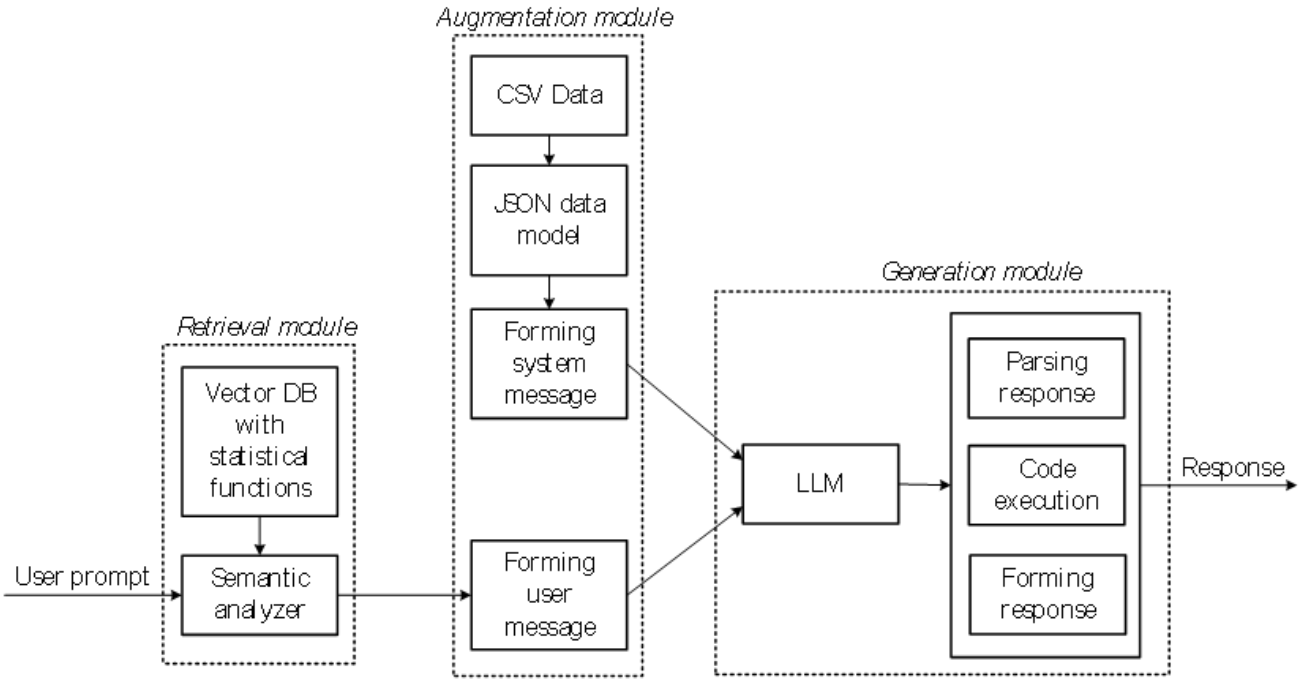

Figure 2: Block diagram of the Data exploration LLM-RAG agent

Retrieval

The retrieval module is responsible for identifying statistically relevant methods to guide the analysis. When a user submits a query, the system semantically analyzes the text to find appropriate statistical techniques. It uses a vector database (ChromaDB) to store embeddings of the statistical method descriptions such as t-tests, ANOVA, Pearson correlation, regression models, and clustering algorithms. The user's query as well as embeddings in the vector database are encoded using an embedding model (all-MiniLM-L6-v2) and then compared to the stored vectors.

If a statistically relevant method is found with a similarity score below a defined threshold, the corresponding method name and description are retrieved. This result is used to enhance the prompt sent to the language model, ensuring the generated analysis aligns with appropriate statistical practices. If no relevant method is confidently retrieved, the system defaults to generating general-purpose code, preserving robustness and flexibility.

This retrieval component forms the first stage of the RAG pipeline, enabling dynamic incorporation of external, domain-relevant knowledge into the analysis process.

Augmentation

The augmentation module prepares the language model context by constructing a complete, data-aware prompt. First, the structured dataset is automatically parsed and modeled into a JSON schema, describing each field's name, type, and semantic role. This schema acts as a formalized summary of the dataset's structure and is critical for guiding how the language model interacts with the data.

The system then assembles a system message, which defines constraints and operational guidelines such as instructing the language model to work only with the preloaded Pandas data (a Python library for data manipulation and analysis) and prohibiting external imports or data redefinition. Simultaneously, the user message contains the natural language query, optionally augmented with the statistical method retrieved during the previous phase.

These two messages form the composite prompt sent to the LLM model. Together, they give the model both the semantic intent of the query and the structural context of the dataset, enabling it to generate appropriate and executable analytical code.

Generation

In the generation module, the LLM model returns markdown-formatted output containing one or more python code blocks. These blocks may include data filtering, aggregation, statistical testing, or visualization instructions depending on the complexity of the query and any injected statistical method.

The system parses these code blocks, checks for syntax validity, and then executes them in a secure, isolated environment. Textual results, such as statistical test outputs or numerical summaries, are captured, while any plots generated using Matplotlib (a python library for data visualization) are rendered and encoded as Base64 images for downstream presentation.

To support iterative exploration, the agent also maintains conversational history, allowing users to refine their questions or ask follow-up queries based on previous results. This module completes the RAG loop by producing tangible outputs from semantically guided code generation.

5.2.3 Integration into the TANGO platform

As detailed in the D4.1 TANGO system specifications, the TANGO Ecosystem exposes functionalities to both integrate libraries and models contributions from WP2and WP3 and to access the published AI method via API, giving WP5 use cases the interactive dialog needed to fulfill the specified task.

TANGO's tooling (e.g., model registry, metadata standards) will ensure full compliance with transparency and auditability standards.

The AI policy design support agent will be deployed as a modular RAG agent using ReAct logic and integrate with the TANGO platform through:

- Exposure via TANGO APIs of utilities required for the deployment of the Data Exploration Agent to different domains.
- Deploying within TANGO open-source versions of the Data Exploration Agent.
- Use of TANGO's ecosystem for lifecycle management (via MLOps), ensuring traceability, logging, and feedback incorporation.

The libraries will become part of the TANGO Library by satisfying the following requirements:

- Being published in a public Git repository, accessible to the TANGO community and other potential users, ensuring transparency and ease of access; the library must be released under Apache v2 licence;
- Being delivered with comprehensive and detailed documentation, covering the library's purpose, functionality, installation instructions and usage examples.

The TANGO Library provides interfaces to enable integration of models into the TANGO Ecosystem, allowing contributors to define AI methods that fit the platform functional constraints for being exposed through API.

Based on the features of the models needed by the current use case, one or more interfaces among TangoModel, ExplainableTangoModel, TangoModelMetadata can be chosen to achieve the desired functionality. The integration of the methods is currently being discussed.

5.3 Ethics

5.3.1 Ethical Considerations

The ethical dimension of this use case is fundamentally shaped by its commitment to fairness, transparency and inclusivity in policymaking. The adoption of explainable AI (XAI) and retrieval-augmented generation (RAG) technologies supports what Vredenburgh (2024) terms deliberative transparency, enabling decision-makers to understand not only what decisions are proposed but also why they are proposed. The combination of dashboards, narrative outputs, and interactive prompts fosters procedural visibility and strengthens democratic legitimacy in AI-assisted governance. Importantly, this aligns with the value of informed self-advocacy, a core ethical requirement for individuals affected by state decisions.

Equally significant is the use of human-in-the-loop logic and structured feedback mechanisms involving both public authorities and NGO experts. This mitigates automation bias and addresses concerns over the erosion of human judgement in complex decision-making. The embedding of human oversight aligns with recent ethical frameworks which emphasise the importance of preserving expert deliberation, contextual sensitivity and reflexivity in AI-supported institutions.

Nevertheless, ethical risks persist. As highlighted by Bernini, Ferretti and La Rosa (2024), the deployment of AI in governance often amplifies latent institutional weaknesses rather than resolving them. In Serbia's case, relatively low levels of institutional trust and ongoing political instability raise concerns about whether the ethical safeguards built into the AI system can be meaningfully upheld. Without institutional reform, even well-aligned AI systems may struggle to produce fair or just outcomes. This underscores the need for complementing technical value alignment with broader institutional change, particularly in contexts where public accountability and administrative continuity are fragile.

5.3.2 Legal Considerations

From a legal standpoint, the processing of personal and sensitive data—such as employment histories, demographic indicators, and qualitative information on maternity leave—invokes several obligations under the General Data Protection Regulation (GDPR). In particular, the use of special categories of personal data triggers heightened safeguards under Article 9 GDPR. The legal principle of data minimisation must be respected, and the project will likely require a formal Data Protection Impact Assessment (DPIA) under Articles 35–36, given the potential for significant rights-affecting decisions.

Further legal scrutiny must focus on Article 22 GDPR, which concerns automated decision-making. While the use case emphasises that the system functions as a decision-support tool rather than an autonomous decision-maker, this boundary must be clearly maintained. Where policy interventions have a legal or material effect on individuals—such as eligibility for benefits or research funding—the AI's role must remain advisory and non-binding. The provision of intelligible explanations to affected individuals, as required under the GDPR, is facilitated by the system's XAI functionality, though the legal adequacy of such explanations will depend on their accessibility and interpretability for non-expert users.

Additionally, legal accountability must be clarified. While the technical components are delivered by private and research entities, ultimate responsibility for decisions and outcomes must rest with the MFWD and other public actors. As the literature on GovTech warns (Bharosa, 2022), public-private digital partnerships risk shifting control away from public institutions unless formal accountability structures, documentation protocols, and procurement guidelines are robustly enforced.

5.3.3 Socioeconomic Considerations

Socioeconomically, the project offers both direct and systemic benefits. It targets a historically under-supported demographic—women researchers affected by maternity-related career disruption—and enables more nuanced, evidence-based interventions. This contributes to gender equality in scientific research and can address structural barriers to inclusion. By combining structured administrative data with qualitative insights from custom surveys, the AI system is well-positioned to detect intersectional disparities and inform targeted welfare reforms.

Beyond this immediate objective, the project holds broader significance for governance innovation in Serbia. By demonstrating the feasibility of AI-assisted, transparent and participatory policy processes, it may contribute to rebuilding public trust in welfare institutions. In contexts marked by rule-of-thumb policymaking and opaque decision-making logics, such tools can enhance both responsiveness and legitimacy. Furthermore, the system is modular and integrated into the TANGO platform, making it scalable and potentially reusable in other policy domains.

6 Conclusions

This deliverable, D5.2 - Design of Case Studies, has presented the design and initial development of four real-world use cases that exemplify the potential of the TANGO platform to support hybrid human-AI

decision-making across a range of critical and socially impactful domains: perinatal health and wellbeing, intraoperative surgical support, credit lending, and public policy design.

Each use case plays a crucial role in demonstrating how the theoretical, methodological, and algorithmic components of the TANGO framework can be instantiated in real-world settings. By addressing diverse scenarios and involving different stakeholder groups, from clinicians and patients to financial officers, citizens, and policymakers, these case studies offer valuable insights into the practical integration of AI technologies that are trustworthy, fair, transparent, and human-centered.

The structured presentation of each use case—in terms of its context, data strategy, methodology, and platform integration, lays the foundation for a consistent and replicable approach to hybrid decision support system (HDSS) development. Importantly, the ethical and socio-economic considerations embedded in each case underline TANGO's commitment to developing AI solutions that are aligned with European values such as dignity, equity, and inclusiveness.

These case studies not only serve as demonstrators of TANGO's capabilities but also as testbeds for future refinement and expansion of the framework. They offer a scalable blueprint that can be adapted to additional domains, thanks to the modular and interoperable nature of the TANGO platform. Their diversity enhances the replicability and long-term sustainability of project outcomes across multiple sectors and European contexts.

Looking ahead, the progress of each use case will be further monitored and documented in Deliverable D5.3 - Case Studies Intermediate Report, which will provide updates on implementation, user feedback, and impact assessment. Through the continued evolution of these case studies, the TANGO project will gain a deeper understanding of how hybrid AI systems can empower professionals, reduce cognitive and emotional strain, enhance decision quality, and foster trust in AI technologies. These insights will be instrumental in realizing TANGO's broader vision: to become the reference framework for trustworthy, human-centered AI in high-stakes decision-making, ensuring that no one is left behind.

References

Bartlett, R. P., Morse, A., Stanton, R., & Wallace, N. (2022). Consumer-lending discrimination in the FinTech era. *Journal of Financial Economics, 144*(1), 28–52. https://doi.org/10.1016/j.jfineco.2021.07.008

Berg, T., Burg, V., Gombović, A., & Puri, M. (2018). On the rise of fintechs: Credit scoring using digital footprints. *The Review of Financial Studies, 33*(7), 2845–2897. https://doi.org/10.1093/rfs/hhz099

Bernini, F., Ferretti, P., & La Rosa, F. (2024). Artificial Intelligence's (AI's) Responsible Use: How to Manage Digital Ethicswashing. In *Creating Value Through Sustainability: An Interdisciplinary Perspective* (pp. 29-64). Cham: Springer Nature Switzerland.

Cartolovni, A. (2022). Transparency and explainability of AI systems in healthcare.

Corfmat, E. (2025). Ethical considerations in healthcare AI: Protecting autonomy, privacy, and inclusion.

De Filippis, G. (2025). Privacy and confidentiality in AI-driven healthcare.

Elendu, J. (2023). Patient consent and data protection in AI-driven healthcare.

European Commission. (2021a). Proposal for a regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) (COM/2021/206 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52021PC0206

European Commission. (2021b). Proposal for a directive on consumer credits (COM/2021/347 final). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0347

European Commission, High-Level Expert Group on Artificial Intelligence. (2020). *The assessment list for trustworthy artificial intelligence (ALTAI)*. European Commission. https://ec.europa.eu/digital-strategy/publications/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment_en

Gerke, S. (2020). Transparency in healthcare AI: Challenges and solutions.

Gonzalez, M. (2020). Assistive technologies and AI: Transforming accessibility in healthcare.

Langenbucher, K. (2023). Consumer credit in the age of AI: Beyond anti-discrimination law (ECGI Law Working Paper No. 663/2022). *European Corporate Governance Institute*. https://ssrn.com/abstract=4275723

Martin, D. (2024). AI and healthcare: Balancing innovation with ethical considerations.

Mayson, S. G. (2019). Bias in, bias out. *Yale Law Journal, 128*(7), 2218–2300. https://www.yalelawjournal.org/article/bias-in-bias-out

Mennella, R. (2024). AI in healthcare: Navigating ethical, social, and legal challenges.

Svetlova, E. (2022). From individual ethics to systemic risks: Artificial intelligence in finance. *AI and Ethics, 2*, Article 10. https://doi.org/10.1007/s43681-021-00129-1

Vozna, L. (2025). Artificial intelligence and the aging population: Opportunities and ethical challenges.

World Health Organization. (2021). *Ethics and governance of artificial intelligence for health: WHO guidance*. World Health Organization.

World Health Organization. (2024). *WHO guidelines on the use of large language models (LLMs) in health.*

European Union. (2016). *General Data Protection Regulation (GDPR), Regulation (EU) 2016/679*. https://eur-lex.europa.eu/eli/reg/2016/679/oj

Guidotti, R., Monreale, A., Giannotti, F., Pedreschi, D., Ruggieri, S., and Turini, F. (2019). Factual and counterfactual explanations for black-box decision making. IEEE Intelligent Systems, vol. 34, pp. 14-23.

Karimi, A.H., Scholkopf, B., and Valera, I. (2021). Algorithmic recourse: From counterfactual explanations to interventions. In Proceedings of FAccT 2021, pp. 356-362.

De Toni, G., Viappiani, P., Teso, S., Lepri, B., and Passerini, A. (2024). Personalized algorithmic recourse with preference elicitation. TMLR.

Lenders, D., Pugnana, A., Pellungrini, R., Calders, T., Pedreschi, D., and Giannotti, F. (2024). Interpretable and fair mechanisms for abstaining classifiers. In Proceedings of ECML-PKDD 2024.

Funded by
the European Union