



FINAL DAI-DSS RESEARCH COLLECTION

D3.3

Editor Name	Prof. Stefan Böschen (RWTH-HumTec)
Submission Date	February 28, 2025
Version	1.0
State	FINAL
Confidentially Level	PU



Co-funded by the Horizon Europe
Framework Programme of the European Union

EXECUTIVE SUMMARY

This deliverable shows the “Final DAI-DSS Research Collection” as part of the Horizon Europe project FAIRWork. The deliverable aims to describe guidelines, methods and tools for democratising the production process in the light of their flexibilization utilizing artificial intelligence (AI), Optimisation, Human Factors Analytics, and multi-agent systems (MAS) as mediators in the form of prototypes, physical experiments in laboratories, implemented questionnaires, modelling tools or semantic model of criteria catalogues. Importantly, this collection is published in the FARIWork Innovation Shop and represents the key support features for the Democratic AI-based Decision Support System Support System (DAI-DSS). The FAIRWork Innovation shop is accessible online as Deliverable 3.3. This document is considered the accompanying document of the deployed online version.

<https://innovationshop.fairwork-project.eu/>

This deliverable also presents the scientific basis for the FAIRWork project across seven research tracks: The “**Democratization of Decision-Making in Socio-Technical Settings**” examines the dynamics of democratization in industry through MAS by exploring the contextual conditions for implementing a DSS within socio-technical frameworks. The “**Decision-Making Using Multi Agent Systems**” explores the potential of MAS for decentralized, adaptive decision-making in industry. By balancing technical, human-centric, and ethical aspects, MAS enhances efficiency, inclusivity, and scalability in complex systems. The “**Digital Human Factors Analytics**” outlines the use of wearable sensors to capture critical information on human physiological, cognitive-emotional, and resilience states, including the intelligent sensor box (ISB). It also details a novel framework using Personas as Human Digital Twins for Decision Making in Industry 5.0. The “**Optimization in Decision Support Systems**” is crucial for defining clear goals in AI-driven manufacturing, addressing challenges such as process optimisation, automation, and resource allocation. Various techniques, including AI, new algorithms and heuristics, support decision-making and efficiency in manufacturing companies. The “**AI-Enriched Decision Support Systems**” explores how AI methodologies, particularly machine learning (ML), can optimise decision-making in manufacturing, with an emphasis on dynamic tasks. It also addresses the gap between industry and developers by proposing a structured categorisation of DSS, enabling developers to select appropriate AI methods for industrial applications. The “**Model-based Knowledge Engineering for Decision Support**” presents a structured approach to AI adoption in enterprises by proposing a three-layered framework: Identification, Specification, and Configuration. It highlights the role of conceptual and technical models from the identification of the problem setting to the configuration of AI to ensure the alignment with business needs. Furthermore, it also reflects the integration of different AI techniques, such as retrieval-augmented generation (RAG) and large language models (LLMs), within use-case-specific prototypes, demonstrating how model-based methodologies can support AI configuration. It also investigates how such design models can be reused to support the explanation of decision scenarios on a high abstraction level using OMILAB’s Scene2Model tool. The “**Reliable and Trustworthy AI**” demonstrates the importance of transparency for trusting AI systems and that transparency needs to be adapted to the target group. AI systems have to be understandable, which is why a system-dependent approach that sets the user in the center is recommended. A developed transparency matrix with additional individual consulting workshops for the developers has shown to be successful in implementing transparency and accuracy communication into AI services.

PROJECT CONTEXT

Workpackage	WP3: Research on Method and Tools for DAI-DSS
Task	T3.1: Research on Democratization of Decision-Making using Multi Agent Systems T3.2: Research on Digital Shadows and Twins for Human Experts and Data Driven Algorithms T3.3: Research on AI-Based Decision-Making for AI, Robots and Human Experts T3.4: Research on Reliable and Trustworthy AI
Dependencies	WP2, WP4, WP5, WP7, WP8

Contributors and Reviewers

Contributors	Reviewers
Lucas Paletta, Herwig Zeiner (JR)	Roland Perko (JR)
Gustavo Vieira (MORE)	Wilfrid Utz (OMILAB)
Sylwia Olbrych, Alexander Nasuta, Johanna Werz, Johannes Zysk (RWTH-WZL-IQS)	Anas Abdelrazeq (RWTH-WZL-IQS)
Noushin Qeybi, Stefan Bösch (RWTH- HumTec)	Roland Sitar (FLEX)
Marlene Mayr (BOC)	
Christian Muck (OMILAB)	

Approved by: Robert Woitsch [BOC], as FAIRWork coordinator

Version History




Version	Date	Authors	Chapters Affected
1.0	February 28, 2025	Stefan Bösch et al.	All

Copyright Statement – Restricted Content

This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content.

This is a restricted deliverable that is provided to the community under the license Attribution-No Derivative Works 3.0 Unported defined by creative commons <http://creativecommons.org>

You are free:

	to share within the restricted community — to copy, distribute and transmit the work within the restricted community
Under the following conditions:	
	Attribution — You must attribute the work in the manner specified by the author or licensor (but not in any way that suggests that they endorse you or your use of the work).
	No Derivative Works — You may not alter, transform, or build upon this work.

With the understanding that:

Waiver — Any of the above conditions can be waived if you get permission from the copyright holder.

Other Rights — In no way are any of the following rights affected by the license:

- Your fair dealing or fair use rights;
- The author's moral rights;
- Rights other persons may have either in the work itself or in how the work is used, such as publicity or privacy rights.

Notice — For any reuse or distribution, you must make clear to others the license terms of this work.

This is a human-readable summary of the Legal Code available online at:

<http://creativecommons.org/licenses/by-nd/3.0/>

TABLE OF CONTENT

1	Introduction.....	9
1.1	Purpose of the Document	9
1.2	Document Structure	9
2	The FAIRWork Innovation Shop	11
3	Research Methods and Services.....	13
3.1	Overview of Methods and Services.....	13
3.2	Democratization of Decision-Making in Socio-Technical Settings	14
3.2.1	Overview.....	14
3.2.2	Results.....	15
3.2.3	Outlook	17
3.3	Decision-Making Using Multi Agent Systems.....	18
3.3.1	Overview.....	18
3.3.1	Motivation and Reference to FAIRWork Use Case.....	19
3.3.2	Innovation and Research Activities around Multi-Agent Systems.....	19
3.3.3	Description of Functionality and Results.....	20
3.3.4	Outlook	21
3.4	Digital Human Factors Analytics	21
3.4.1	Overview.....	21
3.4.2	Acceptance Analysis on Wearables Sensors in Production Environments.....	25
3.4.3	Study Plan for Resilience-Oriented Field Trial in Production Environment.....	26
3.4.4	Service: Extended Resilience Score Computing: Integration of Recovery-Stress State and Bounce-Back	27
3.4.5	Relevance of Fairness And Transparency in Digital Human Factors Analytics	28
3.4.6	Outlook	28
3.5	Optimization in Decision Support Systems	29
3.5.1	Overview.....	29
3.5.2	Optimise the Check of Calibration Documents	30
3.5.3	Work Schedule by using Human Factors.....	31
3.5.4	Transport Optimisation	32
3.5.5	Outlook	32
3.6	AI-Enriched Decision Support Systems	33

3.6.1	Overview.....	33
3.6.2	Research on Existing AI Applications: ML Catalogue.....	35
3.6.3	Guidelines and Recommendations for AI Developers.....	36
3.6.4	AI-Based Optimizing Solutions for Industry.....	37
3.6.5	Outlook.....	38
3.7	Model-based Knowledge Engineering for Decision Support.....	38
3.7.1	Overview.....	38
3.7.2	Modelling for AI: An Approach for Model-based AI Configuration.....	41
3.7.3	Using Conceptual Models to Support Explanations within Decision Support Systems.....	47
3.7.4	Outlook.....	52
3.8	Reliable and Trustworthy AI.....	53
3.8.1	Overview.....	53
3.8.2	Qualitative Focus Groups about AI Transparency.....	55
3.8.3	Quantitative Experiment Comparing AI Transparency Methods.....	56
3.8.4	Matrix and Guidelines on the Application of Transparency from a Lay User Perspective.....	57
3.8.5	Evaluation of Requirements and the Status Prior to the Introduction of a DAI DSS to the Use Case Partners.....	58
3.8.6	Practical Application of Transparency in Different DAI-DSS Services.....	59
3.8.7	Outlook.....	61
4	Explainability and Fairness in AI Services.....	62
4.1	Overview.....	62
4.2	Explainability and Fairness Introduction.....	63
4.2.1	Explainability in FAIRWork.....	63
4.2.2	Explainability in CRF Use Cases.....	64
4.3	Fairness in FAIRWork.....	64
4.3.1	Application in FAIRWork Approach 1: CRF-Example.....	64
4.3.2	Application in FAIRWork Approach 2: FLEX-Example.....	66
4.4	Model-based Framework Supporting Trustworthiness in AI and Data.....	70
4.5	Ethical Watchdog.....	72
5	The Democracy Question.....	74
5.1	Democracy in Companies.....	74
5.1.1	Worker Expectations Reflecting Features of Democratic AI.....	77
5.2	The Question of Representation.....	78
5.2.1	Conceptual Questions.....	78

5.2.2	Empirical Insights.....	80
5.3	Legitimacy via Social Embedding and Procedural Implementation of AI Tools.....	81
6	Summary and Conclusions.....	82
7	Annex A: List of Abbreviations.....	83
8	References	85

LIST OF FIGURES

- Figure 1: Research tracks underlying the outline of the research collection..... 13
- Figure 2: Extensive research plan for the investigation of human aspects in AI-guided decision-making in FAIRWork. 14
- Figure 3: Dimensions explored in the qualitative evaluation of FLEX Althofen..... 15
- Figure 4: Stages of the computation of the resilience score that underlies the risk stratification model (RRSM). . 23
- Figure 5: Wearable biosignal technologies for the production environment for studies and daily monitoring (Credit: JR). 25
- Figure 6: Acceptance votes for watches (activity trackers), smart biosignal shirts, and eye tracking glasses..... 26
- Figure 7: The overview of the research track AI-enriched DSS..... 33
- Figure 8: Overview of the three-layered approach. 39
- Figure 9: Items to support the “introduction of AI into companies”. 43
- Figure 10: Models to Introduce AI into companies. 45
- Figure 11: Examples of model-based configuration and orchestration..... 46
- Figure 12: Conceptual overview for mapping decision results to conceptual models..... 49
- Figure 13: Outlook example for a model repository..... 52
- Figure 14: Core results from the quantitative focus group analysis: What does transparency mean for lay users and which factors influence the transparency requirements towards AI transparency..... 56
- Figure 15: Requirements and effects of AI transparency differ for user groups like AI experts and end users. 63
- Figure 16: Model-based framework to support AI and Data trustworthiness..... 71
- Figure 17: Features of Democratic AI in companies..... 76
- Figure 18: First task in democratic AI workshop..... 76
- Figure 19: Second task in democratic AI workshop..... 78
- Figure 20: Levels of decision-making processes with MAS giving a socio-technical structure to DAI-DSS. 79

LIST OF TABLES

- Table 1: Innovation shop in FAIRWork..... 12
- Table 2: Results from workshops with different service partners with regard to the transparency in their services. 61

1 INTRODUCTION

1.1 Purpose of the Document

The goal of this document is to use the latest status of comprehensive DAI-DSS research collection to show the importance and form of an innovation shop model as the key outcome for supporting the implementation of DAI-DSS in companies. These research tracks have been schematically sketched in the Deliverables “Deliverable 3.1 DAI-DSS Research Specification” and “Deliverable 3.2 First DAI-DSS Research Collection”. Therefore, this deliverable is about two aims:

- First, to map the specific features of MAS and AI in Decision Support Systems (DSS).
- Second, to showcase the innovation shop model, developed on the basis of the named conceptual as well as empirical analysis.

The first goal of mapping the specific features for DAI-DSS is to provide a detailed account about the most recent developments of the respective research tracks, under investigation of the technical factors inherent in the given use cases, including the application of AI and MAS in DSS, as well as the examination of human aspects, such as the reliability and trustworthiness of AI.

The second goal is to showcase the FAIRWork's Innovation Shop, which is an online platform, where projects result of various *Technology Readiness Levels (TRL)* can be published. This allows supporting the dissemination, communication and exploitation of FAIRWork's results. To do so, we created Innovation Items within our innovation shop to make them publicly accessible. The research tracks and its results described in this deliverable, correspond to published innovation items, published on:

<https://innovationshop.fairwork-project.eu/>

This enables interested parties to easier access the project results and contact the responsible partners. Additionally, the project partners themselves can use multi-media content to describe their results and use these descriptions during and after the FAIRWork project. At the beginning of the deliverable an overview with the created Innovation Items and which section refer to them is provided.

1.2 Document Structure

The document is structured as follows: Section 2 contains an introduction to FAIRWork's Innovation Shop and how we used it to disseminate, communicate and exploit FAIRWork's research results. Additionally, this section contains an overview of the Innovation Items created for this deliverable. Section 3 focuses on the research collection in terms of the concrete research methods and services employed to investigate the technical aspects of decision-making processes, the human aspects in the process, and finally the digital human factors measurements. It shows the 'methodological backstage' for the successful implementation of AI and MAS-based technologies into DSS. Methods such as data-driven modelling, prototyping, and testing are proposed within the AI and MAS domains. Section 4 covers conceptual as well as empirical results on the quest for explainability and fairness in FAIRWork from an algorithmic point of view. We particularly focus on the transparency in algorithms that human IT experts would be able to understand. This transparency has been taken up by the framework of explainable AI (XAI), often

overlapping with Interpretable AI, or explainable machine learning (XML). Related to this, but with a specific focus on representation, Section 5 addresses the question of democracy in companies and offering a heuristic of conducting this type of research and practice. Doing so, the conceptual questions regarding representation and its empirical insights, as well as the legitimacy of AI tools through social processes are presented. Finally, the report concludes with a summary Section 6, where the authors summarize the key points debated in the deliverable and emphasize the importance of incorporating the human perspective into decision-making processes and the need for reliable and trustworthy AI.

2 THE FAIRWORK INNOVATION SHOP

The main contribution of this deliverable is the established research collection, consisting of the research results created by the partners. These results can take various forms, like tools, prototypes, experiments, concepts, questionnaires or similar. These created artefacts are not only used within the FAIRWork project and its DAI-DSS, but these results are also used by the individual partners who created them within FAIRWork's individual exploitation efforts. To support the communication of these research results during the project and their exploitation after the project, FAIRWork deployed its own innovation shop:

<https://innovationshop.fairwork-project.eu/>

The FAIRWork Innovation Shop is an online platform for publishing project results as self-contained and individual artefacts, which we call innovation items. The Innovation Shop differs from marketplaces or web shops, as its primary goal is not to sell the items but to make them accessible to more interested parties.

Therefore, published innovation items can be on varying TRL, ranging from tested software products to early prototypes or described concepts. Innovation items are not only software but all research results created, like methods, concepts, studies, questionnaires or success stories. As innovation items can be any project result, the term *Exploitation Item* is used for a subset of innovation items that are used for the individual or joint exploitation of the project partners. Exploitation items, therefore, tend to have a higher TRL, as they will be used by the project partners after the project ends. The strategy for defining Innovation or Exploitation Items for research results prioritizes the enhancement of existing items, by ensuring that the contributions contribute and add value to these areas. New items are created when the research results do not align well with the existing Exploitation items and explore new topics.

The innovation shop and the concept of the innovation or exploitation item are used to make the project results easily accessible to third parties and to support the project partners in disseminating and exploiting their results. They are self-contained, meaning they are described so that they are understood without a comprehensive understanding of the FAIRWork project. Therefore, a description of the item is provided, as well as links to further materials, e.g., the link to a scientific paper, to the source code where it can be used or data that was used or generated out of it. The self-contained description eases the adaptation of the defined innovation items, as they can be changed independently without the need to check for consistency with all the others.

Using the innovation shop additionally supports to tackle problems which often occur when developing exploitable artefacts out of research results within a research project. One is that the time when the items are requested typically not match the time when they are provided, therefore there is a need to safely store and evolve the research results until they are needed.

Further, research results often possess a different maturity and level of detail, needed for artefacts that can be exploited. Here additional effort must be made, to make the artefacts usable and maintainable, which do not directly result from core research contribution. Here the innovation shop enables to publish the research results on an early stage to already use them to draw attention towards the results, as each item contains the responsible project partner and contact information.

Defining the research results in a self-contained way enables the dissemination within the FAIRWork project and allows the reuse of the description to publish on other platforms and communities in the future. Examples could be

Adra's AI-on-Demand platform (<https://www.ai4europe.eu/>) or experiments within OMiLAB's Community of Practice (www.omilab.org).

The rest of this document contains description of research results, created during the second half of the project. Table 1 contains an overview of the Innovation and Exploitation Items which are influenced by the blow described research results. The first two columns contain the item name and its link, the third column the project partner responsible for the item and the last column the sections, where research results are described which influence the item.

Exploitation/Innovation Item	Innovation Item Link	Owner	Section Reference
Democratization in Industry via MAS: Case Study Approach	https://innovationshop.fairwork-project.eu/items/19/	RWTH	3.2 and 5.2
Process Maestro	https://innovationshop.fairwork-project.eu/items/9/	MORE	3.3 and 4.5
AI models research and development.	https://innovationshop.fairwork-project.eu/items/8/	RWTH	3.6.2, 3.6.3 and 3.6.4
Optimisation Toolbox	https://innovationshop.fairwork-project.eu/items/10/	JR	3.5.3 and 3.5.4
Consulting Services with Human Factors Lab for Production Environments	https://innovationshop.fairwork-project.eu/items/12/	JR	3.4
Intelligent Sensor Box	https://innovationshop.fairwork-project.eu/items/11/	JR	3.4
Service to extend Process Modelling for AI	https://innovationshop.fairwork-project.eu/items/2/	BOC	3.7.1
OLIVE Microservice Integration Framework	https://innovationshop.fairwork-project.eu/items/1/	BOC	3.7.2
User Centric Services to introduce AI into companies	https://innovationshop.fairwork-project.eu/items/3/	BOC	4.4
AI Transparency for Trust	https://innovationshop.fairwork-project.eu/items/7/	RWTH	3.8.2, 3.8.4 and 3.8.6
Scene2Model	https://innovationshop.fairwork-project.eu/items/14/	OMiLAB	3.7.3

Table 1: Innovation shop in FAIRWork.

3 RESEARCH METHODS AND SERVICES

3.1 Overview of Methods and Services

Providing the scientific foundation for the FAIRWork project, the research tracks in Figure 1 outline the context in which its diverse research activities unfold.

The human is at the centre of the decision support system- as a decision maker, a worker, and most importantly, a driving force in the democratisation process within industrial socio-technical settings. The **Democratisation of Decision-Making** is a fundamental, central and highly innovative research theme that is addressed from the very beginning. Various aspects of Human Factors are essential to human-centered socio-technical setting within a digital system architecture. In FAIRWork, we are particularly focusing on the benefits of the **Digital Human Factors Analytics** based on the collection of **wearable biosignal sensor data** in various environments, such as, the exploratory ambiance of the Human Factors Laboratory, however, with the objective to measure directly at work in the manufacturing settings.

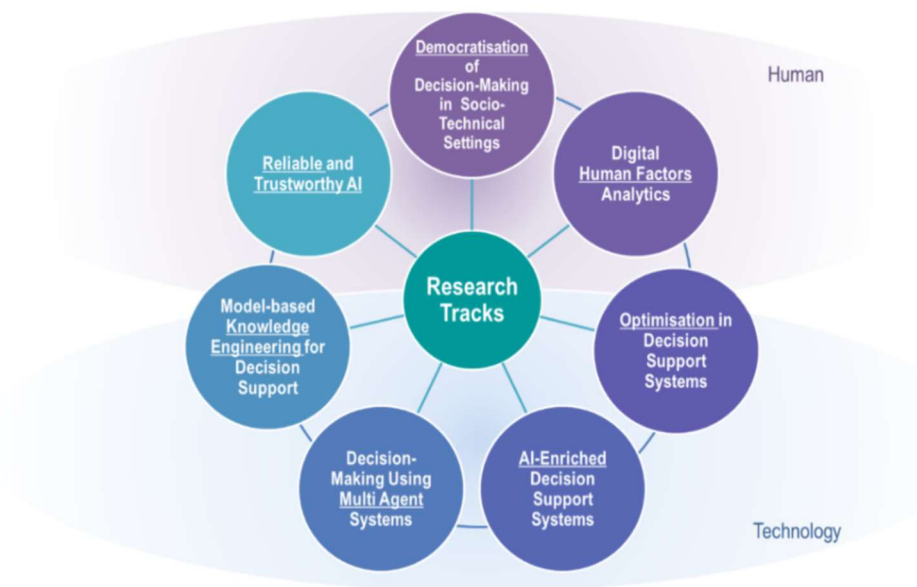


Figure 1: Research tracks underlying the outline of the research collection.

The technological aspect of the project is primarily reflected in the **Optimisation in Decision Support Systems**, a central research area that integrates relevant input data on human behavioural status and various system data with objective functions to provide meaningful guidance for decision-making and system processes. The **AI-enrichment of the Decision Support Systems** provides intelligence, such as adaptiveness, reasoning and machine learning solutions to the decision-making process. Another approach for the development of intelligent systems is represented by the research track provided by **Decision-Making using Multi-Agent Systems**. In addition, the project FAIRWork focuses on **Model-based Knowledge Engineering for Decision Support** and, with this strategy the project enables to complement AI-based and multi-agent-based systems.

Finally, there is a strong focus on **Reliability and Trustworthy AI**. This research track explores and identifies the requirements for trust and acceptance of AI and ethical integration of technologies within socio-technical settings.

3.2 Democratization of Decision-Making in Socio-Technical Settings

3.2.1 Overview

In the realm of AI, democratizing decision-making entails promoting democratic practices throughout the development, implementation, and utilization of technologies. This process necessitates an analytical approach that considers both social and technical factors. Its aim is to ensure that AI technologies contribute to enhanced democratic decision-making processes. Achieving this involves exploring methods for democratic control over these technologies and understanding how they can foster democratic practices (Noorman & Swierstra, 2023)¹. To achieve this goal, FAIRWork project has designed and implemented the DAI-DSS, integrating various technologies to support decision-makers (Woitsch et al., 2023)².

To delve into democratic decision-making within a socio-technical framework and to explore the contextual situation for implementing a DSS within a company, a case study approach was adapted, which facilitated an in-depth analysis of the human factors in the project. The case studies enabled a comprehensive investigation of our use cases and explored the potential demand for the democratic design of the decision-support tool.

For this purpose, a three-step procedure (see Figure 2) was identified, which includes steps of onboarding, precision and contextualization. In the **Onboarding Phase**, the focus was on an on-site visit from the lead case to set the foundation of the empirical case study, gain an overview, and collect primary data for initial analysis. The case study focused on a detailed exploration of the use case involving document analysis, on-site observation, and worker interviews. Through the analysis of the case-study, three key dimensions were identified: Decision-Making, Involvement and Expectations³, which were explained in detail in Deliverable 3.2, and can also be found as an innovation item in Section 2.

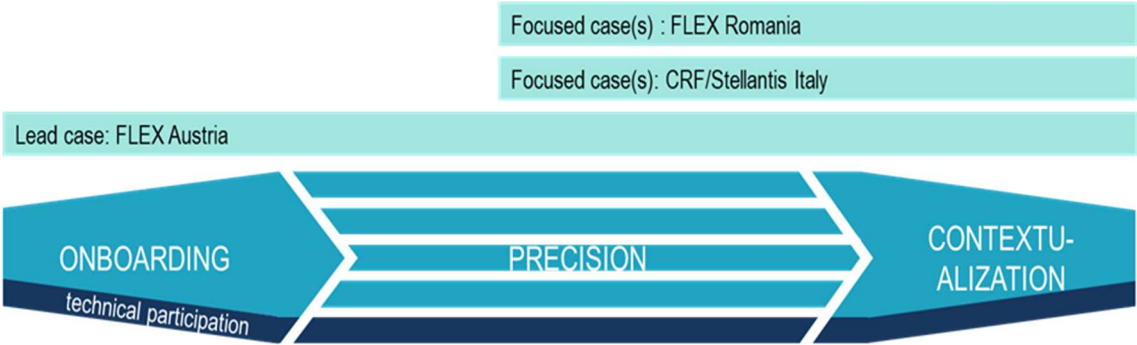


Figure 2: Extensive research plan for the investigation of human aspects in AI-guided decision-making in FAIRWork.

To gain deeper insights into the human perspective and to explore additional options for supporting the democratic implementation of the DSS, the **Precision Phase** was conducted on July 15, 2024. This phase centered on FLEX, our lead case located in Althofen, Austria, to ensure alignment and address potential language barriers. The objective was to deepen our understanding of the company and investigate the dynamics of democratization in industry through MAS. To this end, a comprehensive case study was planned, including on-site observation, worker interviews, and a workshop.

Building on the dimensions explored during the onboarding step, a fourth dimension - challenges - was identified through the analysis of the findings from the second case study (Figure 3). Using each method employed in this study, we recognized a broad spectrum of prominent challenges in the company, ranging from production line issues to interpersonal matters. To clarify these challenges, we grouped the identified issues into three categories: Resource Allocation, Production Management, and Personal competences. Particularly in the workshop, held in the form of a focus group interview, some underlying issues faced by the workers and the company were uncovered, which will be explained in more detail in the results section.



Figure 3: Dimensions explored in the qualitative evaluation of FLEX Althofen.

The final step in our empirical research is the **Contextualization Phase**, where we aim to assess whether the findings are systematic and determine their feasibility for implementation by industry partners. This phase focuses on validating the results from the previous phases to understand how they can support both industry and service partners. While the validation with industry partners is still ongoing, we are currently focusing on the feedback from our service partners. Given that the critical question of the project was how the DSS can be democratic, we conducted a workshop where partners could actively participate in providing an answer. The online workshop was held using a virtual collaboration board, consisted of two tasks:

- The first task asked the collaborators to match the democratic AI features (identified through the case study) with their services and explain how each feature could be provided.
- The second task listed the expectations gathered from the workers and asked our colleagues to indicate which of these expectations could be met through their designed services. This research allowed, moreover, to specify the expectations regarding criteria of democracy in companies (see Section 5).

3.2.2 Results

The evaluation of the interviews led to the identification of the new dimension of challenges, which was further categorized into three main areas. The first category, Resource Allocation, was broken down into four subcategories: Responsibility Allocation, Worker Allocation, Training Allocation, and Component Allocation. The second category, Production Management, was divided into two subcategories: Workflow and Product. Finally, the third category, Personal Competences, encompassed subcategories related to Skill Sets, Social-contextual Learning, and Supply Chain Insight.

Resource Allocation

Starting with the first category, we identified several issues and desires related to Resource Allocation. Among these, the issues related to the Responsibility and Worker Allocations were the most significant. Regarding Responsibility Allocation, employees from different hierarchical levels emphasized the importance of balanced team contributions. Additionally, some experienced workers noticed the need for better alignment between responsibility and competence, particularly among the younger generation. One of the prominent issues regarding responsibility assignment is the employees' reluctance to make decisions independently, which was repeatedly mentioned in the first step of the case study as well. Meanwhile, the employees stated that they do not trust the tool's decisions when made autonomously.

Apart from the challenges related to responsibility allocation, Worker Allocation in the company also faces specific issues. As we discovered in the previous step, worker allocation is a time-consuming task for supervisors, and reallocating workers due to frequent and sudden Paid Time Off (PTO) makes this task even more challenging. Unclear workforce information further complicates worker allocation. Worker allocation could also impact employees differently, depending on task difficulty, perceived fairness, or interpersonal dynamics. Due to the mentioned issues, supervisors and managers play a key role in transparently communicating their decisions, especially when it comes to worker allocation. Additionally, the loss of temporary employees was a concerning topic for all staff, which came up several times in the interviews, as well as in conversations during on-site observation and even in the company canteen.

Resource allocation issues extended into Training Allocation, with some aspects interpreted as conflicting with each other. We found that some experienced employees were highly motivated to improve themselves, learn new skills, and engage in diverse tasks, though opportunities for further training were limited. Meanwhile, employees on the top floor highlighted their efforts to encourage cross-training. Another aspect in this context is the opportunity to train operators across different areas, enabling every operator to confidently handle medical devices and manage the production line related to the medical hall. The issues surrounding training allocation are often made more challenging by the overload amount of theory and practice that newcomers and apprentices must take in.

The final factor in resource allocation that caught our attention was Component Allocation, identified as a significant issue within the company. Employees across different hierarchical levels acknowledged the challenges in component placement, emphasizing the necessity of applying the four-eyes principle. They also suggested that a digital tool could help resolve this issue by providing notifications about the necessary components or actions.

Production Management

To delve deeper into the challenges faced by the staff in our case, we turn to the second explored group: Production management encompasses various tasks, including planning, overseeing, and controlling the production process, with the goal of delivering the right quality at the right time while minimizing costs. Based on our findings, which stem from employee criticisms, obstacles, and limitations, we categorized these challenges into two subgroups: workflow and product.

Focusing on the Workflow, inadequate system functionality was identified as a well-known problem. This issue occurs so frequently that employees consider it a default condition within the company, asserting that machines cannot completely replace humans. Another issue related to the workflow is dealing with multiple systems with different procedures, which makes handling and familiarizing with processes difficult. This results in delayed adaptation to the workflow, particularly for younger employees. The diverse operations of each production line may

cause quality issues due to minor mistakes, as the machines might produce the same product but operate differently. The workers, especially those at the entry level, faced serious problems regarding the operation of the production lines, particularly when it came to errors and defects. Further on this topic, the workers pointed out the extended error identification in machinery, as the machines sometimes remain in error for an extended period before being detected. This issue, seen as trivial by younger employees, was thought to be easily solvable with system support. However, implementing a support system requires certain prerequisites, such as the availability of digital forms of necessary parameters. Therefore, limited digital data availability can be considered another challenge in the workflow.

The products in the company have a wide variety, ranging from industry to automotive and healthcare sectors. Thus, it is not surprising that some challenges in production management are related to the Product. Besides the diversity of production lines, the different types of products are also considered by employees as a difficulty in the workflow. As mentioned above, this process variation could affect the quality of products, particularly if the quality checkers do not master the entire system. The issues related to the product can be traced back to unclear quality boundaries, which results in a product interruption. The workers suggested that providing support for cross-checking and identifying whether the products are good or defective would be a good idea.

Personal Competences

Challenges at this company extend to personal competencies, including employee skills, the organization and management of those skills, and the quality of skills needed to address production line issues. According to the findings, three subgroups – Skill Set, Social-contextual Learning, and Supply Chain Insight – were recognized, which encompass all the mentioned items from the employees.

In terms of Skill Set clarifying employee competencies, in line with the previously mentioned workforce information, would support more effective decision-making. Limited access to workforce competencies was also pointed out by skilled workers, who mentioned that access to competencies is possible by asking. The need for a skilled workforce, particularly in the medical area and quality control, was also emphasized as an area for improvement, as discussed earlier in the training allocation section.

Focusing on the fact that environmental factors can shape the level and quality of skills, challenges regarding Social-contextual Learning attracted our attention. The important issue in this category was to harmonize learning approaches, as different mentors had varying perspectives on what their operators should or should not do. The influence of context in learning or adopting a new method is so significant that employees in management view early involvement as essential for the success of any innovation; otherwise, it may fail.

Working closely with this company during the first and second steps of the research led us to conclude that managing the complexities of different production lines requires the competence of Supply Chain Insights. Therefore, nuanced quality knowledge is essential for being recognized as a skilled worker. Apart from quality knowledge, employees need multifaceted system knowledge to address issues arising on production lines.

3.2.3 Outlook

While our findings from the empirical research contribute to a deeper understanding of the democratization of decision-making in socio-technical settings, further research, particularly case studies, could build on these insights to complete the validation step and assess the feasibility of our results in collaboration with industry partners. As FLEX undergoes its first evaluation step with the DSS, it would be highly beneficial for FAIRWork's empirical

process to examine the democratization process after the implementation of MAS for decision-making and explore whether workers' expectations and challenges are effectively addressed.

3.3 Decision-Making Using Multi Agent Systems

3.3.1 Overview

The rapid advancement of Industry 4.0 has ushered in an interconnected environment characterized by the decentralization of computational power. Decision-making processes in industrial environments have increasingly integrated MAS as a core framework for addressing decentralized and complex challenges. Industries can enhance system efficiency and worker participation. Recent research highlights the transformative potential of MAS in various domains, with a strong focus on technical, and human-centric, and ethical dimensions.

Ethics is an important aspect in the design of MAS, particularly within industrial cyber-physical systems (CPS). Ethical behavior encompasses both system operations and stakeholder interactions. Addressing the growing challenge of integrating ethics into industrial systems is a fundamental point in human-centered systems (Trentesaux et al., 2022)⁴. Integrating ethics into agent systems allows these systems to support decisions that are not only efficient but that also fosters trust. The ethical dimension extends to autonomous agents as decisions can have significant impacts on individuals (Cervantes et al., 2020)⁵. Agents are equipped with mechanisms to manage ethical issues across various contexts, supporting human decision-making processes by providing support to decision. Ensuring that MAS align with human values and ethical aspects is essential for fostering trust and mitigating risks associated with their integration into human collaboration in the industry. MAS formalism has been increasingly adopted for the realization of decentralized control systems, providing a robust framework for modeling, simulating, and optimizing inherently decentralized systems such as supply chains (Răileanu & Borangiu, 2023)⁶. This interconnected framework facilitates dynamic decision-making, fostering adaptability and scalability in complex industrial environments.

Industry 5.0 highlights the need for integration of human factors with CPS in order to develop cooperative sustainable environments, emphasizing ethical and human-centric innovation. The symbiotic relationship between human intelligence and cognitive computing aims to enhance human capabilities while embedding ethical principles in technological design (Longo et al., 2020)⁷. Through the positioning of humans as active collaborators alongside autonomous systems, Industry 5.0 redefines the factory of the future, ensuring that automation complements rather than replaces human expertise. The synergy between human and agent decision-making capacities is a core challenge in MAS development (Gal & Grosz, 2022)⁸. Humans excel in reasoning and contextual understanding, whereas agents are adept at processing vast data sets. Effective MAS design respect and leverage these complementary strengths, enabling collaborative decision-making processes that integrate human values and computational capabilities. Algorithms that combine the performances of humans and agents have demonstrated superior outcomes compared to isolated autonomous systems. The increasing collaboration between humans and machines introduces ethical risks, including potential dependence on automated decisions and the erosion of human expertise, it is critical to mitigate these risks by ensuring transparent and explainable systems (Pacaux-Lemoine & Trentesaux, 2019)⁹. MAS prioritize human oversight and provide mechanisms for evaluating and validating automated decisions to maintain trust and effectiveness in human-machine cooperation.

The adoption of MAS in decision-making processes offers significant opportunities to enhance industrial systems through decentralization, human-centric design, and ethical integration. MAS enables industries to create systems that are efficient, inclusive, and aligned with human values, addressing ethical challenges, balancing human and agent capabilities, and fostering symbiotic relationships. As technological innovation in the factory of the future

continue to reshape the industrial landscape, the integration of ethics with MAS into human and cyber decision processes, design will remain a relevant factor in achieving sustainable and equitable progress.

In this research, we aim to address the ethical aspects through the exploration of a watchdog agent that is capable of monitoring parameters regarding the decision-making process that are ethically relevant to the final decision and all individuals involved. The incorporation of a watchdog agent aligns with the proposition of ethics-aware systems, where ethical behavior in autonomous systems must be subject to verification and validation. When monitoring the decision-making parameters, the watchdog agent ensures that the system operates within predefined ethical boundaries, promoting a human-centered design approach to the decision-making processes. The exploration of a watchdog agent aims to embed ethical considerations into the decision-making processes using MAS. This ensures that as these systems become properly integrated into industrial applications, they operate responsibly and ethically, ultimately contributing to human well-being and adhering to societal values.

The developed service promotes resource allocation in industrial settings using MAS. Its purpose is to support decision-making in a decentralized manner while prioritizing a human-centric approach by incorporating human-relevant data in a practical use case. The service retrieves pertinent information from the Knowledge Base and integrates with the Orchestrator to facilitate workflows and data exchange in a cohesive architecture. Positioned within the optimization domain, it accounts for supporting decision-making in multi-agent perspective while takes advantage of human factors for a human-oriented approach.

3.3.1 Motivation and Reference to FAIRWork Use Case

The adoption of MAS for supporting decision-making in industrial processes arises from the increasing demand for efficient, scalable, and adaptive solutions capable of addressing the application of human values in real scenarios. Industrial processes inherently comprise a network of interconnected components and diverse stakeholders, each with distinct and sometimes conflicting objectives. Achieving robust outcomes in such settings necessitates sophisticated approaches that can holistically address these challenges. In this context, MAS stands out by offering a decentralized and dynamic framework that models and manages the interactions between stakeholders adding value through the digitalization of these interactions. With its ability to enhance decision-making by digitalizing and structuring the interaction processes between stakeholders, integrating human-centric considerations, such as individual preferences, roles, and contextual factors, become facilitated into the decision-making framework. MAS can balance technical efficiency with the requirements of human actors. In essence, MAS provides a transformative approach to decision-making in industrial processes, facilitating collaboration, improving resource allocation, and fostering human-centric outcomes while addressing the complexities of modern industrial ecosystems.

3.3.2 Innovation and Research Activities around Multi-Agent Systems

The integration of MAS incorporating human factors data into decision-making into workload balance represents a significant innovation, advancing beyond the state-of-the-art. Unlike traditional systems focused solely on productivity, MAS enables a human-centered approach by modeling diverse stakeholders and their interactions, aligning resource allocation decisions with both production goals and worker well-being. When leveraging autonomous, decentralized processes, MAS simulates complex social interactions and integrates a broad spectrum of inputs, including human conditions, preferences, and well-being metrics. This fosters decision-making that is not only efficient but also fair, adaptive, and inclusive, aligning with Industry 5.0's vision of socially responsible and technologically advanced industrial environments.

This novel application of MAS introduces a paradigm shift, transcending operational optimization to prioritize equitable workload distribution and increased worker satisfaction. MAS enhances human participation in governance processes, creating democratic decision frameworks even in constrained industrial settings through balancing relevant human considerations with production demands. It demonstrates the potential to redefine industrial decision-making by harmonizing technological efficiency with human-centric values, paving the way for sustainable and socially responsible advancements in resource allocation and workforce management.

Additionally, the development of ethical watchdogs within MAS represents an innovation that introduces mechanisms for embedding ethical oversight directly into decentralized decision-making processes. These watchdogs act as autonomous agents designed to monitor and alert to the ethical implications of decisions made within the system, ensuring alignment with predefined human-centric values and societal norms. Ethical watchdogs help mitigate biases and safeguard against unintended consequences in resource allocation or workload distribution, where computer agents and humans (Gal & Grosz, 2022)¹⁰ share decision-making in order to address ethical conflicts (Belloni et al., 2015)¹¹. This capability allows MAS to uphold ethical standards in dynamic and complex industrial settings. The integration of the watchdog advances the field by operationalizing ethical principles in decision-making, offering a novel approach to fostering accountability and trust (Woodgate & Ajmeri, 2022)¹² in technology-driven environments, and reinforcing the alignment of industrial processes with broader societal and human values.

3.3.3 Description of Functionality and Results

The service is integrated into the system's architecture where it exchanges data with the Knowledge Base and Orchestrator. Utilizing an algorithm embedded within the MAS, the service suggests to the decision-maker potential solutions to the allocation challenges presented. Prior to each shift, it provides a recommendation for worker allocation by analyzing relevant data pertaining to the available workforce, the specific requirements of the tasks to be performed, and the characteristics of the products and production lines involved. An Ethical Watchdog, configured to track relevant desired parameters in the decision process, alerts to the infringement of such parameters allowing the decision-maker to trigger a renegotiation of worker allocation. This data is processed within the algorithm to identify a feasible allocation configuration that abides by the defined threshold in order to fulfill the established ethical requirements while considering the multitude of possible alternatives, given constraints and objectives.

The MAS-based resource allocation service demonstrated its capability to support decision-making in industrial workforce management by dynamically matching worker profiles with production line requirements. The system consistently generated allocation recommendations by ranking workers based on resilience and preference metrics. This approach not only provided resource allocation through agent interactions and negotiations but also ensured that human-centric values were maintained throughout the decision-making process, thereby reinforcing the system's relevance in complex industrial environments.

The development of the multi-agent allocation process within an industrial manufacturing setting has demonstrated potential improvements in workload distribution and decision-making support. The approach integrates principles that emphasize fairness, worker considerations, and digital representation of relevant factors in the allocation process. Prior to assignments, factors such as worker availability and physical condition where operational requirements are assessed to determine workforce distribution. Oversight mechanisms are in place to monitor and adapt worker placements according to ethical relevant aspects. A multi-agent-based approach considers factors that influence task suitability, incorporating representation of human stakeholders in an interactive decision-making

process. If a worker does not meet the conditions necessary for a given assignment, alternative allocations are explored. The process can be adjusted dynamically based on relevant parameters.

The decision-making process culminated in an integrated approach that combines multiple metrics to promote fairness and balance. Workers' preferences for different assignments are factored into the allocation process, ensuring that individual needs are considered alongside task suitability. Moreover, the incorporation of negotiation mechanisms enhances accountability, solidifying a robust allocation framework. Overall, the incorporation of human-centric considerations in a multi-agent approach for decision-making processes offered support in workforce management. The approach offers insights into balancing operational needs with workforce factors, contributing to more informed decision-making in dynamic environments.

3.3.4 Outlook

MAS are relevant for modelling individual users (agents). This is highly relevant in the context of industrial human resource management, where employee profiles are dynamically matched to the requirements of the production line. This service uses an algorithm to propose staffing solutions before each shift, analyzing the availability of workers, the requirements of the tasks and the characteristics of the products/production lines.

It is interesting to pursue the idea of an 'ethical watchdog' to check the ethical suitability of individual functions in an agent system. These can be parameters, but also outcomes. They should help decision-makers to become aware of possible violations. As a consequence, they can also trigger a renegotiation of allocations, as in our example, taking into account ethical requirements.

The integration of multi-agent systems with agent workflows and generative AI is a broad and promising area of research. MAS can break down complex tasks into smaller, more manageable subtasks that can be handled by individual agents. Agent workflows define the interaction and coordination between these agents to ensure a smooth process. Through the use of generative AI, these agents can be empowered to develop new solutions and adapt dynamically to changing environments.

3.4 Digital Human Factors Analytics

3.4.1 Overview

The work in FAIRWork on digital human factors analytics emerged from the conceptual framework on Industry 5.0 to connect with an inherently socio-technical dimension, demanding attention to the wellbeing of workers, the need for social inclusion and the adoption of technologies that do not substitute but rather complement human capabilities. Building upon the digitalisation of the industrial processes we focus on the workers' well-being, empowering workers using unobtrusive digital devices, endorsing a human-centric approach to technology.

Sustainability, human-centricity, and resilience are the hallmark features of Industry 5.0 (European Commission, 2020)¹³. The worker is not to be considered as a 'cost', but rather as an 'investment' position for the company, allowing both the company and the worker to develop. This implies that the employer is interested in investing in skills, capabilities, and the well-being of its employees, to attain its objectives. Mental health and well-being must be considered on an equal footing when designing digitalized workplaces.

While there are new risks associated with digitized ways of working, such as the risk of burnout due to the always-online and always-available working culture, digital technologies could be used to support workers in better

controlling and managing the risks and impact of the new working environment on their mental health and well-being. Digital solutions and wearables could open new channels for alerting workers and their general practitioners about critical health conditions, both physiological and mental. They could also support workers in adopting healthy behaviors in the workplace. This is, moreover, likely to bring economic benefits and savings due to productivity gains and avoidance of accidents, long-term illness, and absenteeism.

FAIRWork brings human, AI, data, and robots together by supporting decision-makers in making decisions thus positively affecting the work balance between workers and machines. One key aspect for the daily decision-making on worker allocation in production processes is to consider the resilience of individual workers in the context of fostering well-being and avoiding illness and absences (Paletta et al., 2023)¹⁴.

Resilience is a meaningful adaptation in persons' psychological traits and experiences that allows them to regain or remain in a healthy mental state during crises without long-term negative consequences (Southwick et al., 2014)¹⁵. Resilience has shown that it plays a crucial role in promoting mental health and well-being: resilient people are better equipped to navigate situational challenges, maintain positive emotion and motivation, and recover from setbacks. They demonstrate higher levels of self-efficacy, optimism, and problem-solving skills, which contribute to their ability to adapt and thrive in adverse situations.

In the proposed work we principally understand resilience as the ability to respond to stress (Smith et al., 2008)¹⁶, however, we particularly focus on the impact of chronic stress on the reduction of resilience resource capacity (Schetter et al., 2011)¹⁷. The technical objective of our work is to compute resilience scores from long-term strain tendencies being estimated from wearable biosignal sensors. For this purpose, the architectural construct of the Intelligent Sensor Box (ISB; Paletta et al., 2023)¹⁸ enables the measurement of worker's psychophysiological strain while performing tasks and provides information about the workers' estimated resilience. It consists of a framework for a set of stationary and wearable sensors, AI-based analytics for assessment and optimization functions. It can be applied to evaluate the ergonomics and design of industrial training and work environments.

The innovative contribution by digital human factors analytics firstly focussed on the estimation of human resilience as a functional of **long-term stress monitoring**, and its relation to specified use case of worker allocation in the industrial environment (FAIRWork Deliverable 3.2). In this context, we presented an initial stage of a complete model on resilience (Paletta et al., 2024)¹⁹. This model was augmented based on further research on wearable biosignal data, would include additional sensors, such as, a smart shirt as well as eye tracking glasses, for further refinement based on multisensory-based assessment of resilience scoring.

The conceptual framework of the **resilience risk stratification model** (RRSM) is presented in Figure 4. It illustrates our hypotheses on how the accumulation of the negative consequences of stress has a cyclical nature and how it can contribute to a loss spiral. This framework is based on the Transactional Model of Stress and Coping (Lazarus & Folkman, 1987)²⁰, the Job Demands-Resources Model of Burnout (Bakker & Demerouti, 2007)²¹, the Effort-Recovery Model (van Veldhoven, 2008)²², the Conservation of Resources Theory (Hobfoll, 2001)²³, and the WearMe project (deVries et al., 2019)²⁴.

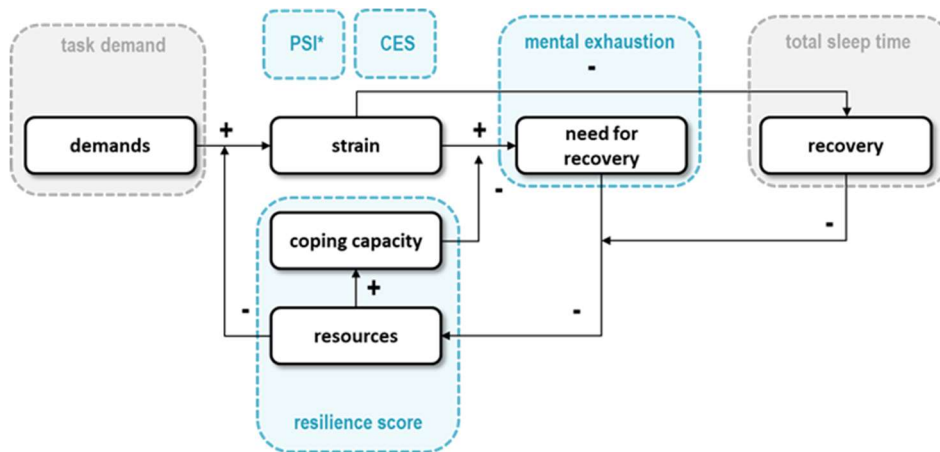


Figure 4: Stages of the computation of the resilience score that underlies the risk stratification model (RRSM).

The RRSM model was developed in FAIRWork to represent a measure of mental exhaustion in terms of the daily total strain score as a function of the strain data from wearable sensors (Figure 4). The accumulating effect of mental exhaustion is then represented by another functional that integrates daily score contributions within a predefined extent of recency. The resilience score underlying the risk stratification is then further outlined by an inverse function of the mental exhaustion. This score implicitly represents a tendency of the long-term stress dynamics rather than a short-term response-based construct. In this context, the framework includes a cyclical nature that is supported by the Conservation of Resources theory (Hobfoll, 2001)²⁵, which states that long-term loss of resilience resources increases one’s vulnerability to stress, and, since additional resources are necessary to battle stress, this may lead to a depletion of resources in a loss spiral. The motivation of the development of this RRSM framework is to prevent this loss spiral for the benefit of the worker as well as the economic impact of the manufacturing company.

The RRSM is of central importance for the allocation of workers for specifically stressful work. Persistent stressful work can have an impact on the mental exhaustion, and this is an important parameter for the overall resilience risk stratification as a key objective in the work of Digital Human Factors Analytics. The resilience score would indicate levels of risks for decision support to the manager that assigns work to workers and can have an important impact on the entire economic situation of the manufacturing company. Finally, these scores can provide a relevant input to optimization routines that would provide higher long-term benefits to the worker, to the company and ecologically relevant aspects (Paletta et al., 2024b)²⁶.

In the novel and second major contribution to the modelling of resilience we aim at the quantification of the **recovery status** (Kellmann & Kallus, 2024)²⁷ via monitoring strain and subsequent fatigue in the individual worker. Based on these observations, recovery-related processes, such as, sleep, and activities can be initiated. By measuring the frequency of current stress symptoms along with the frequency of recovery-associated activities, the recovery-stress state can be defined. As a set of complex processes in time, recovery takes place physiologically and psychologically with the aim of regaining a balanced psychophysiological state (Heidari et al., 2019)²⁸ Because the need for recovery varies inter-individually and intra-individually and depends on internal and external influences, the monitoring of recovery requires a well-defined and structured framework. Recovery does not only play a pivotal role in the restoration of lost resources and fatigued states to guarantee readiness for performance (Halson,

2014)²⁹. It is also important in minimising the risk of negative outcomes such as overstraining and psychological disorders, such as, depression.

The importance of monitoring recovery can best be explained by its relationship with performance and the negative consequences of too much strain without appropriate recovery. Continuous non-functional overreaching (NFO) together with an emerging state of under-recovery characterises a development of decreasing performance and well-being. Although NFO and under-recovery share many commonalities, they should be considered as distinct concepts (Kellmann & Kallus, 2024)³⁰. Under-recovery describes a broader condition of insufficient recovery related to general, psychophysiological stress aspects. If this downward spiral of excessive application of stressors, such as, psychological and social stress, and insufficient recovery is not identified and stopped early enough, a state of overstraining may manifest as an ultimate consequence. High-level motivated work is no longer possible while affected workers experience severe psychological – e.g., a-motivation, irritability, anger - and physiological - such as, immunosuppression, cardiac disturbances – symptoms (Meeusen et al., 2013³¹; Jiménez et al., 2016³²).

The “**bounce-back effect**” (Smith et al., 2008)³³ is a critical aspect of resilience that reflects the ability of an individual to recover and return to a stable or functional state after experiencing stress, disruption, or adversity. In particular, it relates to the following aspects that are highly relevant to resilience computing:

- **Measurement of Core Resilience Aspects:** Resilience is not just about enduring challenges; it's about how effectively and quickly one can recover from them. The bounce-back effect captures the recovery phase, which is the most telling indicator of true resilience. For example, a resilient individual recovers emotionally and mentally after a personal setback, regaining productivity and well-being.
- **Key to Adaptability:** The bounce-back effect highlights the adaptability of a person. The faster and more effectively the bounce-back occurs, the better the ability to adapt to changes and unforeseen challenges. An individual, adapting to new circumstances, such as adjusting to a major life change, shows a strong bounce-back capacity.
- **Indicator of Systemic Strength:** In resilience engineering or organizational contexts, the bounce-back effect indicates the underlying robustness of systems. It demonstrates whether the system can withstand shocks without long-term detrimental effects.
- **Predicts Long-Term Success:** The ability to bounce back is often predictive of long-term sustainability and success. Individuals that can recover quickly are more likely to thrive in dynamic and challenging environments. In athletes, the bounce-back effect from injuries or defeats often predicts their future performance and longevity in their careers.
- **Mitigation of Secondary Risks:** The bounce-back effect prevents cascading failures or additional negative consequences that might arise if recovery is slow or incomplete. After an illness, a rapid return to normal physiological functioning reduces the risk of complications.
- **Enhances Mental and Emotional Well-being:** In humans, a strong bounce-back effect is tied to better mental health outcomes. It promotes positive coping mechanisms and prevents chronic stress or burnout. Resilient individuals may experience stress but recover quickly, maintaining overall psychological stability.

The bounce-back effect is central to resilience as it encapsulates recovery, adaptability, and the capacity to thrive after adversity. By focusing on this effect, resilience strategies can be designed to strengthen individual workers, ensuring they can navigate challenges with minimal long-term impact.

In the modelling of resilience risk stratification, we will finally model the quantification of bounce-back effects using wearable biosignal sensors as well as machine learning methodologies. Based on a field trial that will be applied at the use case partners, i.e., at the plants of Stellantis and FLEX, we will gain sufficient data in a long-term measurement in order to provide rather precise estimates of resilience in terms of bounce-back as well as long-term exhaustion aspects.

The concept of “Intelligent Sensor Box (ISB)” (ISB; Paletta et al., 2023)³⁴ was introduced to integrate relevant human-centred digital sensing into a larger framework of decision-making in production environments. The human data are firstly determined by a dedicated body sensor network, including low-cost sensors, such as, biosignal sensors, wearables, human sensors, or even virtual sensors. Specific attention is dedicated to developing the “Digital Human Sensor” (DHS) applying AI-enabled Human Factors measurement technologies. Each instantiation of a DHS may provide a digital vector of Human Factors state estimates, such as digital biomarkers representing the assessment of physiological strain, affective state, cognitive workload, fatigue, situation awareness, etc. The internal architecture of the ISB supports the data flow from human- and workplace-mounted sensor data to higher abstractions of Human Factors, i.e., dominantly ergonomic, and psychophysiological constructs that determine the mental state and behaviours of the human, and from this impact the sociotechnical systems within the production process. The integration of human-centred data into the overall decision-making process needs the anonymization of the highly vulnerable psychophysiological data. However, personalized data are of interest to the individual worker or decision-maker and are available for individual insight on a security- and privacy-preserving basis.

3.4.2 Acceptance Analysis on Wearables Sensors in Production Environments

Introduction

We present the concrete development of a set of wearable sensor technologies together with the ISB-dedicated software architecture that enables monitoring and analytics to study resilience scores at the production site. The wearables (Figure 5) include a “Garmin vivosmart 5” fitness tracker to provide heart rate (HR) and heart rate variability (HRV), the greenTEG (“CORE”) core body and skin temperature sensor to be attached to the chest, as well as, optionally, Pupil Labs “Neon” eye tracking glasses to provide eye tracking data with 200 Hz sampling rate as well as a “QUS” biosignal shirt of sanSirro GmbH for measuring HR, HRV and breathing rate.



(a) vivosmart 5 fitness tracker, Garmin Ltd.



(b) QUS smart shirt, sanSirro GmbH



(c) Eye tracking glasses, Pupil Labs GmbH

Figure 5: Wearable biosignal technologies for the production environment for studies and daily monitoring (credit: JR).

Study

A decisive issue is the acceptance of using the wearables for biosignal data acquisition by manufacturing companies' workers and managers. We firstly presented the wearables to 19 (m=11, f=8) employees of the manufacturing company FLEXTRONICS INTERNATIONAL GmbH, in the city of Althofen, Austria, and then issued a questionnaire about the future use of these wearables during daily work shifts. 13 of the participants were shop floor workers and 6 from management.

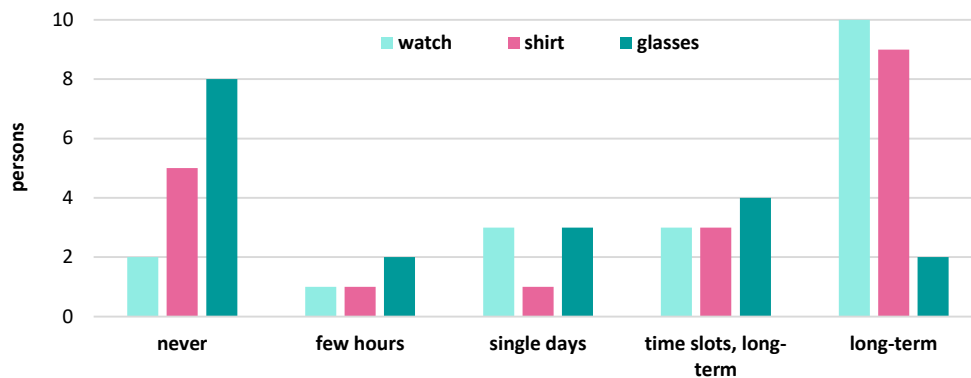


Figure 6: Acceptance votes for watches (activity trackers), smart biosignal shirts, and eye tracking glasses.

Results

Figure 6 presents the results of the investigation that demonstrate that there is a high acceptance towards long-term use of smartwatches (or activity trackers) and biosignal shirts. The rejection for eye tracking glasses was withdrawn by several participants when we assured that the embedded miniature cameras for outward environment video capture would be removed. Although the concrete number of participants that expressed this opinion was not documented except for qualitatively reporting by oral communication this demonstrates that there is a potential to exclusively use eye tracking glasses for egocentric measurements.

3.4.3 Study Plan for Resilience-Oriented Field Trial in Production Environment

To estimate recovery, stress, and resilience in production workers, a system leveraging wearable biosignal-based technologies will be implemented at a manufacturer. Wearable devices such as activity trackers / smartwatches or arm straps would collect physiological data, including HRV and sleep patterns, providing indicators of stress and recovery levels. These data points will be continuously monitored and transmitted to a central platform equipped with machine learning algorithms. The platform will analyse trends, flagging stress episodes, insufficient recovery periods, or signs of resilience. Insights from the analysis will provide means for future presentations via a user-friendly dashboard for workers and supervisors, offering actionable recommendations like rest breaks or targeted interventions. Data privacy and compliance with workplace regulations would be ensured, fostering trust and sustainable usage. This implementation aims to enhance workers' well-being and optimise productivity.

Implementation Details

The study plan envisions to equip 20 workers with wearables and to continuously monitor their data about 30 days. There will be pre-and post-study psychological tests, questionnaires and interviews. The results will be forwarded exclusively to the workers. High-level statistical information will be forwarded to the management that could draw conclusions from the anonymized distribution of resilience scoring on the working population about whether there might be a risk for increased sickness outages.

3.4.4 Service: Extended Resilience Score Computing: Integration of Recovery-Stress State and Bounce-Back

The computational method to estimate the recovery-stress state will integrate biosignal data, such as HRV, cortisol levels (via wearable sensors or proxies), and sleep quality metrics. The method will apply signal processing techniques to extract relevant features, such as stress peaks or recovery phases, and use machine learning models (e.g., support vector machines or neural networks) to classify the recovery-stress state dynamically. Time-series analysis will identify trends and deviations indicative of chronic stress or insufficient recovery. Resilience will be inferred computationally by assessing the ability of the system to return to baseline after stress, quantifying the frequency and depth of recovery periods relative to stress episodes. The output will provide actionable metrics like resilience scores or risk indicators for burnout, enabling targeted interventions to enhance well-being and performance.

Implementation Details

Based on the implementation of Paletta et al. (2024)³⁵ on the computation of resilience scores for decision support using wearable biosignal data with requirements on fair and transparent AI, we will extend a component that measures the recovery-stress state of the individual worker.

Measuring the recovery-stress state involves assessing physiological and psychological parameters that reflect a person's stress level and their ability to recover. The approach will include, as follows,

Physiological Measurements:

- **Heart Rate Variability:** HRV is a primary indicator of autonomic nervous system balance. High HRV typically reflects good recovery, while low HRV indicates stress or fatigue.
- **Cortisol Levels:** Cortisol, a stress hormone, can be measured through saliva, blood, or wearable proxies. Elevated cortisol levels indicate stress, whereas normalizing levels suggest recovery.
- **Sleep Patterns:** Sleep duration, quality, and stages (e.g., REM, deep sleep) are critical for recovery. These can be monitored using wearable devices (activity trackers).
- **Respiration Rate:** Controlled or reduced respiration rates are often associated with recovery states.

Psychological Assessments:

- **Questionnaires and Surveys:** Tools like the Recovery-Stress Questionnaire (RESTQ; Kellmann & Kallus, 2024)³⁶ or Perceived Stress Scale (PSS; Cohen et al., 1983)³⁷ capture subjective stress and recovery experiences.
- **Mood Tracking:** Daily mood logs will provide insight into emotional recovery.

Behavioural and Environmental Contexts:

- **Activity Tracking:** Monitoring physical activity levels and sedentary behaviour can provide context for stress-recovery balance.
- **Lifestyle Indicators:** Nutrition, hydration, and social interactions are qualitative data points relevant to recovery.

Integration and Analysis:

These data points are combined using computational models to classify stress and recovery states. Time-series analysis, machine learning, or statistical tools will track trends, identify imbalances, and provide personalised feedback for intervention.

3.4.5 Relevance of Fairness And Transparency in Digital Human Factors Analytics

Explainable AI and fairness of AI services in the context of socio-technical environments are key to enable future, ethically approved applications of AI for the optimization of production services with human-machine interaction.

Fair algorithms will prevent decisions to reflect discriminatory behaviour. The aim is to gain a better understanding of collective decision-making processes to tackle new socio-technological challenges where aspects of decision-making and fairness are important. We need to ensure that people in similar situations are treated equally and not discriminated against. Examples of unfair decisions are situations where people are discriminated against based on protected characteristics, such as, race, gender or age. here are various techniques for implementing fairness, including methods based on distributed consensus of resources and fair distribution of similar resources. Surash and Guttag (2019)³⁸ mention sources of bias in machine learning – measurement bias, representation bias, longitudinal data fallacy, statistical parity, etc. – with their descriptions in order to motivate future solutions to each of the sources of bias, for fair resource allocation (Jiang & Lu, 2019³⁹).

Fairness has to be investigated on any machine learning services that underlie the resilience scoring in the FAIRWork project. Concretely, fairness measures have to be operated on the probability distributions on gender-related aspects (sex, age, race, etc.) that provide data for decision-making towards resilience risk stratification.

Transparent solutions enable users to introspect processes to understand how software arrives at a solution to a problem. Transparency is provided by methodologies of eXplainable AI (XAI; Longo et al., 2024)⁴⁰ that are counteracting a tendency of "black box" in AI, where even the designers of the AI system cannot explain why it arrived at a specific decision. The focus is usually on the reasoning behind the decisions or predictions made by the AI which are made more understandable (Vilone & Longo, 2021)⁴¹. XAI should enable users to introspect dynamic systems as well as control options using XAI tools, such as, LIME, SHAP and WIT. These tools enable to explain and to interpret the predictions of machine learning models and can be further used to track the specific influence of vulnerable parameters - such as, gender, age and country of origin - on the generated results.

Finally, the context information about the worker's resilience makes it possible to dynamically generate suitable, transparent and fair recommendations for work allocations.

3.4.6 Outlook

In the future, we will conduct a wearable biosignal study. Such a study will gather real-world data on workers' physiological responses to their tasks and environment. Following this, we will refine the cognitive-emotional strain model, leveraging the controlled environment and detailed data from a study conducted at the Human Factors Lab.

We will collaborate with partners with to outline the persona-based model for digital twin applications. This model will allow us to create virtual representations of workers, enabling simulations and analyses of different scenarios. All of these activities, including the studies, model development, and optimization efforts, will culminate in scientific publications.

3.5 Optimization in Decision Support Systems

3.5.1 Overview

Making decisions in manufacturing involves many challenges, from allocating resources, to optimising a process in the manufacturing industry, to optimising processes and supply chains. Optimisation plays a crucial role in many areas. Optimisation, a fundamental principle in applied computing and mathematics, deals with the systematic search for the best or possible solution to a problem under given conditions. Examples of practical applications can be a process optimisation (e.g., document analysis) or resource optimisation (e.g., how best to load containers onto flatbed trucks). Historically, operations research (OR) has primarily concentrated on economic factors. However, since the beginning of the 21st century, human considerations have become increasingly important. Therefore, we consider to integrate human factors in our optimisation strategies (Prunet et al., 2024)⁴².

There are various AI based techniques that provide support for making informed decisions and increasing efficiency. First, when making decisions, it is important to consider the scenario. Second, complexity theory plays a fundamental role in optimization. Since most problems in combinatorial optimization are NP-hard, heuristics are typically required for their solution. Significant progress has been made in the last four decades in developing metaheuristics based on local search and various hybridisation schemes (Fraga, 2015)⁴³. Third, several modelling paradigms from a high-level perspective, examining the interrelationships between multiple elements. Decision analysis provides a valuable framework for structuring and solving complex problems involving both soft and hard criteria, behavioural operations research and dynamic elements of a process. In recent times, ethical and fairness issues have become increasingly important in decision-making.

From the methodology approach, there exist many ways to support decision making with optimisation techniques:

- AI offers innovative approaches to solving complex optimisation problems where conventional methods reach their limits. In particular, for high-dimensional, nonlinear or stochastic problems, AI methods enable efficient exploration of the solution space and approximation of optimal solutions. Combining AI methods with domain-specific knowledge and classical optimisation techniques promises more efficient and robust solutions for a wide range of applications.
- Mathematical programming is a central methodology in operations research. The simplex method, first published by Dantzig (1951)⁴⁴, is considered the most significant development in this area. Other areas of focus include optimization, combinatorial optimization, and stochastic programming. The most commonly used techniques for solving mathematical programs are branch-and-bound, branch-and-cut, branch-and-price (column generation), convex optimisation, and dynamic programming.
- Heuristics, based on Laguna et al. (2013)⁴⁵, are an important tool in production for solving complex problems and optimising decisions. However, it is important to be aware of the advantages and disadvantages of heuristics and to use them wisely. In combination with other methods, such as mathematical optimisation, heuristics can contribute to a significant improvement in production processes. Possible applications are, e.g., sequence planning. This involves determining the sequence in which orders are processed in order to minimise throughput time and maximise efficiency. Heuristics in resource allocation support the optimal distribution of resources (e.g. machines, personnel) to different tasks.
- A very relevant technique for FAIRWork is constraint programming (CP). CP offers a unique approach to optimising production processes by focusing on finding feasible solutions that satisfy a set of defined constraints. Unlike traditional optimisation techniques that seek a single optimal solution, CP can identify multiple solutions that satisfy constraints imposed by factors such as machine capacity, material availability and delivery dates. This flexibility is particularly valuable in complex manufacturing environments where finding

a single "best" solution may not be feasible or desirable. CP allows the exploration of different options that meet all the necessary requirements, enabling production managers to make informed decisions based on additional criteria, such as minimising lead times or prioritising specific customer orders.

As already defined, optimisation in manufacturing is an essential aspect of applied mathematics and computer science that aims to systematically search for optimal solutions under given constraints. In manufacturing, optimisation manifests itself in a variety of problems, ranging from the efficient allocation of resources to the process optimisation of specific operations. It is always important to define a clear objective. An appropriate method is then selected and implemented. The text analyses the challenges of decision making in complex manufacturing environments and highlights the role of optimisation as a solution approach. Different aspects are discussed and used. These include artificial intelligence techniques (e.g. for the FLEX application) or heuristics motivated by the Stellantis application. The following services provide information about the importance and challenges of optimisation in manufacturing and discuss different approaches.

3.5.2 Optimise the Check of Calibration Documents

One example of the use of AI in a business is the automation of the quality control of digital calibration documents (Nummiluikki et al., 2023)⁴⁶. It offers a significant improvement over manual review. Using various AI-based methods such as optical character recognition (OCR) and natural language processing (NLP), key data points such as instrument details, calibration equipment information, test results and signature verification can be automatically extracted from various document templates.

Quality control of calibration documents is essential to ensure the validity and traceability of measurement results and to fulfil the requirements of quality standards such as ISO 9001. Effective quality control includes checking the documents for correctness of date, completeness, accuracy, consistency and availability of signatures. Among other things, the following points should be checked: the clear identification of the calibration item and the standard used, the traceability of the standard to national or international standards, the documentation of the environmental conditions during calibration, the specification of the measurement uncertainty and the clear and unalterable documentation of the results. In addition, the calibration documents should be regularly checked for up-to-dateness and validity to ensure that they comply with current requirements and standards. Implementing a documented procedure for the quality control of calibration documents helps to minimise measurement errors, improve the quality of products and processes and increase confidence in the measurement results.

Motivation of the Use Case: After a device is manufactured, it must undergo validation to ensure its accuracy and reliability. Calibration issuers are responsible for testing the device and issuing a certificate containing all the essential details about the tested device and the calibration test cases. Flex processes approximately 3,000 such certificates annually. Currently, these certificates are manually reviewed for formal errors and test results, a process that can take upwards of 10 minutes per certificate depending on the document's length. This time-intensive task involves data comparisons from various sources to ensure that only certificates with accurate information are deemed valid. By enhancing the efficiency of this process, the certificate validator significantly streamlines Flex's workflow.

The format of a calibration certificate varies depending on the issuing laboratory. Presently, the majority of certificates Flex receives are issued by the laboratories MicroPrecision and TESI. Despite differences in layout formatting, the core information contained in the certificates remains consistent and typically includes:

- Details about the device undergoing testing
- Information on the calibration equipment used
- Confirmation that all calibration tests have been passed
- Signatures of authorized personnel
- Verification of certificate completeness

Key details about both the tested and calibration devices are accessible through the database.

Details about the device undergoing testing encompass the device name, manufacturer, model, asset number, the date of measurement, and the certificate's expiration date. Information specific to the certificate includes the issue date and the total number of pages.

Implementation Details: The calibration documents are text documents. The analysis of PDF documents using AI methods and an implementation of rule-based lists to check the text or calibration documents. Human review reaches its limits here, especially when processing large volumes of documents or complex layouts. The developed libraries can extract text passages, identify keywords and recognise semantic relationships between words and sentences.

3.5.3 Work Schedule by using Human Factors

A worker schedule based on stress levels aims to optimize the workload of employees and prevent burnout. Various factors are taken into account, such as the complexity of the tasks, the number of deadlines and the general workload (Berti et al., 2021)⁴⁷. These factors are used to determine a stress level for each employee. The worker schedule then distributes the tasks in such a way that the stress levels are as balanced as possible, and no employee is overloaded. Breaks and recovery times are also integrated into the schedule in order to relieve employees and maintain their performance. This system promotes the health and well-being of employees and can lead to higher productivity and employee satisfaction (Tropschuh et al., 2024)⁴⁸.

Furthermore, a stress level and resilience core could also be incorporated, reflecting an individual's capacity to handle stress. This score might be based on factors like experience, personality traits, or even training in stress management techniques. The worker schedule then distributes the tasks in such a way that the stress levels are as balanced as possible, and no employee is overloaded, taking into account both their current stress level and their resilience. For example, an employee with a lower resilience score might be assigned slightly less demanding tasks, even if their current stress level is similar to a colleague with a higher resilience. This approach not only promotes employee well-being but also enhances productivity and job satisfaction. By considering both stress levels and resilience, the schedule aims to create a sustainable workload for each employee. Breaks and recovery times are integrated into the schedule to further alleviate stress and maintain performance. This system fosters a healthier work environment, leading to increased employee engagement and reduced burnout rates. By dynamically adjusting workload based on both stress and resilience, the system can better cater to individual needs and contribute to a more robust and adaptable workforce.

Implementation details: The implementation was done through a simple exemplary implementation of a work plan. An assignment algorithm was developed that takes into account the stress level of the employees and the current workload. Subsequently, frequent changes in task allocation are made, such as swapping, shifting or splitting tasks.

The change that improves the stress level the most is adopted. This process is repeated until an acceptable balance is achieved for all resources. In addition, where possible, additional breaks are built into the schedule to relieve employees. Longer recovery periods are scheduled for high stress levels. The innovation lies in the fact that human resources can change more frequently over time, resulting in a more dynamic and varied distribution of resources that takes into account people's stress levels.

3.5.4 Transport Optimisation

Industrial companies are under increasing pressure to improve their performance and measure themselves against key performance indicators. This development poses major challenges for many companies (Adenipekun et al., 2022)⁴⁹. More and more companies are faced with the almost impossible task of guaranteeing high quality standards while at the same time minimising resources. One of the most important logistical processes, the transport of goods and materials, is particularly affected by rising costs. A large number of departments and functional areas within an industrial company, such as production, warehousing and dispatch, are linked to the transport process. The efficiency of transport, therefore, has a direct influence on the productivity of the entire company. Despite the trend towards shorter throughput times, goods and materials must be transported to the right place at the right time ever more quickly and precisely. In an industrial environment, various players are involved in the transport process: Production employees, warehouse staff, drivers and logisticians must work together efficiently to ensure a smooth process. Companies are therefore intensively looking for optimisation options that enable cost savings without compromising the quality of transport and on-time delivery. The optimisation of transport, including external logistics services, is therefore a decisive factor in reducing operating costs and increasing the competitiveness of industrial companies.

Implementation Details: The research implementation of an algorithm for dynamic route planning that is characterised by universal applicability and extensive configuration options. In contrast to previous approaches, it enables the adaptation of loading patterns and time windows during runtime. The focus is on the optimisation of the insertion heuristics, considering different vehicle occupancies and optional time windows.

3.5.5 Outlook

We plan to continue our research in the following different areas for the manufacturing industry:

Future research shows great potential for integrating human factors into decision models for production and logistics problems. Of particular interest are scalable modelling approaches that incorporate human factors. Furthermore, we would like to extend our research by applying human factors to the routing literature, e.g. by considering fatigue. The work of Fu et. al. (2022)⁵⁰ is one of the few papers that considers human factors in the planning of driver breaks.

For the quality assurance task (e.g. checking the calibration document) in production, it will probably be interesting in the future to combine the chosen rule-based approach with the possibilities of generative AI as offered by LLMs. This approach could offer advantages in terms of more flexible adaptation to different formats and types of calibration documents. In particular, generative AI could be used for making better decisions. By combining rule-based systems with the flexibility and adaptability of generative AI, companies could significantly improve quality assurance in production and achieve time and cost savings. It should be noted, however, that the use of LLM in this area will require careful validation and monitoring to ensure that this approach works.

3.6 AI-Enriched Decision Support Systems

3.6.1 Overview

In this section, we explore the potential of AI methodologies for optimising decision-making and scheduling processes in manufacturing contexts. Our research is structured around four research questions addressing the integration of AI into industrial practices as described in (Figure 7). First, we examine existing AI applications (I). Next, we evaluate their potential to enhance traditional decision-support systems in complex and dynamic environments (II). Then, we investigate how optimization metrics and constraints impact industrial scheduling (III). Additionally, we assess the extent to which AI techniques can be utilised for optimisation in production contexts by developing small-scale demonstrations (IV).

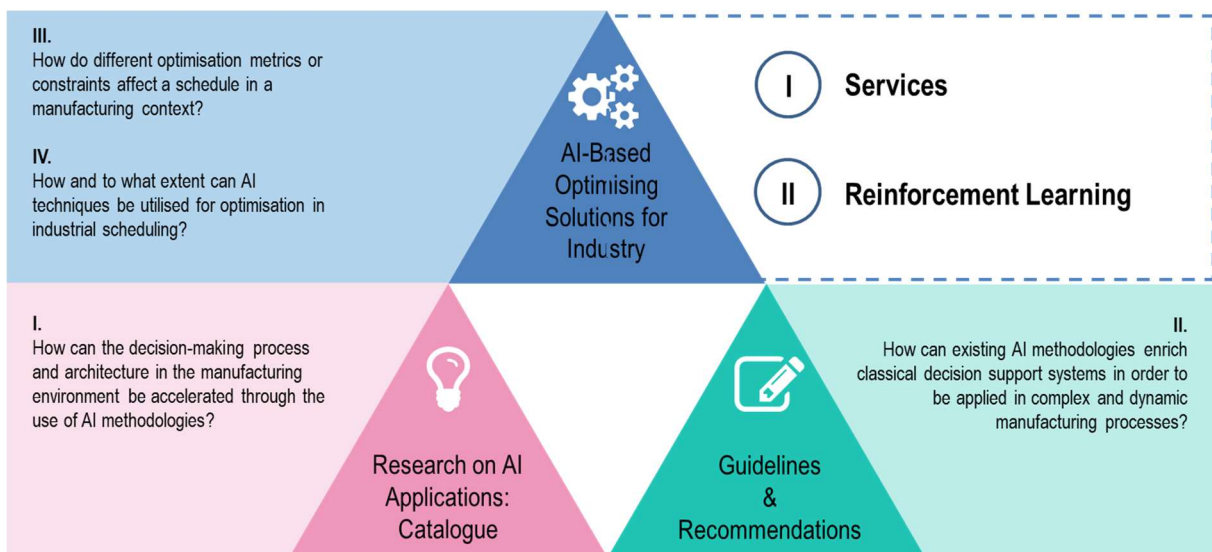


Figure 7: The overview of the research track AI-enriched DSS.

During the given time frame for this periodic report, our efforts in WP3 were dedicated to investigating the research on machine learning methodologies applied in resource and production planning (research question I) as well as further research in the domain of reinforcement learning (RL) applied to optimise industrial scheduling (research question III). The research questions and a short summary of the outcomes are presented in below.

Research Question 1: How can the decision-making process and architecture in the manufacturing environment be accelerated through the use of AI-based methodologies?

To address the first question, we conducted a scoping review. Over 200 scientific articles published within the past 14 years were identified. Each article's abstract was reviewed by two researchers, resulting in a final selection of 70 articles that were reviewed in full scope, as available free of charge. The analysis of 48 accepted papers revealed 61 methods that we classified into RL, supervised learning, and unsupervised learning. An additional classification layer was applied based on the use cases in which these methods have been applied, such as production planning, resource planning, process optimisation and control, and maintenance.

The possible outputs of the presented ML methods were broad. They used regression models to include predictions of power consumption, process duration, demand forecasting, remaining useful life, tool wear rates,

operating conditions, and material distortion forces. On the other hand, RL focused on more dynamic tasks, such as scheduling, resource allocation, maintenance planning, and order acceptance. By utilising classification techniques, studies investigated an improvement of key performance indicators, environmental limits, and scheduling options. Also, forecasting demand, cycle times, and predictions of machine failure were examined. Advanced models, including hybrid approaches, e.g., adaptive neuro-fuzzy inference system, deep RL, and multi-agent RL, were used for specialised tasks like power plant performance, policy optimisation, and scheduling recommendations.

Question 2: How can existing AI methodologies enrich classical decision support systems in order to be applied in complex and dynamic manufacturing processes?

In the scope of the second research question, a research gap in DSS classification was identified, where the present literature underlined the need for adequate resources to guide developers in selecting an appropriate method. As one of our goals in the project is to create guidelines for developers, we proposed a structured categorisation of DSSs into four distinct classes: rule-based, optimisation-based, simulation-based and learning-based. This classification serves as a tool for developers and end-users to find common ground in understanding real-world use cases and possible technical solutions. A detailed description of this study can be found in Olbrych et al. (2024)⁵¹ and in Deliverable 3.2.

Question 3: How do different optimisation metrics or constraints affect a schedule in a manufacturing context?

Resource allocation is a fundamental problem in manufacturing and production. It involves assigning tasks to resources such as machines in an efficient and timely manner. This problem is crucial for optimizing productivity and minimizing costs, but it presents significant challenges due to its complexity and the interdependencies between tasks. Recent advancements in RL have shown promise in tackling such optimization problems, particularly when traditional heuristic methods struggle to scale. However, RL applications in this domain are often hindered by sparse reward signals, which provide limited feedback to guide the agent's learning process. One key metric in this context is the makespan - the total time required to complete all tasks, from the start of the first task to the completion of the last -. While optimizing the makespan is central to improving scheduling efficiency, its sparseness as a reward signal complicates RL training. To address these challenges, two studies were conducted to investigate the impact of reward signal design and exploration techniques on the effectiveness of RL for job shop scheduling.

The first study, "*Reward Shaping for Job Shop Scheduling*," found that dense reward signal yields better results with the same training resources. Specifically, constructing a dense reward signal based on the metric of machine utilization proved more effective than directly optimizing a sparse reward signal tied to the makespan. The makespan alone results in a sparse reward signal because it is only revealed once a production plan is complete, which corresponds to the end of an episode in an RL setting. The reward signal in reinforcement learning represents an optimisation metric or a combination of optimisation metrics. This means that a higher reward indicates better performance of the corresponding optimisation metrics. For more information, please refer to Deliverable 3.2 and Nasuta et al. (2024)⁵².

The second study builds upon the environment used in the first and investigates curiosity-driven exploration techniques for RL. Curiosity-driven exploration has emerged as a promising approach to systematically guide RL agents in exploring state spaces, especially in environments with sparse rewards. This technique transforms

a sparse reward signal based on the makespan metric into a dense reward signal. The study demonstrated that curiosity-based approaches achieved comparable, and sometimes even slightly better, results within the available timestep budget compared to formulations based on machine utilization. Moreover, RL agents incorporating curiosity mechanisms can potentially escape local optima and discover better solutions given sufficient computational resources.

Question 4: How and to what extent can AI techniques be utilised for optimisation in industrial scheduling?

Within this research question, we explored the literature and, based on the possible use cases in the project, developed small-scale demonstrators that showcase the practical application of AI techniques. Its scope extends beyond industrial scheduling, which is primarily addressed in Questions 1 and 3. The demonstrators discussed here are publicly accessible through GitHub platform, providing valuable insights into diverse optimisation strategies and hands-on learning opportunities.

*AHP-Fuzzy Logic for Safety and Efficiency*⁵³

This demonstrator evaluates workspace safety and efficiency by calculating customised ratings. Inputs include robot mode, cycle time, layout, robot movement, gripper choice, space requirements, human traffic, worker automation skills, safety requirements, and quantity. These parameters allow for tailored assessments of workspace performance.

*Genetic Algorithm for Allocation*⁵⁴

This demonstrator uses a genetic algorithm to optimise resource allocation, specifically worker-order assignments. It considers worker preferences and central allocation strategies to maximise a score, representing various objectives such as efficiency, cost-effectiveness, or overall performance.

*Reinforcement Learning for Inventory Management*⁵⁵

Within this example, RL for inventory optimisation was applied. It adapts inventory levels to seasonal demand fluctuations using the Stable Baselines3 library. The project also integrates Weights & Biases (Biewald, 2020)⁵⁶ for hyperparameter tuning and experiment logging, ensuring effective model performance.

3.6.2 Research on Existing AI Applications: ML Catalogue

3.6.2.1 Motivation and Reference to FAIRWork Use Case

Going one step further from the categorisation of DSSs, we created an extensive catalogue of data-driven methods applied to production and resource planning in the industry. Conducting a literature review on ML methods for production planning and resource planning in the industry is highly valuable, particularly for use cases like worker allocation, warehouse management, and production scheduling. The overview of the technical landscape not only establishes knowledge but also identifies research gaps and promotes the adoption of ML in the industry. This study aims to connect academic research with practice by presenting the newest technological achievements and building trust in industries to implement promising ML solutions. The review aims to discuss key topics, provide recommendations, and highlight the evolving landscape of AI-enhanced DSS in manufacturing.

3.6.2.2 *Innovation beyond the State-of-the-art*

In the context of Industry 4.0, deploying advanced data-driven technologies often faces the burden of high complexity, the number of algorithms that could be utilised, and a lack of trust. Studies like Berlimini et al. (2021)⁵⁷ and Ivanov et al. (2021)⁵⁸ presented literature reviews that provide the orientation for developers in industrial operational management and research perspectives on Industry 4.0. Our study, on the other hand, provides an overview of the latest algorithms applied not only in production planning but also in resource management, covering topics like cost estimation, energy consumption forecasting, predictive maintenance and dynamic demand planning.

3.6.2.3 *Description of Functionality*

The result of this study is a comprehensive catalogue of validated ML methods that can be used as a foundation for developing AI services. It presents various learning algorithms that can be further improved and adapted for specific industrial scenarios. Additionally, it highlights future directions and identifies existing gaps in the research field.

3.6.2.4 *Results*

Our review focused on publications from 2010 to mid-2024, selecting conference papers and journal articles while excluding reviews. We targeted studies on decision support, resource planning, and machine learning in manufacturing. The search was conducted using three platforms: Scopus, Web of Science, and IEEE Xplore. After removing duplicates and screening abstracts, 78 papers were accepted for the full-text screening. Due to access limitations, 47 full texts were analysed, summarising ML methodologies.

In the reviewed research positions, supervised learning is the dominant approach in industrial applications, with models like Random Forest, Linear Regression, Decision Trees, and Multi-Layer Perceptrons used for cost estimation, energy forecasting, demand prediction, and production optimization. Neural networks, particularly Multi-Layer Perceptron and Long-Short Term Memory models, excel in time-series forecasting and predictive maintenance. Another widely applied method was RL, which aids real-time decision-making with algorithms like deep q-network, proximal policy optimization, and actor-critic optimizing scheduling, resource allocation, and maintenance. Unsupervised learning, though less common, adds value to demand segmentation and risk assessment. Methods like k-means and hierarchical clustering help uncover hidden patterns, highlighting the potential for enhanced anomaly detection and early warning systems.

From 2019 to 2024, ML methods evolved from static regression-based predictions for cost and quality estimation to real-time adaptability and intelligent decision-making, driven by the rise of RL. This shift reflects the growing demand for dynamic systems that respond to industrial changes. Additionally, hybrid models like adaptive neuro-fuzzy inference systems and boosted decision trees enhance predictive accuracy by combining multiple techniques, proving valuable in complex scenarios.

In conclusion, the research in this area reflects a robust and growing reliance on machine learning to address the multifaceted challenges of modern industrial systems. The progression from traditional regression-based models to advanced RL methods underscores the increasing complexity and dynamism of production environments.

3.6.3 **Guidelines and Recommendations for AI Developers**

FAIRWork project aims for transparent and explainable use of AI methodologies that often function as "black box" systems, making it challenging for developers to explain processes to end users and slowing down their

implementation in real-world scenarios. Therefore, the focus of this study was to develop a novel DSS classification that integrates current technologies and allows for accessible explanations and discussions on selected methods for a specific industrial use case. The main outcome of this study is a categorization of DSS into four distinct classes: rule-based, optimization-based, simulation-based, and learning-based. It provides guidelines for selecting methodologies based on use-case requirements. The classification aligns with user needs and industry evolution, ensuring adaptability and effectiveness while addressing challenges in explaining methodologies to end users. More information about the research can be found in Olbrych et al. (2024)⁵⁹ and in Deliverable 3.2.

3.6.4 AI-Based Optimizing Solutions for Industry

In the first study, we explored how RL can optimise scheduling processes, focusing on the Job Shop Scheduling Problem (JSP). By investigating different reward functions' impact on solution quality and evaluating case studies, we seek to provide valuable insights for enhancing manufacturing productivity and competitiveness. (Nasuta et al., 2024)⁶⁰.

A second study investigates curiosity-driven exploration techniques to enhance RL performance in environments with sparse rewards, using the JSP as a case study.

3.6.4.1 Motivation and Reference to FAIRWork Use Case

Curiosity-driven exploration was initially developed to address challenges in RL for video games like *Montezuma's Revenge*, leading to significant breakthroughs in handling sparse reward settings. Optimization problems, such as resource allocation in industrial use cases, often lead to sparse reward settings when formulated in an RL framework. This study explores whether curiosity-driven exploration can offer similar benefits in optimization contexts as it did in video games.

3.6.4.2 Innovation beyond the State-of-the-art

While curiosity approaches have been extensively studied in video games and RL for optimization problems like the JSP, to the best of our knowledge, applying curiosity-driven exploration to JSP has not been explored before. A key distinction between video game environments and optimization problems lies in the observation space: video game states are typically represented as images (screenshots). In contrast, optimization problems often use directed graphs. As a result, video game approaches often employ convolutional neural networks, while optimization problems are more suited to graph neural networks or dense models.

3.6.4.3 Description of Functionality

Curiosity approaches add an additional reward signal to the environment's original reward function. This signal incentivizes the RL agent to explore new states and system mechanics by rewarding novel states and the agent's ability to predict the effects of its actions. By introducing a curiosity signal, the agent first learns how its actions affect the environment. Once it has an accurate model of the environment dynamics, it shifts focus to solving the original task. In the video game domain, this approach allowed agents to first learn complex moves, such as a double jump, which they could then use to play the game more effectively.

3.6.4.4 Interfaces

The two most promising curiosity approaches were implemented as Gym environment wrappers, making them easily adaptable to RL environments that follow the Gym standard. These wrappers also integrate experiment

tracking with the Weights and Biases platform, allowing for the logging of intrinsic and extrinsic rewards, as well as the encountered states.

3.6.4.5 Experiments

The study evaluates the impact of curiosity modules in an RL setup for the JSP using benchmark instances from Fisher and Thompson. The JSP environment is configured with a sparse reward function, providing feedback only upon the completion of a full schedule, creating a challenging exploration scenario. Hyperparameter tuning for the proximal policy optimization algorithm was conducted in two stages for both instances, involving random and grid search. The best-performing configurations were evaluated with varying numbers of timesteps.

3.6.4.6 Results

By transforming the sparse makespan-based reward signal into a dense one, the study achieved results comparable to, and occasionally better than, those based on machine utilization. Additionally, the curiosity-driven approach enabled RL agents to escape local optima and discover potentially superior solutions when given sufficient computational resources.

3.6.4.7 Integration into the DAI-DSS architecture

Curiosity Gym wrappers can be utilized to train RL agents deployed as AI services. Since curiosity approaches emphasize exploration, they are most beneficial during the training phase of an RL agent. In production, the RL agent should focus on exploitation, and the curiosity Gym wrapper should not be used. However, training an RL agent with the curiosity Gym wrapper is advantageous for improving exploration capabilities during the learning process.

3.6.5 Outlook

Future research should investigate the scalability of curiosity-based RL in more complex scheduling scenarios, including larger problem instances with increased jobs, machines, and constraints. Additionally, the adaptability of these approaches should be tested in dynamic environments where job arrivals are unpredictable, and processing times vary due to machine breakdowns or other uncertainties. Further optimisation of hyperparameters across diverse problem settings could enhance the robustness and efficiency of RL models, ensuring their applicability in real-world industrial applications.

Additional research should focus on implementing resource and production planning methodologies in real-world industrial scenarios, particularly in the context of Industry 5.0, which emphasises human-centric, sustainable, and resilient systems. Application-oriented research is crucial for developing practical benchmarks for ML methodologies tailored to modern industrial environments. These benchmarks could guide the integration of advanced technologies.

3.7 Model-based Knowledge Engineering for Decision Support

3.7.1 Overview

Modelling is used to externalize the knowledge of decision-makers and experts familiar with the use case scenario and problem setting and to detail the steps towards configuring AI algorithms. Due to the advantage that models

are interpretable by humans and machines, modelling is applied to assist starting from the definition of requirements to the implementation of suitable AI solutions. E.g. Domain-specific models can serve as the basis for configuring the DAI-DSS aiming to enhance the transparency and explainability of the involved AI applications or technical models could be used to describe orchestration between AI services in the DAI-DSS.

To support the configuration of the use-case-specific AI applications, a systematic three-layered approach based on models ranging from problem identification to the executable AI solution for decision support was proposed in Deliverable 3.2. As a recap, the framework and description of the layers are illustrated in Figure 8. The layers are described as:

- 1) **Identification:** This layer captures the identification of the current status and the concrete problem settings of business processes, but also corresponding success factors, existing IT systems and architecture components, compliance requirements and relevant KPIs through harvesting, and modelling domain knowledge. This can be supported with approaches such as co-creative or hybrid workshops, or interviews.
- 2) **Specification:** In the second layer the initially defined properties are described in more detail and mapped with AI applications. The abstract decision logic, methods for knowledge-based mechanisms e.g. input and output data types and underlying expectations for mechanisms as well as desired solution outputs are defined. For example, training and test data that enable the training of AI, Excel and decision models describing the reasoning, or methods for certification and approval of AI results are specified.
- 3) **Configuration:** In configuration, concrete AI applications and used techniques, frameworks or computational models for execution or calculation (e.g. concrete rules that are executed by rule engines, fuzzy logic that can be interpreted or semantics that can be inferred) are configured.

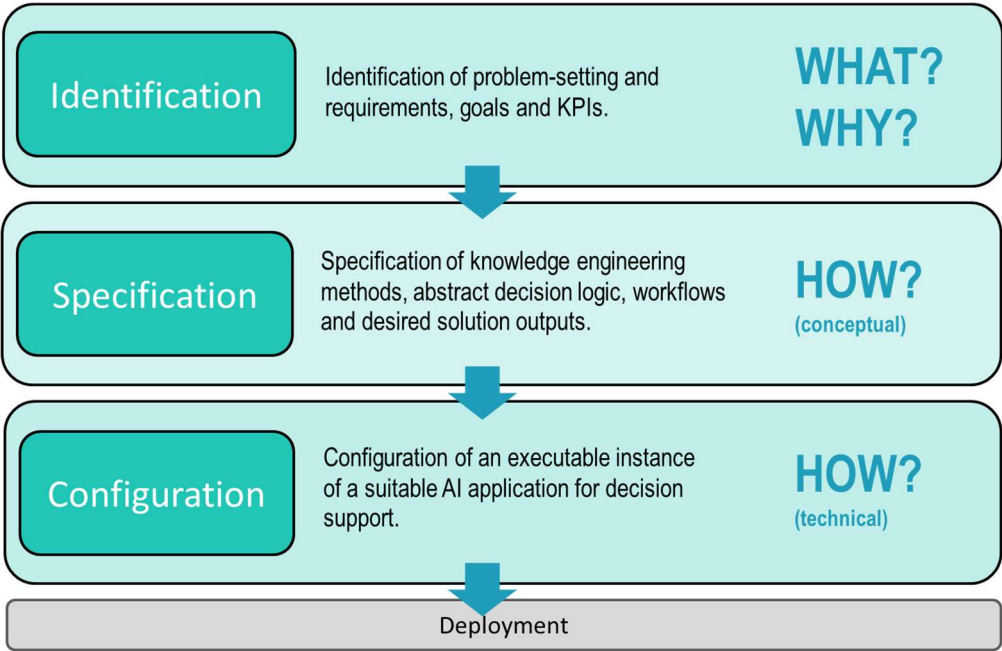


Figure 8: Overview of the three-layered approach.

Within this research track the aim is to use the three-layered methodology and the prototypes and materials gathered and elaborated during the FAIRWork project and point out possibilities to introduce AI into businesses in a structured way based on models.

Current trends in AI suggest that, on the one hand, a shift from single AI methods to compound AI systems to fit increasingly complex tasks is needed (Jaffri, 2024)⁶¹. Especially, the progress in LLM, generative AI (GenAI) and the advantage of compound AI systems in their modularity and the possibility of combining and orchestrating multiple components and different AI approaches (e.g., hybrid AI) to perform more complex tasks stresses the future need to develop whole intelligent systems and applications instead of single AI services. On the other hand, compliance requirements in terms of legal, ethical and technical aspects for organizational processes using AI are getting more and more relevant (AI HLEG, 2019⁶²; AI HLEG, 2020)⁶³. As more complex AI solutions emerge, new legislation e.g. the AI Act, forces companies to comply in legal, ethical or technical dimensions when offering or applying AI-based applications. Therefore, besides deciding on a specific AI solution to enhance organisational efficiency, companies are required to accompany the AI introduction with organizational processes for AI usage and governance.

Thus, the three-layered approach is adapted to a broader focus, highlighting two aspects:

- 1) In Deliverable 3.2, the three layers are investigated for specific AI approaches and use cases (e.g. rules, fuzzy rules, agents, artificial neural networks etc.). Chosen AI approaches are analysed to identify underlying rules, conditions and requirements for the three layers. In Deliverable 3.3, the three layers are summarized in a company view in contrast to a very specific AI-service view. This might cover whole AI applications and solutions instead of single (AI) algorithms including multiple frameworks, algorithms or concepts such as RAG combined with LLM or Graph RAG. Potentially, also using and evaluating existing AI applications and platforms on the market.
- 2) For a successful introduction of new AI applications into companies, AI usage and governance aspects must be considered too. AI usage includes the definition of processes, roles, skills, responsibilities and how AI is used in the business processes, the legacy infrastructure, and data and considering applicable regulatory requirements like GDPR, AI-Act or Data Resilience Act. In addition, AI governance must be established for AI introduction, training for the employees must be provided and measures to ensure compliance and the expected impact and benefit of an AI application need to be defined. Sample questions are highlighted for the reflection of existing AI solutions at the market in Section 3.7.2.3. A model-based method on how to bring different (compliance) principles into AI and support AI trustworthiness through modelling will be presented in Section 4.

The overall goal of BOC's research in this research track is to define how modelling can support companies navigating through increasingly complex business environments with a focus on introducing and applying AI. Therefore, the following two questions are asked:

- How can models support getting AI into companies?
- What models are needed to identify, specify or configure AI?

It must be noted that the two questions capture and summarize the context of the research questions proposed in Deliverable 3.2. To answer, the methodology to support the introduction of AI into the FAIRWork use case partner organisations CRF and FLEX is described. Additionally, “other” use cases are reflected. The methodology is based on the three-layered approach presented above and proposes a mapping of all elaborated items during the project such as documents, context information, processes and prototypes. Modelling methods describing the item’s relevant information are suggested to identify, specify and configure AI.

For example, models can be used to “identify” the need for AI in terms of finding existing AI solutions on the market to support business processes, such as an AI-based spend management platform for a company’s invoice management or a GenAI-powered marketing platform to increase lead rates. Second, models can be used to “specify” the needs and expectations of AI. For example, models can be used to represent organisational process landscapes, business processes or decision processes and to derive use-case-specific requirements or solution outputs. Also, to capture expectations towards AI and its requirements when being developed can be illustrated with concept models. Third, models can be applied to “configure and influence” AI. These may include describing relevant features for the linear sum assignment solver or constraint programming or using technical models to describe the configuration of orchestration of AI such as the rule-based Netflix Conductor orchestrator or the orchestration of AI using LLM and RAG.

Within this research track, OMILAB researches how information from conceptual models can be used to support the design, deployment and support explanation of proposed solutions from the decisions support system. The focus till Deliverable 3.2 (Paletta, 2023)⁶⁴ was set on the model-based design approach and how the modelled knowledge can be reused in the configuration of the DAI-DSS. The focus in this deliverable is set on how models can be used to explain decisions on higher abstraction levels, supported through the models created during the design and instantiation of decision services within the DAI-DSS.

These design models are used to support explaining proposed and made decisions by the DAI-DSS. In this way the created models themselves will be further utilised, improving the model value on one hand and support the explainability within the DAI-DSS through diagrammatic models, which are easier to understand by humans (Larkin, 1987⁶⁵; Mayr & Thalheim, 2020)⁶⁶, on the other hand. Therefore, the models from the design time are not used as is but are enhanced. The enhancement can be based on information from made decisions or functionality added to the models and the used modelling methods in supporting the understanding. The goal is to support explanations through interpretable diagrammatic models and in doing so not only explain the solutions to experts and decisions makers, but to support the understanding of all stakeholders involved or influenced by the proposed solutions to the decisions.

Therefore, the models from the three-layered approach introduced above will be used to feed information from the made decisions back to the models and in doing so add a second information flow to the three layers. As the models describe the created decision support on different levels, the models are representing already parts of the decisions scenario and are therefore semantically adjacent to the information created by the decision support system. This proximity will be used to adapt the models to integrate information from the decision support system and capture and visualise them within the models.

3.7.2 Modelling for AI: An Approach for Model-based AI Configuration

This section presents the method of mapping: 1) items such as created prototypes, documents and materials and 2) models and modelling methods to the three layers Identification, Specification and Configuration. The

methodology illustrates how AI was introduced based on the three-layered structure into the use case partners CRF and FLEX, while analysing the underlying models. Also, examples and thoughts for other use cases are given.

All items created during the project such as business process models, decision models, deliverables capturing the problem statement, use case and decision processes as well as prototypes are mapped to the layers. Then each item for each layer is described as a model, as the underlying idea is that all items can be represented through models (e.g., BPM for the business processes, DMN for the rule-based decision service, ER-Models for describing data and its relations, etc.). This is described under the Experiment Section 3.7.2.3. There is no restriction to a certain modelling type as long as the model is suitable to encode knowledge or relationships in a simple way. Models can range from conceptual to computational but might also include informal and less structured techniques like a description of relationships in natural language. The described approach aims to contribute to a so-called model-repository as an outlook (Section 3.7.4) to enable a flexible combination of the appropriate models depending on the user's needs.

3.7.2.1 Motivation and Reference to FAIRWork

Conceptual models are interpretable by human beings and machines. Their visual aspects aid human understanding, while the conceptual and semantic representation contribute to machine interpretation. Thus, these models can assist in bridging the gap between human-oriented and machine-oriented approaches. Domain experts' problem settings as well as their decision logic for solving the problem can be represented in the form of domain models (e.g., business process models and notation (BPMN)) to configure AI solutions. Additionally, domain models can be used to provide context information to the AI or as a medium to describe AI-generated information to the user. On the technical side, an AI catalogue containing solutions that need to be configured to meet the specific use case needs. This research explores how models serve as an instrument to communicate and represent requirements, and capabilities as well as describe AI configuration and orchestration logic. The proposed method serves as a starting point to investigate the utilization of models to identify requirements coming from use case scenarios, to specify the logic for AI solutions to address the use cases, and to configure AI services and their orchestration. Additionally, due to the increasing demand for being compliant when utilizing AI solutions for organisational tasks and processes, initial thoughts supporting AI usage and governance aspects through modelling are reflected.

This research can be positioned into the recent conceptual modelling (CM) with AI domain. The combination defines the CMAI domain, which aims to improve the strengths and address the weaknesses of each separate domain (Fettke, 2020)⁶⁷. Bork et al. (2023)⁶⁸ indicate that based on their review conceptual modelling for AI Ethics as well as model-based code generation especially for recent technologies in Machine and Deep Learning or NLP is gaining in relevance. Also, Shlezinger et al. (2020)⁶⁹ highlight the importance of CM and AI to leverage the reliability, interpretability and robustness of deep learning approaches by applying hybrid approaches. Mattioli et al. (2022)⁷⁰ emphasize hybrid AI approaches to achieve trustworthy AI and propose multiple steps for engineering AI services.

3.7.2.2 Interfaces

As different types of materials and, therefore, multiple models and modelling techniques can be used for the three layers, the hybrid modelling tool Bee-Up⁷¹ is suitable. Bee-Up implements multiple modelling languages into a single prototypical modelling tool, which can be downloaded and used for free. It allows the creation of models in commonly used modelling languages such as BPMN, decision model and notation (DMN) or unified modeling language. While Bee-up already supports multiple modelling languages, it may be necessary to explore new modelling tools that address additional languages needed for this research effort, as the modelling method must

be usable and fitting to the information system with which it should be used. Bee-up was already applied in Deliverable 3.2 for some of the concrete examples like rules and fuzzy rule-based modelling.

The Bee-Up modelling tool is implemented on the open ADOxx⁷² metamodeling platform, which supports the development of various domain-specific modelling tools. The ADOxx platform was not only used in the context of Bee-Up within FAIRWork, but the Scene2Model tool, which is used model-based method, and the certification prototype are implemented on ADOxx. ADOxx itself offers a GUI for creating and interacting with the created models and functionality to process the models. Therefore, the platform’s own scripting language AdoScript, can be used to implement the needed functionality. Such functionality must not be implemented within ADOxx itself, but AdoScript offers the possibility to send and gather information from external systems through HTTP calls. Within FAIRWork this is used to combine the modelling tools with the services within the DAI-DSS environment.

3.7.2.3 Experiment

This section deals with the identification, specification and configuration steps to introduce and apply AI for CRF, FLEX, and “OTHER” use cases. The framework focuses mainly on FAIRWork but also includes under “OTHERS” further AI solutions as examples of 3rd party providers on the market.

In the first step, materials, models or prototypes are mapped to the corresponding “Identification”, “Specification” and “Configuration” layers. All items can be clustered per layer (horizontally) and company partner (vertically) and might serve as a consulting “package” and solution example for a similar use case. The clusters are in the following referred to as “pattern”. Patterns of one layer might vary for companies. E.g. for the “Identification” layer of FLEX and CRF the pattern is the same, whereas in “Specification” or “Configuration” not. This suggests that certain materials are optional for the Configuration of AI solutions. The framework and its vertical and horizontal dimensions as well as all relevant items, documents, Excels, and prototypes mapped to the layers are illustrated in Figure 9.

	CRF	FLEX	OTHERS
Identification	<ul style="list-style-type: none"> • Scene2Model • Process Landscape • Business Process • Deliverable 2.1 • ... 	<ul style="list-style-type: none"> • Scene2Model • Process Landscape • Business Process • Deliverable 2.1 • ... 	<ul style="list-style-type: none"> • ...
Specification	<ul style="list-style-type: none"> • Data Samples • Decision Process • Deliverable 5.1 • Rule Excel • ... 	<ul style="list-style-type: none"> • Data Samples • Decision Process • Deliverable 5.1 • IT Architecture • ... 	<ul style="list-style-type: none"> • ...
Configuration	<ul style="list-style-type: none"> • Decision Support through Decision Tree • Resource Allocation MAS-based • Production Planning Service with a Hybrid Approach • Resource Allocation using Linear Sum Assignment Solver • Decision Support through Decision Tree • ... 	<ul style="list-style-type: none"> • Support Machine Maintenance using RAG and LLM • Document Transformation using LLM • Support Compliance for Clean Room using RAG and LLM • Calibration Certification Service • ... 	<ul style="list-style-type: none"> • AI-based spend management • AI-based BPM creation • UI/UX prototyping • ...

Figure 9: Items to support the “introduction of AI into companies”.

To formalize and depict the knowledge of a company, multiple steps were needed. For both CRF and FLEX, the first step resulting in the “**Identification**” was to depict the problem settings composed of a modelling workshop including handwritten notes, the second step was to capture the semantics with Scene2Model and depict it digitally. Then, the process landscape was proposed for a broader view of the company processes and to embed the final relevant business processes describing the ultimate scenarios for which AI should be developed. (e.g., worker allocation to different production lines, production optimization, support machine maintenance). The materials used in this pattern for CRF and FLEX are described in detail in Deliverable 2.1.

The “**Specification**” differs slightly for FLEX and CRF. In addition to the data samples, and the derived decision processes detailing the business process models from both companies, Excels were used for CRF’s Worker Allocation depicting the rules and data points as modelled in the decision-making process for worker allocation. This Excel describes the decision process model depicted logic when taking a rule-based approach and illustrates understandably how data, human sensor data, worker preferences and “AI” mechanisms can result in a certain allocation. The decision process models and data samples for CRF and FLEX mapped to this layer are detailed in Deliverable 5.1. Additionally, for FLEX the IT infrastructure for Machine Maintenance was depicted, as different data sources are accessed.

Based on the problem understanding, the decision process and the Excel of the previous layers, several AI solutions (e.g. Support Understanding of Decisions through Conceptual Modelling (DMN), Resource Allocation using Linear Sum Assignment Solver, Production Planning Service with a Hybrid Approach (CP and RL) etc.) were developed for CRF “**Configuration**” pattern. The identified gateways of the decision process model of the previous layer were used as data points and considered for the configuration of the AI models. Also, for the configuration and orchestration in the overall DAI-DSS prototype between the AI and other components such as the knowledgebase, the rule-based Netflix conductor orchestrator and the MAS-orchestrator, the decision process model and its Excel implementation served as basis. A detailed technical description and the configuration of AI models, orchestrators and knowledgebase can be found in the documentation of the final DAI-DSS Prototype in Deliverable 4.3.

The “**Configuration**” pattern of FLEX was approached differently. For CRF, all three layers from top to bottom were strictly followed, for FLEX not all items in the “Identification” and “Specification” patterns were necessary for the configuration of AI techniques. For example, the problem and use case descriptions of Deliverable 2.1 in the “Identification” pattern as well as data examples, their format and the expected outputs solutions of the AI service of the “Specification” pattern were used for context understanding and for selecting an appropriate AI approach. However, the decision process models and the business process models had no direct impact on the configuration of the AI solutions (e.g. Support Machine Maintenance using RAG and LLM, Document Transformation using LLM, Support Compliance for Clean Room using RAG and LLM, Calibration Certification Service etc.) and their orchestration logic within the DAI-DSS. This example shows that not all items of previous layers are equally crucial to introducing AI into companies but some can be seen as add-ons.

Lastly, the category “OTHERS” aims to extend the framework with AI solutions from the market, which should find consideration due to the emergence of a plethora of AI tools for other use cases besides manufacturing. Solutions with the potential to enhance organisational processes range from AI-supported spend management platforms and AI-based business process generation to no-code UI/UX design platforms. For such existing AI solutions, the question is which models are necessary in the identification and specification layer to determine AI introduction with regard to usage and governance in companies. For example, when reflecting on the **AI-based spend management platform**⁷³ offering AI-supported invoicing, travel expenses and company card management potential questions for the identification and specification layers must capture information on:

1) current processes including problem statements or business processes describing the current situation and requirements in a company. E.g. which tasks are currently done manually for invoice management and which part of the process can be done by AI, how often is the process repeated, what are skills and roles required by employees, who is responsible for tasks and where are approvals needed.

2) used IT systems and architecture. E.g. where and which business administration systems are used in the business process e.g. SAP, Oracle or Microsoft Dynamics 365 and how they need to interact with the AI solution. What are data flows, relevant data and metadata, how must it be deployed for which costs.

3) compliance aspects and risks. E.g. which standards or legal acts are applicable (e.g., AI-Act, GDPR, DORA...). How to ensure compliance e.g. through employee training, definition of governance processes, and how to ensure robustness and seamless integration of AI applications or data sharing principles of different systems. Also, what are the application investments and their benefits over time.

In the second part of this research, models and modelling techniques to get AI into companies are mapped to the items consisting of materials and prototypes from Figure 9 above. An overview of the different models and their usage for the items is illustrated in Figure 10.

	CRF	FLEX	OTHERS
Identification	<ul style="list-style-type: none"> Storyboarding based on scenes BPMN Natural Language ... 	<ul style="list-style-type: none"> Storyboarding based on scenes BPMN Natural Language
Specification	<ul style="list-style-type: none"> Natural Language ER Model BPMN DMN ... 	<ul style="list-style-type: none"> Natural Language BPMN Archimate
Configuration	<ul style="list-style-type: none"> Decision Support through Decision Tree Resource Allocation MAS-based Production Planning Service with a Hybrid Approach Resource Allocation using Linear Sum Assignment Solver Decision Support through Decision Tree ... 	<ul style="list-style-type: none"> Support Machine Maintenance using RAG and LLM Document Transformation using LLM Support Compliance for Clean Room using RAG and LLM Calibration Certification Service ... 	<ul style="list-style-type: none"> AI-based spend management AI-based BPM creation UI/UX prototyping ...

Figure 10: Models to Introduce AI into companies.

In the “**Identification**” layer for FLEX and CRF natural language to describe the problem settings during the modelling workshops was used referring to item “Deliverable 2.1” of Figure 9. Also, Scene2Model⁷⁴ is a less structured and formal modelling technique and is described as using the modelling method of Storyboarding based on scenes. This technique uses physical modelling and haptic paper figures to represent ideas and concepts. The “process landscape” and “business process” models are outlined using the BPMN format.

In the **“Specification” layer** natural language is used for data samples (e.g., machine repair instructions) but also for describing the decision models in textual form as done in Deliverable 5.1. In addition, the modelled decision-making processes were specified using the BPMN standard. The rule-based Excel that incorporates the data perspective might be described with entity-relationship (ER) models detailing how data is processed and capturing the flow of data. DMN was applied for describing the decision point e.g. of FL or the service Support Understanding of Decisions through Conceptual Modelling (DMN) in Deliverable 3.2. Also, an ArchiMate model is used to describe the IT infrastructure of FLEX and to depict the context which data source describes which aspects for maintenance.

In the **“Configuration” layer** models correspond directly to the AI models developed in FAIRWork such as the AI solution using CP and RL models for production planning to assist decision-makers in assigning orders and workers to production lines or the AI solution using RAG and LLMs to support machine maintenance. Besides that, the AI service itself can be described as a model, the instantiation of the DAI-DSS for one of the many different use cases can be interpreted as one form of configuration of the overall system using different domain and technical models.

For example, an overview of three instances or prototypes of the DAI-DSS configured for different scenarios is given in Figure 11. For each prototype, the AI services use case configurations, orchestration and data and user interfaces differ. To describe the prototypes, domain models and technical models are used. Domain models which are derived in the previous layers (e.g. decision-making process to assist fair worker allocation or the data-source model to improve the information access for maintenance) can be used to configure while the technical models represent the AI and use-case specific workflow triggered by receiving input and finish by showing the output in the user interface (UI). The sequences of the technical models describe how the orchestrator behaves to generate the output. Based on the technical models presented in Figure 11, the suggestion is that AI solutions retrieve data, access configuration information etc. differently. The domain models are described in BPMN or ArchiMate and can be used explicitly or implicitly to configure the AI service and prototype. For some domain models like Machine Maintenance, the model serves as input to provide context information, for other use cases the model can serve as an output to store results in a structured way.

AI-Service	Scenario	Technology	Domain Models	Technical Model
Production Planning Service with a Hybrid Approach	Assist Decisions about Fair Worker Allocation and Production Planning	RL, CP		
Support Machine Maintenance using RAG and LLM	Improve Information Access to Support Maintenance	LLM, RAG		
Document Transformation using LLM	Improve Reliability of “Documentation about Quality Check”	LLM		
...

Figure 11: Examples of model-based configuration and orchestration.

The technical models can also be determined for the services from 3rd parties like the AI-based spend management platform, the AI-based business process generation service or the UI/UX prototyping platform is also seen as a computational model consisting of multiple components such as workflows or AI services.

3.7.2.4 Results

The proposed methodology shows an initial approach to support the introduction of AI with the help of models into companies within the FAIRWork project and its reflection of OTHER use cases. Different items are used to identify, specify and configure AI solutions. These items can also be described through models of various types ranging from mathematical, and computational to conceptual. The proposed framework shows one method on how different stages of getting AI into the companies can be supported through modelling techniques. For Identification, techniques such as natural language description, Storyboarding, or BPMN are used. For Specification BPMN, ER, DMN and natural language are applied. Domain models can be used to support configuration, provide information or store output. For configuration, the configured computational model of the AI application is mapped. Furthermore, the configuration and orchestration of overall prototypes combining multiple components like RAG and LLM, data and user interfaces, aim to be described with technical models. The technical models describe different sequences of combining and orchestrating the individual parts of an AI application to receive a use-case solution.

3.7.2.5 Integration into the DAI-DSS Architecture

The results of the present research contribute to the DAI-DSS configurator and DAI-DSS orchestrator components present in FAIRWork's High-Level Architecture. The domain models use the knowledge depicted with models for the configuration of the specific DAI-DSS instances with the general aim of supporting suitable AI solution selection and its configuration. The technical models analyse the different sequences of the orchestrator and aim to support the identification of orchestration alternatives depending on the use case and the applied AI solution.

3.7.3 Using Conceptual Models to Support Explanations within Decision Support Systems

In FAIRWork's Deliverable 3.2, one focus regarding conceptual modelling was on how conceptual modelling can be used within FAIRWork to capture important knowledge about decision scenarios and to later configure the DAI-DSS. The approach is based on the Design Methodology introduced in our Deliverable 2.1 and in (Woitsch et al., 2024)⁷⁵, and the three-layered method that was introduced in our Deliverable 3.2 and further discussed in the beginning of this section. In this approach conceptual models are used on different abstraction layers to support the configuration of the DAI-DSS to enable the system to propose solutions to decision problems. The abstraction layers are defined on the method's layers (*Identification, Specification and Configuration*), as each layer has a different purpose and therefore needs different abstraction and different modelling languages.

In the Design Methodology the models start on a high-abstraction level and are created in physical workshops, using Scene2Model (Muck & Palkovits-Rauter, 2022)⁷⁶. Later it uses modelling languages like BPMN or DMN to capture more detailed information about the decisions. The added information can be become as detailed to automatically instantiate decision services, as described in the prototype of Woitsch et al. (2024)⁷⁷. In this context, it was mostly looked at on how to use conceptual models in the direction from the scenario understanding towards the decision services.

This section will now discuss how conceptual models and functionality offered through corresponding modelling tools can support the understanding of decisions scenarios, e.g., by reusing data from decisions suggested by the

DAI-DSS and mapping this to conceptual models. The focus will be set on models on a high abstraction level, created with the Scene2Model tool, showing the decision scenarios on a high abstraction level. This idea was also published in (Muck et al., 2024)⁷⁸.

3.7.3.1 Motivation

Conceptual, diagrammatic models use a meaningful and abstract visual representation to make information easy understandable by humans (Larkin, 1987⁷⁹; Mayr & Thalheim, 2020)⁸⁰. They are used in a wide array of domains to capture knowledge. Further, by using metamodeling and implementing them in tools (Bork et al., 2018)⁸¹, their value can be further enhanced, as this enables machines to understand and process them, enabling to further support users.

Conceptual modelling is often used in computer science to support the design and implementation of software systems (Mayr & Thalheim, 2020)⁸² or to describe enterprises, supporting users to better understand and adapt them (Vernadat, 2020)⁸³. Independent for what the models are used, they must be created before they can be used. This creation must not be done manually but they modelling tools can be used to create or adapt the models automatically, easing the work for users. For example, models can be automatically fed with information from information systems, to ease the understanding of the current state or problems within a running information system (Szvetits & Zdun, 2016)⁸⁴.

Models with a semantically rich visual representation can further ease the understanding of the human users (Moody, 2009)⁸⁵. Therefore, in the beginning of FAIRWork's Design Methodology we use physical workshops and digital models to capture the important aspects of the decision scenario and represent in comprehensible way. This was done with the Scene2Model tool, supporting the physical workshops and the digitalization of the workshop results (Miron et. al., 2019)⁸⁶. Discussing and visualising scenarios on a high-abstraction level allows people with different backgrounds to communicate and understand effectively.

The models created for the configuration of the DAI-DSS contain semantics, which will be used to explain the decisions to involved stakeholder. Here the models cannot only show the general scenarios which are captured for the configuration, but they can be enhanced with information and DAI-DSS's results to show different variation. This section will focus on how the high-level models created with Scene2Model can be processed to show more information about the decision scenario.

The expressability of models can only be enhanced by adding the information to the model, but to also visualise within the models. Further, tailoring the concepts and the visualisation to the users supports the understanding and interpretability of the concepts. Therefore, it is important to adapt the modelling methods to the needs and context of the current users. But not only the adaptation of the concepts can improve the interpretability, but also additional functionality can support this and increase the model value. For example, to influence the visualisation by providing additional textual description or animating the objects.

3.7.3.2 Innovation beyond the State-of-the-art

Models are used in the design time of systems to capture the information needed to understand and build them. These models usually belong to pre-defined modelling methods, providing the semantic to the models, facilitating a better understanding of the models from people who know the method. On the other hand, people not familiar with the modelling method, may not understand the complete meaning of the models that easily.

Model@Runtime is a research domain, investigating how updating models with information about the current state of the system they represent can help the users to better understand the system and its current state. Often system near modelling methods are used to represent that data, which further limits the pool of people who can easily comprehend the models and their meaning.

In this part of the project, we investigate how models on high abstraction levels, which are currently getting increasingly attention for designing complex systems, can be enriched with information from the decision they are representing. The basic concept of mapping created data to models from the Models@Runtime approach can be used, but in our research, we do not need to show current runtime data, but to add the data when the user wants to understand it. What is novel in this approach is, that the mapping should not be done to a specific set of models which are defined in advance, but to enable the mapping of the created information to different modelling methods without the need to implement a new endpoint for each of the modelling methods.

3.7.3.3 Description of Functionality

The idea and functionality described in this section were implemented in a research prototype, which will be introduced in Section 3.7.3.5 (Experiments). The idea is to use information created during the use of the DAI-DSS to enrich models and show concrete decision scenarios. The conceptual overview of this idea is visualised in Figure 12.

The results created by the DAI-DSS are saved within its knowledge base. To visualise the created information in models, the information must be transformed in a structure that can be processed by the modelling tool. Therefore, an endpoint is needed which is aware of the metamodel of the models, the metamodel of the decision results and transform the information for the decision results to fit to the metamodel.

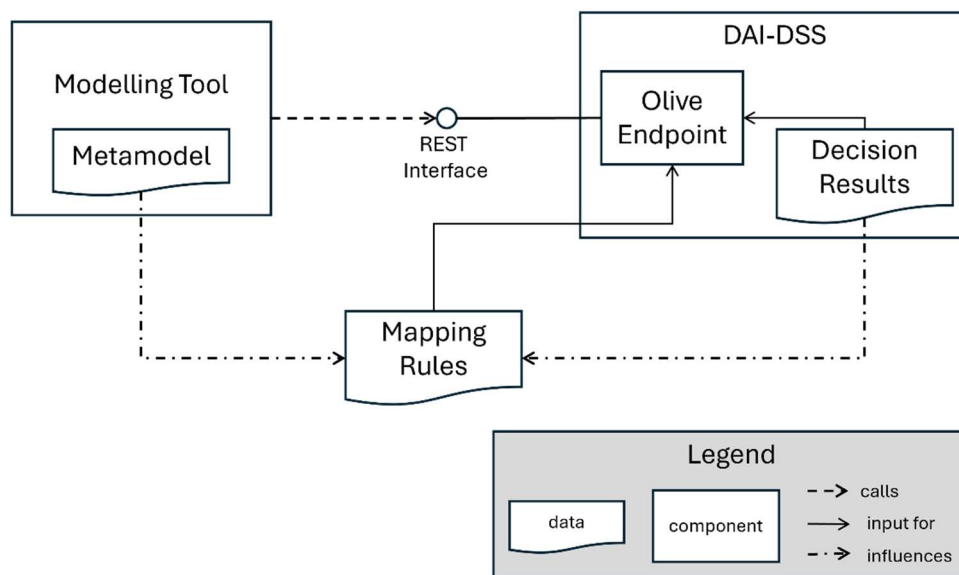


Figure 12: Conceptual overview for mapping decision results to conceptual models.

The mapping rules must be created, containing which concept of the decision result fits to which modelling object. For example, the resilience of a worker provided in the decision result, must be mapped to the resilience attribute of the modelling object representing the worker. Therefore, rules must be defined which define how the mapping

should be done. Further, rules are needed to define what information from the decision results should be shown how in the model. For example, this could be to create new objects, add attributes to modelling objects, create relations between objects.

These two sets of information are saved in the **Mapping Rules** and are influenced and need information about the metamodels of the modelling tool and the data structure. To make this mapping workable a dedicated service must be instantiated, which knows where the data can be decision results can be gathered and to which the mapping rules can be uploaded. Then an endpoint must be offered, which can be called from the modelling tool and once the endpoint is called, the service then collects the needed decision results, applies the mapping rules and returns data that can be consumed by the modelling tool.

The service is able to offer different endpoints for different decisions and ways to represent them in the models. To enable this, the design decision was made that an Olive connector is implemented, which can be configured with the mapping rules. The call must contain the URL from where the decision result information can be collected. In this way, the mapping rules must not be shared with the end user of the modelling tools but can be reused by reusing the endpoint.

With the modelling tool a version of the model must be created that can be changed, to not change the basic model describing the identified scenario on a generic level. Afterwards the functionality must be triggered within the modelling tool, which calls the REST endpoint. Afterward, the provided information applied to the current model, which can be done as the mapping to the metamodel was integrated into the mapping rules. The resulting model can then be used like any other model and further enhanced, processed and so on.

When defining the mapping rules, it is important that an order of rules can be defined, because this can influence the results. For example, if objects should be created from the decision results and then enriched with attributes. The rule to create them must be fired before information can be added.

The modelling tool itself must also be able to store all the information needed, which may call for the need to adapt the metamodel, as not all attributes and concepts from the made decision may be available in the modelling method. Here the need is to enhance the metamodel to show this information and support the models' comprehensibility. The modelling tool can also offer additional functionality to support the understanding, which are tailored to the concrete modelling method.

3.7.3.4 Interfaces

The standalone modelling tool has a graphical user interface (GUI), allowing to create and manipulates models and to trigger the functionality to gather information that should be integrated. Olive has a GUI to configure the specific endpoints and upload the mapping rules. After the configuration Olive offers a REST interface which can be called by the modelling tool to gather the information.

3.7.3.5 Experiments

The focus of the prototype is based on the Scene2Model tool, which is implemented on the ADOxx metamodeling platform and used in the beginning of the project to gather the first insights of the decision scenarios and is used within the proposed Design Methodology for the DAI-DSS.

This tool enables the creation of conceptual model on a high abstraction level and tailoring the available concepts to the current domain and needs. Semantically rich pictures are used to represent the decision scenarios, with

visualisations that are understandable by the users. This enables an intuitive understanding of the general aspects of a decision. The adaptations can also be applied during the runtime of the modelling tool, meaning that the metamodel must not be fixed during the design time, but can also be adapted later to tailor it to support the explanation.

For the integration of the decision results into the modelling tool, a prototype extension was created for the Scene2Model tool, and the Olive controller was enhanced with an additional connector, allowing the instantiation of endpoints for the mapping.

To use the Scene2Model extension a metamodel must be used that is fitting to the domain, which can be adapted with the Scene2Model functionality. Then the extension must be included and configured to connect it to the Olive endpoint. Afterwards, the functionality can be triggered, and the mapping results can be shown in the Scene2Model tool.

The Olive endpoint consumes a JSON file for the configuration, containing the defined rules and can then apply it to the data gathered for a .csv saved as output for an made decision. The rules can map the header of .csv to concepts or attributes within the model. Then each line of the .csv is processed mapped and based on the rule influences the model, e.g. by creating objects, adding attributes or create relations.

The JSON starts with the mapping of the objects to the headers of the .csv, including the attribute that the value in the .csv should be mapped to in the modelling object. For defining what to change in the modelling tool, the JSON file contains than rules that should be applied. These rules are saved in an array to provide an order in which they are called. If something cannot be mapped, because information is missing on in the data or the mapping rules, then it is skipped.

The endpoint provides another JSON containing all the information which should be applied to the model, also in an array so that the order is kept. The modelling tool then applies these definitions and adapts the models accordingly.

After the information is added to the Scene2Model models, the modelling tool can be used to further enhance the models to support the explanations. For example, the animation add-on of Scene2Model can be used to animate the models and emphasis a time axis represented within the model.

3.7.3.6 Results

The result is the framework that was designed to integrate the data into modelling tools and the corresponding research prototype, that can be used with the Scene2Model tool and instantiated in Olive.

Mapping the results to the model enriches the models and allows to reuse them to show case concrete scenarios. Integrating this information automatically eases the manual work and allows the user to focus on understanding or representing a scenario and not to fill it in all by hand before it can be started.

For the high-level models no way was found to easily position created modelling objects in a meaningful way, as therefore understanding of the objects and their relation must be known. But this shortcoming can be overcome by reusing the models from the design time. These already have the position for each object. Then only the additional information must be added and visualised. Therefore, it is important that a copy of the base model is kept, so that different cases can be applied to it.

The creation of objects, adding of attributes and the creation of relations is possible and the Olive connector was implemented to be extendible, if new kinds of rules are needed in the future. The olive endpoint and the Scene2Model extension is kept generic, so that the mapping is done through mapping rules. In this way the results could also be applied to other ADOxx based modelling tools, with some adaptations.

The research prototype and additional information can be found on the FAIRWork Innovation Shop:

<https://innovationshop.fairwork-project.eu/items/14/>

3.7.3.7 Integration into the DAI-DSS Architecture

The prototype uses data from the DAI-DSS as input for the adapting the models. The base data structure of the decision results is the .csv structure which was created for the DAI-DSS. The standalone modelling tool provides the GUI for user to interact with the models and provide the visual representation with for the decision scenario.

3.7.4 Outlook

The presented model-based approach collects all items and models that are relevant to introducing overall AI solutions for CRF and FLEX. The aim is to add and elaborate further use cases and AI approaches as well as underlying models enabling a flexible and combinable approach to support organisations in answering “how to introduce AI” ranging from identification to configuration. This effort should result in a model repository (e.g. Figure 13) to enable the mapping of models and items based on use case and company information so that complete patterns can be transferred to similar use cases. Each model or combination of models can be grouped into focus areas e.g. clustered by type Model Editor, enterprise architecture management or AI solutions. These are again associated with certain “functions” or “services” that can be provided to companies like solving optimization issues with CP.

The mapping of the focus areas and functions to support companies in their organisational processes and management, is not part of the current research in FAIRWork and provides an outlook for a possible future research path. This methodology and the collection of items, domain and technical models should be seen as a starting point and will be further developed. Also, the reflection of models to introduce 3rd party AI solutions including the consideration of governance and usage issues can be seen as ongoing work.

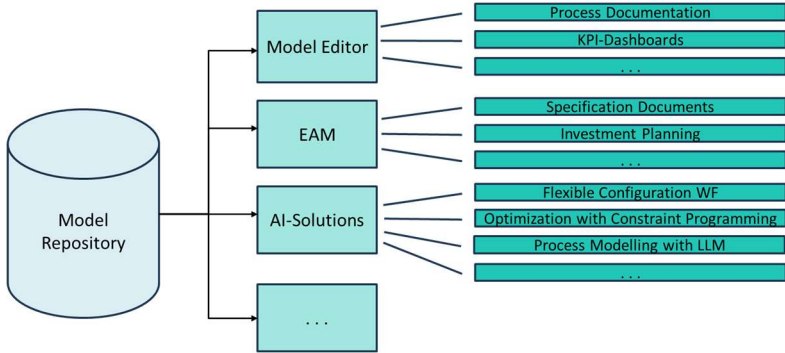


Figure 13: Outlook example for a model repository.

Having a repository of models from different modelling languages also enables the investigation if specific modelling languages can utilize specific mapping from the data created through an information system. Utilizing such language specific aspects could improve the understanding of users familiar with this language. Additionally,

integrating real data from the information system of an organization can not only support the understanding of the scenarios afterwards. But this investigation can is not part of FAIRWork, but an outlook for the future.

3.8 Reliable and Trustworthy AI

3.8.1 Overview

AI systems, especially those that aim to follow democratic approaches, have to be reliable and trustworthy. To maintain human autonomy, meet their requirements, and be of ideal use, the systems, therefore, need to be tailored to human needs rather than forcing humans to adapt to the technology (Shneiderman, 2020⁸⁷, 2022⁸⁸). To achieve this, it is important to pay attention to and respect the end user's perspective and opinion on the systems to be developed. This is why this kind of **human-centered approach** to technologies like AI is one of the three key aspects of the **Industry 5.0** concept (Nahavandi, 2019)⁸⁹.

To foster trust in AI, users typically consider two key questions: Is the system functioning effectively? And is it operating correctly? Addressing these concerns requires transparency regarding both performance and the underlying mechanisms of AI. This aligns with the **Key Requirements for trustworthy AI** identified by the European Commission's high-level expert group on artificial intelligence (AI HLEG 2019a⁹⁰, 2019b⁹¹), which include human agency, technical robustness, privacy, fairness, and transparency. While most of these aspects must be ensured from a technological side, conveying their successful implementation to users hinges on transparency (Arrieta et al., 2020⁹²; Felzmann et al., 2019⁹³; Mohseni et al., 2021⁹⁴; van Nuenen et al., 2020⁹⁵).

To explore ways of building trust in AI, we specifically examined how **transparency** influences users' trust in an AI system. In Deliverable 3.1, we provided a comprehensive overview of the current research on trust in AI systems. In Deliverable 3.2, we expanded on our approach to enhance trust through the transparency of AI systems. Trust is a crucial prerequisite for technology acceptance, adoption, and usage in general (Venkatesh et al., 2016)⁹⁶ and particularly for AI (Siau & Wang, 2018)⁹⁷. In this final Deliverable 3.3, we describe all studies and measures taken throughout FAIRWork. Their aim was a) to analyze transparency from users' perspectives and b) to enable developers of AI to apply the required transparency in their services.

While developers and AI experts have made strides in interpretability and explainability from a technological perspective (Arrieta et al., 2020⁹⁸; Rai, 2020⁹⁹; Murdoch et al., 2019¹⁰⁰), many solutions categorized under explainability are not easily comprehensible to lay users. Researchers like Paéz, 2019¹⁰¹ and Miller, 2018¹⁰² have emphasized the need to focus on end users. Paéz advocated for research into understandability, arguing that understanding how to make AI truly comprehensible is more critical than providing intricate details about its processes if we want to enhance trust. These calls have spurred an increasing amount of research focused on transparency for end users.

Based on these accounts, our research trajectory can be described under the following **research questions**:

- 1) How can AI decision support systems be designed to foster trust?
- 2) What does AI transparency comprise for lay users beyond the technical approaches of Explainability?
- 3) How do different AI system factors influence the effect of transparency on trust, acceptance, and usage?

- 4) How do different types of transparency influence trust, acceptance, and usage of AI services?
- 5) How can the requirements of lay users towards AI transparency be applied to systems at production lines and Multi-Agent Systems, i.e., at the FAIRWork project?
- 6) How can developers of AI services be enabled to set up transparent AI services for their respective end users?
- 7) How can AI transparency be provided in an understandable way for the stakeholders in FAIRWork?

To address these questions, we conducted different studies on AI transparency, with a focus on the perception of lay users. We conducted a qualitative focus group study, as detailed in Section 3.8.2, to find out more about the users' requirements regarding AI systems' transparency. This approach was combined with a qualitative study, described in Section 3.8.3, focusing on comparing the different types of transparency that can be introduced to an AI system and have different impacts on users' trust in the system. However, as these sections were already detailed in Deliverable 3.2, they are only summarized in this deliverable. Thus, additional detail for these subsections can be derived from Deliverable 3.2.

Additionally, different perspectives need to be taken into account, to answer the questions above fully:

On the one hand, we are working together with the technical partners in the FAIRWork project to ensure that our findings regarding transparency can be implemented in the different AI-services. To achieve this, we created a **transparency matrix** that depicts how the characteristics of subjective system factors affect the need for various transparency measures aimed at fostering trust. This matrix was created as a result of the studies and interviews with lay users mentioned above. It was built to enhance the technical developments of FAIRWork and ensure they are implemented with a user-centered approach. What is more, it can be used beyond the project by developers of AI services to be able to develop their AI systems in a transparent way or to check existing AI services for transparency. More details on this can be found in Section 3.8.4. Additionally, we created a table tailored to the different FAIRWork services and their use cases. This table compares how different aspects of transparency measures are important and could be implemented in the different services, as they are different not only in the use case where they will be applied but also in the type of AI used and how it is implemented. Thus, as a first step, we organized a workshop with all of the service partners to identify optimal methods for incorporating transparency into the AI services. Afterward, as a second step, we expanded on these findings and deepened our understanding through workshops with every service developer separately to discuss and consult on transparency implementations tailored to each service. Thus, these workshops and the developed table will assist computer scientists in developing AI services that adhere to transparency standards. The results of this can be found in Section 3.8.6.

On the other hand, we are collaborating closely with the FAIRWork **use-case** partners. During our visit to FLEX in Althofen, Austria, we gained valuable insights into trust and transparency factors, as well as the acceptance and usage of an AI-based decision support system. We interviewed operators and managers from FLEX Althofen to gather this information. Additionally, during a workshop held at FLEX Timisoara, we collected participants' fears and hopes regarding a democratic AI decision support system. To support the implementation of FAIRWork services, participants from the FLEX Althofen plant completed several questionnaires regarding their current working conditions, including workload perceptions, views on existing decision-making processes, and attitudes toward automated systems. These results can be found in Section 3.8.5.

Through our work in the project, we produced important research findings that can be **applied beyond** the FAIRWork project to improve the design and implementation of AI systems in workplace settings. Additionally, by cooperating with other work packages and service partners, we have ensured that our findings on the transparency of AI systems also improve trust in AI in the project itself. More information about the results of this section can also be obtained through the innovation item “AI Transparency for Trust”, see Section 2, “Innovation Shop”.

3.8.2 Qualitative Focus Groups about AI Transparency

A detailed description of this method can be found in Deliverable 3.2. *First DAI-DSS Research Collection*. This section discusses the importance of transparency to establish trust among users in DAI-DSS, particularly lay users who are not IT experts. The main points will be summarised in the following:

While explainability has long been claimed to be a key factor in enhancing trust (Arrieta et al., 2020)¹⁰³, research shows that the main factors fostering trust in AI systems are good performance and reliable communication about that performance (Kaplan et al., 2023)¹⁰⁴. However, different stakeholders have varying expectations regarding transparency, necessitating an understanding of lay users' perspectives to foster trust. Previous research indicates that transparency can bolster trust (Mohseni et al., 2021)¹⁰⁵, especially when mistakes occur (Werz et al., 2020)¹⁰⁶. However, results regarding its effectiveness have been inconclusive due to a lack of a common definition of transparency and what people require from it.

To address this, this study investigated laypeople's needs for transparency in DAI-DSS depending on given system factors of AI. A qualitative analysis was conducted in which 26 participants in three focus groups discussed three fictitious AI applications. The discussions aimed to uncover what explanations users expect from these apps, which transparency information they require, and how this information depends on different system factors. The analysis revealed that lay users' understanding of transparency extends beyond technical explanations. It identified three main pillars: the domain's relevance and prior experiences with systems, the necessity for background information beyond local and global explainability, and the significance of potential errors in outcomes. On an individual level, participants' experiences significantly influenced their attitudes toward transparency. For example, scepticism towards a finance app arose from negative experiences with financial institutions, while familiarity with music services led to greater openness towards respective AI (see Figure 14 for the core results of the study).

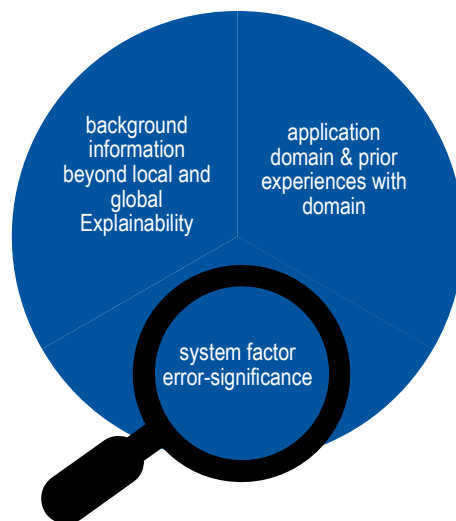


Figure 14: Core results from the quantitative focus group analysis: What does transparency mean for lay users and which factors influence the transparency requirements towards AI transparency.

Transparency concerns were often specific rather than holistic, focusing on security measures or data privacy based on perceived risks associated with each application. Moreover, users did not distinguish between global and local transparency; rather, their demands encompassed broader aspects of system functionality. The significance of potential errors heightened the demand for comprehensive background information across all applications.

The findings emphasize the need for user involvement in designing transparent AI systems in order to match system-dependent influences as well as individual and user group-specific factors. Overall, transparency demands are dynamic and influenced by application types, user backgrounds, and system features. An overview of the results has been published in Werz et al. (2024)¹⁰⁷.

The results from this section can also be called up as an innovation item, see Section 2, “Innovation Shop”.

3.8.3 Quantitative Experiment Comparing AI Transparency Methods

Deliverable 3.2. *First DAI-DSS Research Collection* provides a detailed description of this method. This section explores the complex relationship between AI transparency and user trust. For this, a quantitative study compared different transparency methods. The main points of the study will be summarised below.

Transparency can have paradoxical effects, sometimes even reducing trust based on context, implementation, and target audience (Dasher & Obermaier, 2022¹⁰⁸; Springer & Whittaker, 2018¹⁰⁹). While there is no universal definition of transparency (Ali et al., 2023)¹¹⁰, two main types of transparency can be distinguished: local explanations (specific results; Carvalho et al., 2019)¹¹¹ and global explanations (overall system functionality; Molnar, 2019)¹¹². Local explanations focus on explaining why a single result occurred, while global ones explain the system as a whole. The effects of different transparency types on user attitudes remain inconclusive. Additionally, various technological implementations of transparency have yet to be systematically compared regarding their impact on trust and usage.

To investigate this, a quantitative experiment was conducted with 151 participants to assess how different transparency types affect trust and usage. Four transparency conditions were tested: global functionality, global accuracy, local accuracy, and local functionality and compared with a non-transparent condition. Participants could

use the advice of AI for weight estimations from pictures and provide feedback on their trust in the algorithms. The users evaluated all transparency types. Results indicated that transparency significantly influences both trust and algorithm usage. Trust significantly varied among the four examined types of transparency, to a lesser extent, the usage of the algorithms varied as well. Transparency measures were used and trusted more than the non-transparent algorithm. Global transparency measures proved most effective in building trust.

Overall, the findings suggest that while all forms of transparency enhance trust levels differently, users particularly value general background information about algorithms' developers and testing processes for establishing initial trust. Local explanations also play an important role during algorithm use but may not be as effective in fostering initial trust as global measures.

3.8.4 Matrix and Guidelines on the Application of Transparency from a Lay User Perspective

3.8.4.1 Motivation and Reference to FAIRWork Use case

For a long time, mainly developers and computer scientists have been researching AI transparency, and technological aspects have been the focus. However, different stakeholders need to be addressed differently: IT experts are expecting other information from a system that lay users (Mohseni et al., 2021¹¹³; van Nuenen et al., 2020¹¹⁴). In FAIRWork, the users of the developed systems will mainly be people who are not computer experts but domain experts or workers. For democratic decision-making, informed usage of the systems at hand is a core concern, which is why the understanding of transparency of the user group domain experts and workers has to be established. Transparency that is implemented successfully should increase their trust, acceptance, and usage of a DAI-DSS and provide autonomy to the involved people. This is a central requirement for legitimating their representation through the systems.

3.8.4.2 Innovation beyond the State-of-the-art

Based on the many inconclusive findings as well as the dependence of transparency requirements on system factors, many developers of AI need support in setting up transparency measures for their systems. What is more, the requirements end users pose towards system transparency differ significantly from those of AI experts and developers of systems, as the previous studies have shown (Sections 3.8.2 and 3.8.3). What is more, the previous studies shed light on the aforementioned dependencies: Which aspects of transparency are central for the user group of non-AI-experts? How do transparency requirements depend on system factors?

A transparency matrix was developed to combine the studies' results and provide AI experts with their respective expertise.

3.8.4.3 Method of Development and Description of Matrix

To be able to derive the relevant transparency aspects from given system factors the results of the previous studies were discussed with psychology and AI experts. Based on these workshops, a matrix was set up that lists the given system factors in the first row. These system factors comprise, e.g., the error relevance: How severe is a mistake of the system? Another factor of an AI system could be the sensitivity of the input or processed data. Also, previous experience with the system is one important factor that influences how users perceive a system and its transparency.

Based on the system factors, a user of the matrix can derive 13 implications for AI transparency. These implications show which user needs emerge based on given system factors regarding AI transparency. The transparency implications comprise, for instance, control over decisions, global and/or local explanations, accuracy evaluations, or insights into data processing practices.

The recommended approach for applying the transparency matrix consists of four steps:

1. Decomposing the AI into its process steps for detailed analysis: Especially for complex AI systems, it is beneficial to decompose them into individual process steps before conducting an analysis. For example, one might identify steps such as (a) data input and (b) suggestions.
2. Identification of which subjective system factors apply to the specific system: It is important to note that the system factors are subjective and can be perceived differently by users. For instance, certain individuals may perceive specific data input as being sensitive, while others have no such concerns.
3. Deriving transparency requirements based on these characteristics.
4. Evaluating outcomes to verify the correct identification of these properties and ensure that transparency measures enhance understanding and support usage.

The matrix is currently being finalized and will be published in the year 2025.

The results from this section can also be called up as an innovation item, see Section 5, "Innovation Shop".

3.8.4.4 Integration into the DAI-DSS Architecture

As an exemplary application and an evaluation of the matrix, we conducted several workshops with the service partners of FAIRWork. The goal was to give them a hand in developing transparency for their respective services as well as deriving options for improving the matrix. The results that emerged for the individual services will be described in Section 3.8.6 in more detail.

3.8.5 Evaluation of Requirements and the Status Prior to the Introduction of a DAI DSS to the Use Case Partners

A detailed description of this method can be found in Deliverable 3.2. *First DAI-DSS Research Collection*. This method gathers user input and the status quo before introducing a DAI-DSS, at the use case sites. To achieve the workers' trust and acceptance and as requirements and prerequisites for the development of the DAI-DSS, a status quo study and a workshop with employees regarding their hopes and fears for the DAI-DSS were conducted.

The status quo study was conducted at FLEX in Austria. It consisted of three parts: The employee workload can give important insights regarding specific tasks or employee groups that benefit the most from a DAI-DSS. The workers described their overall workload as moderate. One notable result was that supervisors reported a relatively high mental demand during their week, indicating that a DSS might be especially useful for supervisors. The second part of the questionnaire concerned the employees' attitude towards management decisions. This was done to gauge the employees' current level of trust in supervisors and their attitude toward their decisions before introducing a DAI-DSS. The employees reported highly positive views toward decisions made by their supervisors. Since employees seem content with their supervisors' decision-making, this contentedness must not be allowed to deteriorate through the introduction of the DAI-DSS. The questionnaire closed with a few questions concerning the

attitude regarding automated systems to have a baseline for the employees' attitude in that regard. The employees displayed a moderate level of trust and attitude toward automated systems. However, they also displayed a need to be careful with unknown automated systems. This makes a careful introduction of the DAI-DSS all the more important.

Additionally, we gathered further input from employees at the FLEX company site in Romania during a science fair workshop conducted during the FAIRWork Partner Meeting in November 2023. Here, we gathered employees' fears and hopes for a decision-support system. The main fears were that the system might propose wrong or suboptimal solutions and that, due to a lack of transparency, there might not be a way to be sure that the proposed decision is the best or that the underlying data is correct. Related to a possible lack of transparency was the fear that people might not understand why a decision was made and might, therefore, feel like they were treated unfairly. Other fears contained a lack of choice or a growing dependence on the system. On the other hand, the employees hope for faster and easier answers and decisions for their problems in difficult situations. Lastly, they see a potential to get better and easier insights into company priorities and hope that the system will not only the performance parameters of the workers but also their preferences for different tasks or workplaces. The system might even be more neutral than a human and treat the employees more fairly and unbiasedly.

3.8.6 Practical Application of Transparency in Different DAI-DSS Services

3.8.6.1 Motivation and Reference to FAIRWork Use case

The knowledge gained in the other methods described under Section 3.8 is crucial for the different Services of the DAI-DSS and its application to the FAIRWork Use cases. Therefore, this method aims to provide a practical application to gain from previous results.

3.8.6.2 Innovation beyond the State-of-the-art

The studies conducted and presented under Sections 3.8.2 and 3.8.3, whose results lead to the matrix presented under 3.8.4, show that transparency is an important prerequisite for trust in an AI system and can be applied in different ways. The DAI-DSS presents a novel approach, combining many different systems and services. All of these need to be reviewed individually to determine the best ways to enable transparency and apply the presented findings.

3.8.6.3 Description of Functionality

The goal is to apply the knowledge gained during the studies to the different services developed in the FAIRWork project. First, the different DAI-DSS services were to be better understood and analysed. This happened through overarching workshops, as well as focused one-to-one meetings. Second, options were developed to provide more transparency to the different services and enhance their trustworthiness. As these solutions must be tailored to each service, close cooperation and consultation with partners were essential for this part.

3.8.6.4 Experiments

As a first step, a workshop was conducted with different FAIRWork partners to identify ways of providing transparency to the FAIRWork Services of the DAI-DSS. This workshop raised awareness among the partners about the importance of transparency and the different possibilities of applying it to their specific services. It also provided knowledge about the different services and how to enable them to be more transparent.

With the insights from this workshop in mind, one-on-one meetings were conducted with the different service partners. The goal was to enable the application of transparency for their specific services by answering the following questions: How can transparency be set up for the respective service? How reliable are the services, and how can the extent of that reliability be made transparent? How do we measure transparency and reliability? How can these measures be presented to lay users? How do we measure trustworthiness regarding the services?

At the beginning of these workshops, the findings summarized under Section 3.8.4 and their relevance for the given service were discussed. The services were then gone through step by step to determine the input they needed, at which points users interacted with the service in what way, what kind of output the service produced, and if this differed depending on the user's role. Additional attention was also paid to how reliable and accurate the service is in its answers and how it can be made transparent on a global and local level. All of this information was used to discuss, at which points the service can be made more transparent and understandable for users.

3.8.6.5 Results

In the following, a short summary of the first workshop is provided: For many services, textual explanations are preferred with visual supplements. There are a few parts that could be explained in general or for every service, such as „What is the input data?“ However, differences emerge in the data processing of the services. Therefore, these processes must be regarded in greater detail for each service. The different DAI-DSS services vary in complexity. Some are much more easily understood and, therefore, explained than others. However, the complexity also varies depending on the use case and the amount of data being processed. Two points were identified that are of particular interest to explain: The first is when user expectations are violated. If the system's output differs from what the user expected, then an explanation is especially important. The second point is when different services provide different outcomes. Here, an explanation is necessary to help the user understand why one service would recommend one action and a different service another.

The following table (Table 2) shows the results of the one-on-one meetings with the different service partners and displays how different types of transparency could be implemented into different services and applied in different use cases.

Services	MORE: <i>Resource Allocation MAS-based</i>	BOC: <i>Support Machine Maintenance using RAG and LLM</i>	OMiLAB <i>Support Understanding of Decisions through Conceptual Modelling</i>	JR: <i>Resilience Score</i>	JR: <i>Calibration Certification</i>	RWTH: <i>Production Planning with a Hybrid Approach</i>
Global Transparency	<ul style="list-style-type: none"> How does the service work? What are its limitations? What can I expect from the results? 	<ul style="list-style-type: none"> From low to high details / complexity optionally clickable Details: How many similar vectors are used to create the final document? 	<ul style="list-style-type: none"> Which rules are considered? Priority of rules (available > experience > resilience > ...) 	<ul style="list-style-type: none"> Which part of the individual resilience information reaches whom? How is the information used? 	<ul style="list-style-type: none"> Info that the document is compared with the target document Final decision and responsibility still lies with the user 	<ul style="list-style-type: none"> What information / input does the service use? Setting a prioritization
Local Transparency	<ul style="list-style-type: none"> Why was an agent chosen? Possibility to influence the parameters Explanations in different levels of complexity -> parameter tuning (highest complexity) 	<ul style="list-style-type: none"> Link to source document for the answer Reasoning why a specific document is linked Mark parts of the question that were important for the choice of recommendation 	<ul style="list-style-type: none"> Display for workers which rule lead to the allocation Which rules do not apply to the solution (which worker has too little experience, etc.)? 	<ul style="list-style-type: none"> Meaning and characteristics of resilience parameters Classification of workers scores 	<ul style="list-style-type: none"> Feedback that something is OK or not OK Present relevant parts of document with deviating values or spelling Show two pictures highlighting the difference (in 	<ul style="list-style-type: none"> Why was a user assigned to a position? Information relating to production: What assumptions does the model make about the production of a particular product?

	<ul style="list-style-type: none"> Information about how fairness is ensured 	<ul style="list-style-type: none"> Instructions for prompt design 			<ul style="list-style-type: none"> addition to the text) 	
Other Transparency	<ul style="list-style-type: none"> Who is responsible for the service? How was it tested? Etc. 	<ul style="list-style-type: none"> Background information: Authors, test processes, who else uses it? 	<ul style="list-style-type: none"> Background information: authors, testing, etc.? 	<ul style="list-style-type: none"> Information about privacy and data usage Who tested and approved it? (e.g., workers' council) 	<ul style="list-style-type: none"> Background information: authors, testing, etc.? 	<ul style="list-style-type: none"> Background information: authors, testing, etc.?
Accuracy Information	<ul style="list-style-type: none"> suitability scoring for each result suggestion: <ul style="list-style-type: none"> 1. worker A (best suitable): 75% 2. Worker b (second best suitable): 68% 3. worker c (third best suitable): 44% 4. ... 	<ul style="list-style-type: none"> Certainty/ accuracy for the system overall and for individual answers Indication that the system can also make errors Feedback option for correctness of answers Ideally, the system learns from feedback directly, or the system is evaluated and feedback is considered later 	<ul style="list-style-type: none"> How confident is the system with the answer? Fit to preferences, resilience, and experience. Are all deadlines met? How well does the system work in general? (As an average of the above data over time) Give feedback on the system: How satisfied are you? 		<ul style="list-style-type: none"> Information about accuracy and that for signatures it is only compared, whether there is a signature at all or not 	<ul style="list-style-type: none"> Visualization of how well the target prioritization is met

Table 2: Results from workshops with different service partners with regard to the transparency in their services.

This table and its contents can also be called up as an innovation item, see Section 2, “Innovation Shop”.

3.8.6.6 Integration into the DAI-DSS Architecture

Integration into the DAI-DSS architecture will happen with the service partners. In workshops and one-on-one meetings, solutions tailored to each service were discussed, which will then be implemented by the partner responsible for the corresponding service. Further workshops will accompany this process.

3.8.7 Outlook

The following outlook contains, on the one hand, processes for the methods that are currently still ongoing and, on the other hand, further research.

- Regarding the application of our findings to the DAI-DSS, one-on-one meetings with the different partners will continue during the implementation of the services to support their application of transparency measures, enabling higher transparency and trustworthiness of the DAI-DSS system and its services.
- As part of Work Package 6 “Evaluation”, the introduction of DAI-DSS services will be accompanied by further surveys of employees at the production sites of the use case partners to identify potential changes through introducing the DAI-DSS and evaluate the implementation.
- Further Research: Validation of results found, e.g., further studies with lay users. Identify factors to foster understanding as an important component of trust of non-experts.
- Further Research: Longitudinal studies for the exploration of long-term effects of implementing a DAI-DSS into a company
- Further Research: How to communicate transparency and how to foster trust for those who are indirectly concerned by AI systems?
- Further Research: Reasons for AI aversion and avoidance. How can this be counteracted? What influence does control have?

4 EXPLAINABILITY AND FAIRNESS IN AI SERVICES

4.1 Overview

Explainable AI and fairness of AI services in the context of industrial manufacturing environments are one of the key objectives in the project FAIRWork. In the first part of the project, the focus was firstly on which AI services would support the use cases that the industrial partners provided as key issues that require intelligent solutions. In the second part of the project, we will focus on how “explain-ability” and, particularly, fairness has to be introduced into the socio-technical systems, its services, i.e., the algorithmic decision-making.

In this Deliverable 3.2 we describe basic principles of explainable AI and fairness in AI as well as first general ideas on how to proceed with implementations in the project FAIRWork. In the final Deliverable 3.3 of work package WP3 we will describe then the concrete implementations of explainable AI as well as fairness in decision-making as well as the results of these realisations.

In this Section, we firstly particularly focus on the **transparency in algorithms** that human IT experts would be able to understand. This transparency has been taken up by the framework of **XAI**, often overlapping with Interpretable AI, or XML. XAI either refers to an AI system over which it is possible for humans to retain intellectual oversight or refers to the methods to achieve this (Mihály, 2023¹¹⁵; Longo et al., 2024¹¹⁶). The focus is usually on the **reasoning behind the decisions or predictions** made by the AI which are made more understandable and transparent (Vilone & Longo, 2021)¹¹⁷. XAI is counteracting a tendency of “black box” in machine learning, where even the designers of the AI system cannot explain why it arrived at a specific decision (Castelvecchi, 2016)¹¹⁸.

With the upcoming widespread use of AI systems and applications in industrial environments, accounting for **fairness** has gained significant importance in designing and engineering of such systems. AI systems will be used in DAI-DSS in sensitive socio-technical environments to make important decisions. We investigate various **mathematical measures** of fairness that will provide **quantitative information** about implicit bias in algorithms that render their decisions “unfair. In the context of decision-making,

“fairness is the absence of any prejudice or favouritism toward an individual or group based on their inherent or acquired characteristics” (Mehrabi et al., 2021)¹¹⁹.

Thus, an unfair algorithm is one whose decisions are skewed toward a particular group of people. Thus, it is mandatory to ensure that these decisions do not reflect discriminatory behaviour toward certain groups or populations of either workers or human decision makers.

One challenge that any software must overcome before being integrated into human-centred routines is algorithm bias. Most learning-based algorithms require large datasets to learn from, but several social groups of the human population have long been unrepresented or misrepresented in existing datasets. If the training data is not representative of the variability of the population, the AI tends to amplify biases, which can lead to a lack of generalisation and thus neglect of workers or decision-makers in the case of FAIRWork, e.g. gender or age specific and ethnic minorities that have always been underrepresented in existing datasets, which can intensify inequalities.

4.2 Explainability and Fairness Introduction

4.2.1 Explainability in FAIRWork

Explainable AI or XAI should enable users to introspect a dynamic system as well as control options to understand how software arrives at a solution to a problem. In order to create transparency with regard to possible discrimination by the AI, FAIRWork considers using characteristic, internationally proven XAI tools. Typically used XAI software are local interpretable model-agnostic explanations (LIME; Ribeiro et al., 2016¹²⁰), shapley additive explanations (SHAP; Lundberg & Lee, 2017¹²¹; Aal et al., 2021¹²²) or the what-if tool (Wexler et al., 2020¹²³).

The Shapley value provides a principled way to explain the predictions of nonlinear models common in the field of machine learning. By interpreting a model trained on a set of features as a value function on a coalition of players, Shapley values provide a natural way to compute which features contribute to a prediction or contribute to the uncertainty of a prediction. This unifies several other methods, including LIME, DeepLIFT, and layer-wise relevance propagation.

XAI tools make it possible to explain and interpret the predictions of machine learning models. However, different AI stakeholder require different types of AI transparency. Many forms of explainability have been developed to serve AI experts with the goal to enable them to gain insights into and improve AI systems. End users, on contrary, often require different types of AI transparency (see Figure 15). For them, understandability is more important than technically detailed explanations (Section 3.8.2). An additional concept of explainability could therefore be named understandability: Its goal is to foster acceptance, trust, and usage among users of AI systems that are not AI experts. Therefore, FAIRWork pursues the aim to establish both explainability for the developers and AI experts that are working on the DSS services as well as to enhance AI transparency for end users with a focus on fostering their trust and acceptance towards the system. First, in the service development the results of the previously described studies have been considered to enable users to understand the systems. Secondly, the front-end development that brings together all DSS for users also entails means for transparency that increase users' understandability of the different services and enables them in an informed selection of a service. More information about this can also be obtained through the innovation item "AI Transparency for Trust", see Section 2, "Innovation Shop".

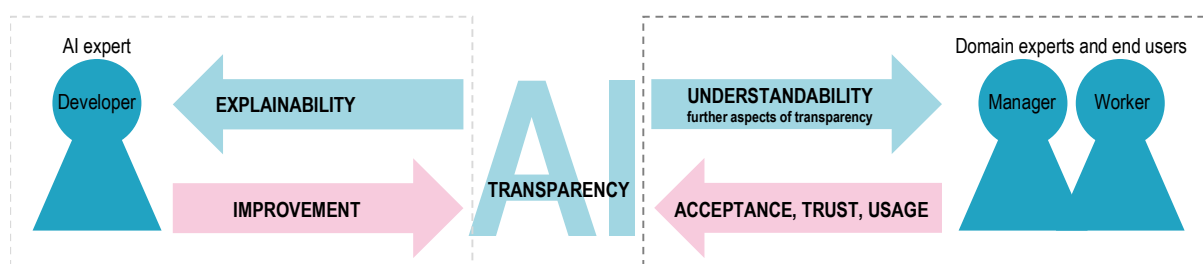


Figure 15: Requirements and effects of AI transparency differ for user groups like AI experts and end users.

Thirdly, explainable technologies can be used to track the specific influence of vulnerable parameters - such as gender, age and country of origin - on the recommendations generated by the intelligent software. FAIRWork considers therefore to track the vulnerable parameters and the context of the respective configuration for each data set. The vulnerable parameters enable the identification of discrimination. The context information makes it possible

to dynamically generate suitable recommendations adapted to the context if the context, such as, for a work allocation, changes over time.

4.2.2 Explainability in CRF Use Cases

In industrial settings, explainability is crucial, especially for scenarios like worker allocation and production scheduling, where the system's outcome directly affects people. Within the FAIRWork project, a linear sum assignment solver method was applied to address the challenge of fair and explainable resource allocation. This algorithm allocates available workers to the lines, using clear objective functions with explicit constraints like worker availability, the geometry of produced parts, workers' medical condition and preferences. For example, breaking down how constraints like worker availability or preferences influence outcomes ensures alignment with operational goals, resulting in a transparent decision-making process. Additionally, visualizing solution steps, such as assigning orders to machines and workers to production lines, allows for clear communication that aims to build trust among the system's users.

4.3 Fairness in FAIRWork

It is of the utmost importance that fairness be a fundamental principle in decision-making processes. This ensures that individuals facing similar circumstances are treated equally and not subjected to discrimination. Examples of unfair decision-making can include situations where individuals are discriminated against based on protected attributes such as race, gender, or age. Unfair decisions can arise when there is a lack of transparency in the decision-making process, leading to outcomes that are perceived as biased or unjust. For example, unfairness can arise when promotions are based on favouritism rather than merit, or when hiring decisions are influenced by personal biases rather than qualifications.

In the project FAIRWork, "fairness" plays a role in several dimensions, as described in more detail, as follows, by several viewpoints.

4.3.1 Application in FAIRWork Approach 1: CRF-Example

Fairness in Decision Making

First, it is clear that numerous decisions are made by groups. Social choice theory (Sen, 1986)¹²⁴ addresses this fundamental aspect, namely the aggregation of individual preferences of group members into a collective decision. The question that must be answered is this: what makes a collective decision a good, i.e. "fair", decision? The aim is to achieve a more precise understanding of the collective decision-making process to master the new technological and social challenges in which aspects of decision-making and fairness are important. We must start with the preferences of individuals, machines, or criteria over a set of discrete objects. Then we must make a "fair" group decision. A major problem of most previous studies is the limited availability of actual preference information. Available information from elections or group decisions is usually limited to the data collected during the election process. Often these are only individual alternatives from a relatively large set of alternatives, such as in plurality voting, where everyone can cast exactly one vote in favour of one alternative. The underlying complete preferences (such as the complete ranking of alternatives) are usually not even recorded. It is therefore difficult to understand or even justify whether the collective result really corresponds to any kind of "collective will".

Second, it is important to analyse the outcome of an AI service, especially for services that deal with the distribution and allocation of resources.

Fairness in Distributed Agent-based System

In agent-based systems, fairness is a crucial factor in ensuring equitable outcomes for all agents (De Jong et al., 2008)¹²⁵ There are various techniques available for implementing fairness, including methods based on decentralised learning, distributed average consensus, and game theory. The objective in all cases is to ensure proportional fairness and envy-freeness. These techniques are essential for achieving fairness in resource allocation (Jiang & Lu, 2019)¹²⁶.

Furthermore, in system dynamics and agent-based modelling simulations, fairness can be conceptualized by considering procedural fairness, which concerns the procedures leading to outcomes, and distributive fairness, which relates to the perception of outcomes as fair or unfair (McGarraghy et al., 2022)¹²⁷.

Fairness in AI Services

We examine the field of algorithm fairness and its objectives. To illustrate the significance of this field, we present examples of unfair models and their implications. Current state and future challenges discuss the challenges of achieving fair algorithmic decision making. The paper explores how bias in the data used to train these algorithms can perpetuate unfairness in real-world decisions (Tolan, 2019)¹²⁸.

However, the use of algorithms for automated decision-making can cause unintentional effects that lead to discrimination against certain specific groups (e.g., in the workload example). In this context, it is crucial to develop AI services that are not only accurate but also fair.

Fairness in Multi-Agent Systems

Fairness in MAS involves more than just designing algorithms, it requires an understanding of human fairness motivations and how these can be modeled and translated into a computational framework (De Jong et al., 2008)¹²⁹ The challenge lies in capturing the complex nuances of human fairness, which often encompasses ethical, social, and emotional dimensions, and embedding these into systems where multiple agents interact. This involves not only ensuring that individual agents operate fairly but also that their interactions lead to outcomes perceived as fair by humans.

The dynamics within MAS often mirror social dilemmas where the interests of the collective clash with the goals of individual agents. In such scenarios, the concept of fairness extends to understanding and balancing these conflicts. Questions about whether agents will cooperate, or act selfishly underscore the importance of designing systems that can manage and ideally reconcile these divergent interests. This requires an understanding of how agents can either contribute to or detract from overall fairness in emergent team behaviors (Gruppen et al., 2021)¹³⁰. The development of cooperative multi-agent fairness thus reframes key questions to focus on whether agents, given incentives to collaborate, can learn to coordinate their actions effectively and fairly. However, the pursuit of fairness in MAS does not come without cost, especially as task difficulty increases. Empirical studies in cooperative multi-agent tasks suggest that while fairness may be relatively "inexpensive" in simpler scenarios — where agent skills are sufficiently high — it can become increasingly costly in more complex situations. As task complexity rises, the challenge intensifies to maintain fairness without compromising the performance or utility of the system, illustrating the delicate balance needed between achieving equitable outcomes and maintaining high performance in MAS.

Furthermore, the broader implications of fairness in decision support systems require a dual perspective that encompasses both algorithmic and societal views. On the one hand, there is a need to develop algorithms capable of balancing different relevant decision factors within a defined context. On the other hand, it is crucial to consider

whether the type of fairness achieved by these algorithms aligns with societal values (Angerschmid et al., 2022¹³¹; Jiang & Lu, 2019¹³²). This distinction highlights the importance of not only designing decision support systems that are fair in a statistical sense but also ensuring that these systems contribute to a form of fairness that is meaningful and desirable within the societal context. This dual perspective underscores the ongoing dialogue and necessary adjustments in how fairness is conceptualized and implemented in both multi-agent systems and broader automated systems.

MAS reproduces these behaviors taking advantage of descriptive models of human fairness that can be further explored with the objective of enhancing decision-making capabilities. In the next steps, we aim to explore fairness aspects in MAS in order to provide a broader socio-technical approach aligned with human values in fact desired in decision support systems.

Fairness and Explainability

Fairness and Explainability in AI-Informed Decision Making explore the relationship between people's perceptions of fairness and how decisions made by AI systems are explained to them. The study suggests that providing explanations can increase trust in the fairness of AI-based decisions (Angerschmid et al., 2022)¹³³.

Fairness and Trust

A Study on Fairness and Trust Perceptions in Automated Decision Making examines the relationship between people's trust in automated decision systems and their understanding of how these systems work. The research highlights that a lack of transparency can lead people to question the fairness of such system.

4.3.2 Application in FAIRWork Approach 2: FLEX-Example

Algorithms for Decision Making that are using statistical and machine learning approaches and that are applied to biosignal sensor-based data highly depend on the modalities of the input data.

In the Human Factors-directed component of the FLEX use case, we intend to sample for the purpose of risk stratification for stress and potential of resilience. Any algorithms in this context should be checked with respect to potential violation of fairness principles.

In the following Sections we give an overview on fairness tools and measures that should be considered when applying decision making methodologies to human-centred data.

Assessment Tools

An interesting direction that researchers have taken is introducing tools that can assess the amount of fairness in a tool or system. For example, Aequitas (Saleiro et al., 2018)¹³⁴ is a toolkit that lets users to test models with regards to several bias and fairness metrics for different population subgroups. Aequitas produces reports from the obtained data that helps data scientists, machine learning researchers, and policymakers to make conscious decisions and avoid harm and damage toward certain populations. AI Fairness 360 (AIF360) is another toolkit developed by IBM in order to help moving fairness research algorithms into an industrial setting and to create a benchmark for fairness algorithms to get evaluated and an environment for fairness researchers to share their ideas (Bellamy et al., 2018)¹³⁵. These types of toolkits can be helpful for learners, researchers, and people working in the industry to move towards developing fair machine learning application away from discriminatory behaviour.

Bias in Data and Algorithms

Many AI systems and algorithms are data driven and require data upon which to be trained. Thus, data is tightly coupled to the functionality of these algorithms and systems. In the cases where the underlying training data contains biases, the algorithms trained on them will learn these biases and reflect them into their predictions. As a result, existing biases in data can affect the algorithms using the data, producing biased outcomes. Algorithms can even amplify and perpetuate existing biases in the data. In addition, algorithms themselves can display biased behaviour due to certain design choices, even if the data itself is not biased. The outcomes of these biased algorithms can then be fed into real-world systems and affect users' decisions, which will result in more biased data for training future algorithms.

Types of Bias

Bias can exist in many shapes and forms, some of which can lead to unfairness in different downstream learning tasks. Surash and Gutttag (2019)¹³⁶ mention sources of bias in machine learning with their categorisations and descriptions in order to motivate future solutions to each of the sources of bias introduced in the paper. Olteano et al. (2019)¹³⁷ prepare a complete list of different types of biases with their corresponding definitions that exist in different cycles from data origins to its collection and its processing. Here we will reiterate the most important sources of bias introduced by Surash and Gutttag (2019)¹³⁸ as well as from Olteano et al. (2019)¹³⁹, integrating the survey of Mehrabi et al. (2021)¹⁴⁰, as follows:

- **Measurement Bias.** Measurement, or reporting, bias arises from how we choose, utilise, and measure particular features (Surash and Gutttag, 2019)¹⁴¹. One should not conclude about people coming from specific social groups are associated with specific feature values different from others and should not apply a difference in how these groups are assessed and interpreted.
- **Omitted Variable Bias.** Omitted variable bias occurs when one or more important variables are left out of the model.
- **Representation Bias.** Representation bias arises from how we sample from a population during data collection process. Non-representative samples lack the diversity of the population, with missing subgroups and other anomalies. Datasets might for example represent more samples from younger than from elder people or being incline in the representation of females in contrast to a majority of data collected from males.
- **Aggregation Bias.** Aggregation bias (or ecological fallacy) arises when false conclusions are drawn about individuals from observing the entire population. Features of various subgroups might differ in many ways, but the model ignores the varieties and makes false conclusions about the diversity in the complete population (such as, in **Simpson's Paradox**; Blyth, 1972)¹⁴².
- **Sampling Bias.** Sampling bias is like representation bias, and it arises due to non-random sampling of subgroups. Because of sampling bias, the trends estimated for one population may not generalise to data collected from a new population.
- **Longitudinal Data Fallacy.** Researchers analysing temporal data must use longitudinal analysis to track cohorts over time to learn their behaviour. Instead, temporal data is often modelled using cross-sectional analysis, which combines diverse cohorts at a single time point. The heterogeneous cohorts can bias cross-sectional analysis, leading to different conclusions than longitudinal analysis.

- **Linking Bias.** Linking bias arises when network attributes obtained from user connections, activities, or interactions differ and misrepresent the true behaviour of the users.
- **Discrimination.** Like bias, discrimination is also a source of unfairness. Discrimination can be considered as a source for unfairness that is due to human prejudice and stereotyping based on the sensitive attributes, which may happen intentionally or unintentionally, while bias can be considered as a source for unfairness that is due to the data collection, sampling, and measurement. Although bias can also be seen as a source of unfairness that is due to human prejudice and stereotyping, in the algorithmic fairness literature it is more intuitive to categorize them as such according to the existing research in these areas.

Definitions of Fairness

Binns (2018)¹⁴³ studied fairness definitions in political philosophy and tried to tie them to machine learning. Authors in studied the 50-year history of fairness definitions in the areas of education and machine-learning. Hutchinson and Mitchell (2019)¹⁴⁴ listed and explained some of the definitions used for fairness in algorithmic classification problems. Saxena et al. (2019)¹⁴⁵ studied the general public's perception of some of these fairness definitions in computer science literature. Here we will reiterate and provide some of the most widely used definitions, along with their explanations inspired from Verma and Rubin (2018)¹⁴⁶.

- **Equalized Odds.** The definition of equalized odds states that the probability of a person in the positive class being correctly assigned a positive outcome. The equalized odds definition states that the protected and unprotected groups should have equal rates for true positives and false positives.
- **Equal Opportunity.** The probability of a person in a positive class being assigned to a positive outcome should be equal for both protected and unprotected (female and male) group members. The equal opportunity definition states that the protected and unprotected groups should have equal true positive rates.
- **Demographic Parity (Statistical Parity).** The likelihood of a positive outcome should be the same regardless of whether the person is in the protected (e.g., female) group.
- **Fairness Through Awareness.** An algorithm is fair if it gives similar predictions to similar individuals. Any two individuals who are similar with respect to a similarity (inverse distance) metric defined for a particular task should receive a similar outcome.
- **Fairness Through Unawareness.** An algorithm is fair as long as any protected attributes are not explicitly used in the decision-making process.
- **Treatment Equality.** Treatment equality is achieved when the ratio of false negatives and false positives is the same for both protected group categories.
- **Test Fairness.** The test fairness definition states that for any predicted probability score S , people in both protected and unprotected groups must have equal probability of correctly belonging to the positive class.
- **Counterfactual Fairness.** The counterfactual fairness definition is based on the intuition that a decision is fair towards an individual if it is the same in both the actual world and a counterfactual world where the individual belonged to a different demographic group.

- **Fairness in Relational Domains.** A notion of fairness that is able to capture the relational structure in a domain—not only by taking attributes of individuals into consideration but by taking into account the social, organisational, and other connections between individuals.
- **Conditional Statistical Parity.** Conditional statistical parity states that people in both protected and unprotected (female and male) groups should have equal probability of being assigned to a positive outcome given a set of legitimate factors.

Fairness definitions fall under different types as follows (Mehrabi et al., 2021)¹⁴⁷:

- **Individual Fairness.** Give similar predictions to similar individuals.
- **Group Fairness.** Treat different groups equally.
- **Subgroup Fairness.** Subgroup fairness intends to obtain the best properties of the group and individual notions of fairness. It is different than these notions but uses them in order to obtain better outcomes. It picks a group fairness constraint like equalising false positive and asks whether this constraint holds over a large collection of subgroups.

Methods for Fair Machine Learning

There have been numerous attempts to address bias in AI in order to achieve fairness; these stem from domains of AI. Generally, methods that target biases in the algorithms fall under three categories (Mehrabi et al., 2021)¹⁴⁸:

- **Pre-processing.** Pre-processing techniques try to transform the data so that the underlying discrimination is removed. If the algorithm is allowed to modify the training data, then pre-processing can be used.
- **In-processing.** In-processing techniques try to modify and change state-of-the-art learning algorithms in order to remove discrimination during the model training process. If it is allowed to change the learning procedure for a machine learning model, then in-processing can be used during the training of a model—either by incorporating changes into the objective function or imposing a constraint.
- **Post-processing.** Post-processing is performed after training by accessing a holdout set which was not involved during the training of the model. If the algorithm can only treat the learned model as a black box without any ability to modify the training data or learning algorithm, then only post-processing can be used in which the labels assigned by the black-box model initially get reassigned based on a function during the post-processing phase.

4.4 Model-based Framework Supporting Trustworthiness in AI and Data

The potential of AI promotes wide application and integration into many business domains. Gartner even predicts an increasing use of Generative AI to be more than 50% by 2027 (Chandrasekaran, 2024)¹⁴⁹. At the same time, companies need to comply with the increasing challenges on trustworthiness aspects especially for problems including system bias, lack of explainability, ethical and privacy issues as well as emerging regulations and legislations e.g. AI Act requires AI to be lawful, ethical and robust throughout the lifecycle of AI systems (Li et al., 2023)¹⁵⁰. The main dimensions to ensure trustworthiness include human agency and oversight, technical robustness and reliability, privacy and data, transparency, algorithmic fairness, societal well-being and accountability (AI HLEG, 2019)¹⁵¹. Especially, GenAI and AI based on LLMs face several disadvantages when applying it in critical applications with regards to a lack of transparency and explainability. As it has no “semantic” understanding and only imitate “understanding” it is biased and hallucinates (Szczuko, 2024)¹⁵². Additionally, LLMs' current reliability, explainability, and robustness might conflict with legal regulations. Therefore, presentations on “Hybrid AI” as by Akhai (2023)¹⁵³ and publications by Mattioli et al. (2022)¹⁵⁴ or Bork et al. (2023)¹⁵⁵, and XAI in combination with conceptual modelling examine the combination of symbolic and sub-symbolic AI approaches with the idea of using well-established symbolic approaches to improve the shortcomings of GenAI and AI based on LLMs. While symbolic AI is transparent, explainable, accountable, reliable, and deterministic; sub-symbolic AI is flexible, abstract, data-driven, and capable of extracting implicit dependencies from data. Conceptual models are seen as symbolic AI, especially if they are used to represent e.g. rules, fuzzy rules, workflows or semantic models.

Thus, besides the initially described AI verification tools in Section 4.2.1 such as LIME or SHAP for transparency and explainability or Aequitas for fairness, a model-based framework to support the assessment and facilitation of robustness or reliability of AI and data, is introduced in this section. One example of supporting AI applications for FLEX services using LLMs to structure information and generate business process models is given. Detailed descriptions of the models used for other services are given in Deliverable 4.3.

The model-based framework, discussed in this section, aims to utilize models to enhance trustworthiness and its underlying aspects e.g. transparency, reliability, explainability, accountability, etc., of AI applications especially data-driven AI such as LLMs. The framework is illustrated in Figure 16.

Conceptual models as discussed in this section and the research track introduced in Section 3.7. are used to represent knowledge in a way that is understandable for humans and machines. The framework, which is introduced below, contains three categories on how conceptual models can be used to support AI and data trustworthiness. Different modelling languages, where semantics are defined in their metamodels, can be used to describe knowledge, which is used to cover different aspects of trustworthiness. For example, trustworthiness can be supported by using the models as input for the decision services, which suggest solutions based on the model knowledge. Here the models themselves can also be used to explain how the decision is made. Or models can be enriched with information on who is responsible for a decision, defining accountability explicitly.

To increase the understandability of the models, well-known modelling languages like BPMN or Archimate can be used. Having a modelling language is not enough, as they must be properly used to increase the value of the models (Karagiannis & Kühn, 2002)¹⁵⁶. Additionally, the semantics that models can represent and understanding depends on how knowledgeable the users are with the modelling language. The interpretability can be improved for domain experts by tailoring the modelling languages and the concepts used to the domain (Karagiannis, 2015)¹⁵⁷. But not only the available concepts must be tailored to the domain, but also the offered functionality must be adapted to meet the needs of the users. Therefore, within FAIRWork we not only used established modelling

languages, but also applied a model-based Design Methodology, described in our Deliverable 2.1, which uses different modelling languages on different abstraction levels to not only describe but also to understand. Therefore, the procedure is started with physical workshops, where high-level physical models are created, which are automatically transformed into digital models, capturing a common understanding of the scenarios that should be implemented in processable way. These models use tailored and semantic-rich representations, facilitating a common understanding by people with various backgrounds. They help to better understand the scenarios and identify problems early. This understanding also supports the creation of domain models of qualified content, as described in Figure 16.

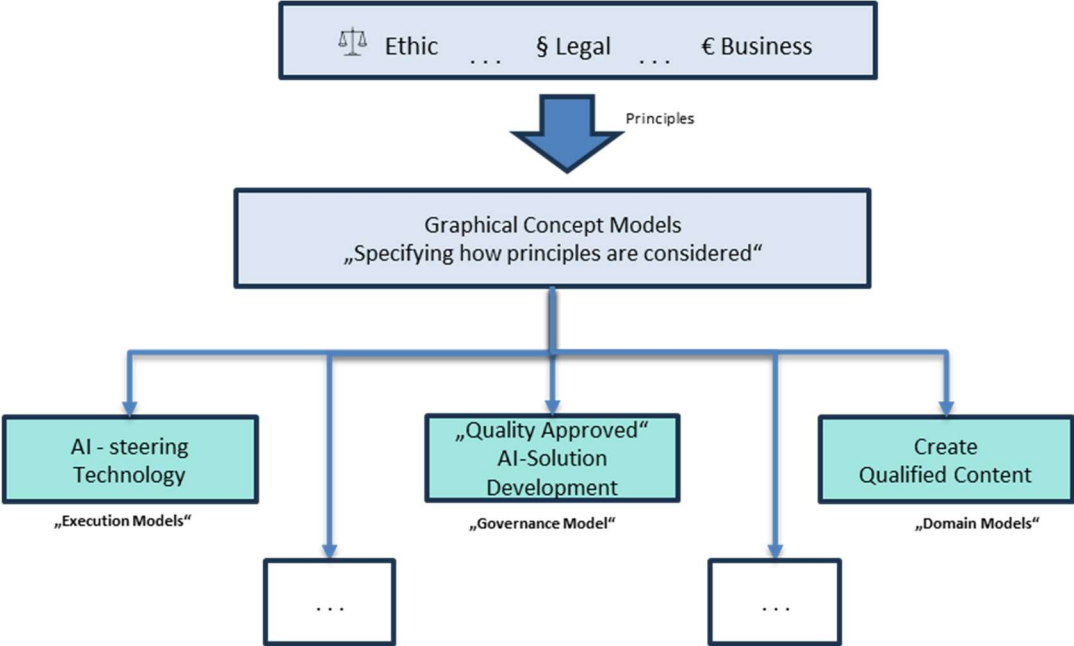


Figure 16: Model-based framework to support AI and data trustworthiness.

For businesses different ethical, legal or business requirements applicable principles emerge depending on the particular use case of AI. For example, the type of AI or its application area like medicine or financial domains determine the level of risk of AI demanding the setup of ethic, legal and technical-specific compliance measures and principles. These compliance aspects can be illustrated with domain-specific concept models and can be approved through different experts like ethical, legal, technical or domain experts. As an example, we introduced a service for certifying models through experts, increasing the trust for the modelled situation. This service was introduced and described in Deliverable 3.2 under the ethical watchdog section. The graphical models aim to support explainability and transparency while the humans in the loop support trust and reliability.

Based on the defined principles different ways to integrate legal, ethical or technical aspects into AI or into governance processes this also includes increasing trustworthiness and reliability of AI are proposed. Three main categories are identified. First, through **AI steering technologies**, second through **quality approved AI-solution development** and third through **creation of qualified content and data**. For the categories different models ranging from strict formal like workflow or rule engines to semi-formal representation such as business process or decision models can be used.

The first category of AI steering technologies covers “execution models” to configure AI solutions during runtime. These can include a combination of goal and rule models to influence agentic based services, but also steering mechanisms in form of workflows, rules or knowledge graphs that specify the execution of AI-solutions.

The second category specifies quality approved AI solution development referring to “governance Models” specifying the development process, used tools required validation steps during implementation and any assessment models. Also, governance processes for AI configuration and building can be captured.

Qualified content and data can be supported by “domain models” referring to BPMN, enterprise architecture models (ArchiMate), or governance, risk and compliance models. These can, for example depict certain business processes, goal and measures or context information. The models can be certified and digitally signed by experts and be used as verified data.

As an example of enhancing the transparency and reliability of AI applications, especially LLMs through AI steering technologies, one approach for the FLEX “document to process model” prototype is introduced. The prototype uses LLM and in particular gpt-4o to generate business process models from word documents. When uploading the document and by pressing the “submit” button, different components and services of the application are triggered. For example, the first step of the workflow is the component “**extract text**”. For this a service is used to interpret different types of documents like docx, pdf, txt, etc. and extracts the textual content and the contained images. The second step of the workflow is “**extract BPMN**” which is a prompt for interpreting text and generating a JSON. There are multiple steps needed to guide the AI to the final result. For each of the steps certain prompts and requirements are defined that the AI-solution must consider or comply with. With this workflow, the LLM based AI in combination with agentic workflows aims to be easier explainable, more transparent and reliable. More technical details about the prototypes can be found in Deliverable 4.3

One part of the research in supporting the explainability through conceptual models is which functionality can support understanding the models and the represented scenarios. For example, how data from already made decisions can be represented within the models to not only show abstract scenarios but concrete ones by enriching them with data, how finding input for defining important aspects of the scenarios can be supported or how models can be animated to visualize dynamic aspects concretely. This enrichment enhances the quality content domain models from the corresponding category, introduced above, as the enriched models are available together with the domain models, supporting a better understanding of the models themselves.

To improve the digital representation and therefore supporting the understanding of the of the scenario, it was researched how the automatic transformation from the physical to the digital models, the adaption of the domain concepts and the tailoring of the functionality of the used modelling tools and the created models can be improved to improve the comprehensibility of the models. This was done by creating prototypes with the ADOxx-based modelling tools, Scene2Model and Bee-Up.

4.5 Ethical Watchdog

The watchdog agent promotes fairness within DSS by continuously monitoring and assessing the system's processes and decisions to ensure they are equitable and adhere to ethical standards. Its primary function is to safeguard against the disproportionate disadvantaged of particular groups, upholding the principle of equity and aligning the system's operations with established ethical guidelines.

A key attribute of the watchdog agent is its ability to enhance accountability, a cornerstone of ethical system design. Accountability is achieved by flagging deviations from ethical standards and ensuring that these deviations are promptly addressed. The watchdog agent prevents the system from operating as a complete "black box," where decisions are made without clarity or recourse for affected individuals, holding the system responsible for adhering to predefined criteria and metrics aligned with ethical principles (Guidotti et al., 2018)¹⁵⁸. The incorporation of fairness metrics, such as demographic parity, further strengthens the agent's role in objectively evaluating outcomes and enforcing predefined thresholds of fairness. This ensures that the system consistently delivers equitable results across diverse groups, fostering trust in its operations in a human-centric approach (Mehrabi et al., 2021¹⁵⁹; Binns, 2021¹⁶⁰; Florridi & Taddeo, 2016¹⁶¹).

Furthermore, the implementation of ethical standards minimizes the risk of the system engaging in harmful or unethical practices. The watchdog agent translates complex algorithmic operations into accessible outputs, empowering decision-makers to make informed choices that also consider ethical dimensions. This informative role not only enhances the system's transparency but also ensures that ethical guidelines are integrated into the final decision-making process.

The watchdog agent's ability to promote both transparency and accountability is fundamental to ensuring that the DSS operates with integrity and trust. Without these attributes, the system risks eroding stakeholder confidence, perpetuating biases, and failing to meet its intended ethical and functional objectives. The watchdog functions as an indispensable component of DSS ensuring fairness by addressing biases, promoting transparency, enabling dynamic oversight, and fostering accountability. Its role not only improves the system's ethical performance but also builds trust among stakeholders by demonstrating a commitment to equity and justice in decision-making processes.

The "Ethical Watchdog" function demonstrates its particular cleverness by being conceptualized as a stand-alone technical tool. The possibilities it offers for identifying inconsistencies in ethically relevant objectives are fascinating. Such inconsistencies represent the starting point for further ethical reflection, but cannot serve as its result. Therefore, we want to emphasize at this point that there are also limits to the automation of this function. For it is precisely the special ambivalences and problems of forming judgments that require social rather than purely cognitive intelligence. Against this background, it is true for the "ethical watchdog" function, as it is ultimately for the forms of democratization of companies through digital tools discussed below, that social processes must be implemented to ensure that it actually works.

5 THE DEMOCRACY QUESTION

The great potential of digital technologies to reconstitute socio-technical-economic orders has, from the outset, triggered a discourse about the associated futures. Two more or less polarizing perspectives have emerged. On the one hand, a discourse of development opportunities that arise from these technologies for the continuation of democratic practices of public-political communication and decision-making. The opportunities for more equal participation, greater inclusion and stronger representation are obvious. At more or less the same time, however, a discourse has emerged about the side effects of these technologies, which, in contrast to the first discourse, focuses on the specific side effects of these technologies and, in particular, their de-democratizing tendencies. Moreover, with regard to recent developments of re-adjusting principles of the use of such technologies more or less under economic reasons, the quest is getting more and more crucial, whether and how in the realm of AI, democratizing decision-making entails promoting democratic practices throughout the development, implementation, and utilization of technologies – or, to the contrary, leads to the opposite dynamic. From a viewpoint of political systems, democracy is the most ambitious form to perform power by, for and with the people. Typically, hierarchical organizations, like companies, are not the place of democratic decision-making. On the contrary, companies as formal organizations are characterized by the fact that they do not make decisions in a participatory way, but rather in a consistently asymmetrical way, through a clear hierarchical structure.

In this context, the question of democratization in companies seems paradoxical. Especially since current developments tend to limit the institutionalized forms of co-determination in companies (via trade unions). A closer look, however, shows that this is not as paradox as expected, for at least two reasons. On the one hand, the innovative action of companies is increasingly developing in the direction of open innovation. This enables new coordination services by opening up to other companies. In this way, innovation processes can be put on a more solid ground. On the other hand, there are ideas about the democratization of innovations, according to which the users of a technology are increasingly involved in the process of creating this product and keep pace with it. These two dynamics reveal that specific opening processes are very much underway in companies. In the process, new actors, typically external actors, come into play. When it comes to the question of democratization, a dynamic of democratization “outwards” and one “inwards” can be identified. These can be interlinked, but they do not have to be. Nonetheless, this indicates a cultural change that is significant with regard to the question of democratization of companies with and through digital tools.

5.1 Democracy in Companies

Processes of democratization in companies face the particular challenge that forms of democratization on the one hand and the hierarchically structured command structure on the other tend to exclude each other. Therefore, a keen eye is needed for the forms and scope of processes of internal democratization of companies - with or without digital tools. Ultimately, formal democratization in companies can be described as a de-concentration of hierarchical power, which simultaneously increases the functionality and legitimacy of workflows within companies by following a more participatory decision-making model and thereby absorbing the de-concentration at the top with a participatory concentration of power at the lower levels of the organization. With regard to our task here, democratic decision-making can be seen as an effort to involve possibly most of the individuals within a group in the decision-making process and to prevent illegitimate centralization and concentration of power in this process. This process necessitates an analytical approach that considers both social and technical factors. Its aim is to ensure that AI technologies contribute to enhanced democratic decision-making processes. For achieving this, it is of utmost importance to explore specified methods for democratic control over technologies implemented for decision-making

support as well as to understand how they can foster democratic practices in companies in general (Noorman & Swierstra, 2023)¹⁶². Since such processes are obviously characterized by a considerable fragility, this cannot be built solely on the hope that the technical systems (in this case, the support with a MAS) will sufficiently structure the processes. Rather, the situation is such that the introduction of such MAS changes the socio-technical configuration of the organization. In order for this to be recognized by the people involved as a legitimate change in the organizational structure, appropriate legitimacy resources are required in the corporate culture, as well as corresponding institutional-procedural safeguards within the operational organization of the company.

To achieve this goal, FAIRWork project has designed and implemented the DAI-DSS, integrating various technologies to support decision-makers (Woitsch et al., 2023)¹⁶³. Therefore, the democratization of decision-making within socio-technical contexts is specifically focusing on Democratic Decision-Making with MAS. Thereby, questions of participation, representation or transparency are decisive. Thus, stakeholders and managers in companies are also striving for fair decision-making models (Charles et al., 2021¹⁶⁴; Hilton et al., 2021¹⁶⁵; Dingwerth et al., 2020¹⁶⁶) Nowadays, the use of new technologies to enhance democratic decision-making models opens up new possibilities in this context, demanding additional examination. One important aspect hereby consists in the fact, that the democratization of decision-making in socio-technical settings can imply an integration of essential democratic features in the implementation process. This ensures that the implementation of the DSS not only enhances efficiency but also improves the company and empowers employees to engage in meaningful participation. One of our key results in conducting the case studies as mentioned above (Section 3.2), consists in identifying additional factors which have been regarded as democratic features in the respective cases (Figure 17).

Besides the primary features, Learning Implementation is recognized as a key factor that fosters democracy because it enables a long-term, comprehensive learning period from scratch, in which all employees are actively involved. Such implementation is less likely to be denied or forgotten, and any limitations of the digital tool can be identified quickly and resolved. Customizability also enables last-minute adjustments for customers and decision-makers, giving them more control over how the system operates and enhancing the system's flexibility. Clear responsibilities (or rather accountabilities) are always a relevant factor in the development of organizations. This is even more the case when organizations are undergoing a more or less fundamental change, like in the case of such an implementation of a digital technology for supporting decision-making. Additionally, accountability is an important democratic feature because it fosters transparency and fairness. This allows individuals to track the decisions made by the system and ensures that decisions can be changed or improved if necessary.

Consequently, these features empower employees by providing tools and opportunities for active engagement in decision-making, ensuring their voices are heard and fostering a more inclusive, collaborative environment that promotes collective decision-making. At the same time, the dimensions of a responsible AI shown in Figure 17 make it clear that a multitude of relevant criteria must be taken into account in processes of democratization. However, the criteria do not behave consistently in a linear fashion (in the sense that achieving one would simultaneously imply improving the other criterion). Rather, the situation is much more complicated. In some cases, these criteria are competing to each other. Therefore, two strategies are important. Firstly, these criteria must be taken into account in the MAS and the procedure must be supported. Secondly, the social quality of the implementation processes of such tools plays a central role. Otherwise, this process would come into sharp contradiction to the articulated relevant criteria for the evaluation of democratization.

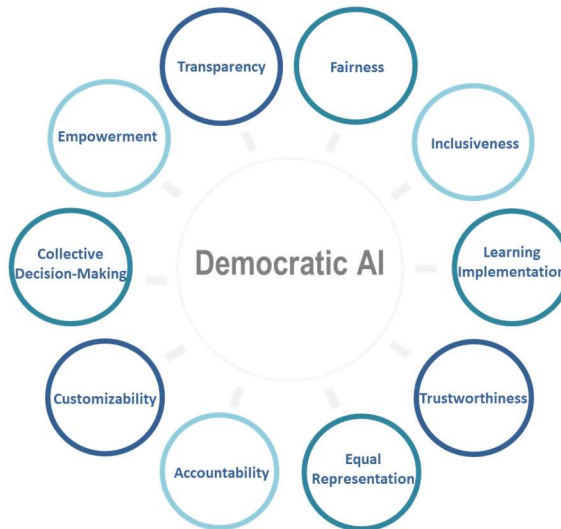


Figure 17: Features of democratic AI in companies.

The criteria mentioned above ultimately identify the central objectives of a model for democratic decision-making in organizations. Nevertheless, the challenge arises as to how these criteria can be operationalized in concrete terms. This form of operationalization is a highly context-dependent activity. It depends on the specific context of application of the respective company. Against this background, a corresponding workshop was held to further operationalize the criteria for democratizing companies through digital tools (see Figure 18).

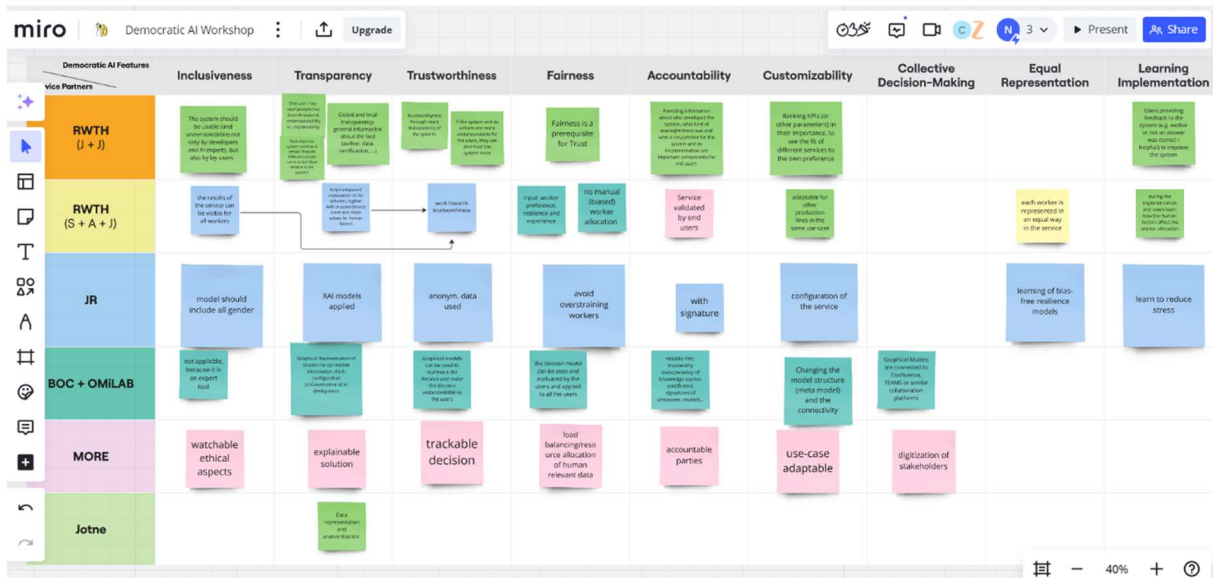


Figure 18: First task in the democratic AI workshop.

5.1.1 Worker Expectations Reflecting Features of Democratic AI

Following the various challenges faced by workers mentioned in Section 3, they also shared their expectations reflected key features of E-Democracy (Council of Europe, 2020)¹⁶⁷, such as Fairness, Transparency, and Trust. During the focus group interview, workers expressed their desire for **Fairness** in the workplace. They emphasized the importance of fair recognition and appreciation, highlighting the value of personal engagement from their superiors, suggesting that even brief check-ins could strengthen their sense of appreciation. Another point highlighted by workers across all hierarchical levels was the importance of fostering fair treatment. They discussed the challenges of ensuring fairness in the workplace, recognizing that varying levels of effort and competencies might require different approaches to treatment. The demands included a focus on improving PTO management, particularly for employees with frequent absences. The interviewees expressed interest in a digital tool that could support more efficient and reliable PTO tracking, helping to streamline planning. Ultimately, employees emphasized the importance of fair and equitable worker allocation, expressing the expectation that assignments should be transparent and well-balanced.

Expectations regarding **Transparency** may not be numerous, but they are significant in quality. The employees from the top-floor level believed that for a digital tool to be successfully implemented in the company, its tangible advantages must be transparent from day one. Apart from the DSS, experienced employees who have been part of the workforce for a long time emphasized the importance of timely and clear communication about changes with workers. They noted that workers expect to be informed about every restructuring as soon as possible, even if it doesn't directly relate to them: *“One of the most important aspects is, of course, communication with employees. How clear is it? How iterative is it? How quickly does it reach them? Does it arrive on time or late? Do others already know more? The earlier, the better—it’s always beneficial to involve every employee and their families as early as possible during changes. That’s obvious, and it’s noticeable that employees are quite demanding in this regard, even if they say they would like it. Currently, there’s a bit of a time-related issue, where it’s clear that employees need to be consistently informed, even if the matter doesn’t immediately affect them—because, in the end, it affects everyone.”*

In order to **Trust** the digital tool, workers needed at least 70% alignment between their preferences, opinions, and choices with the tool's suggestions. They argued that 30% differences are not problematic because the tool cannot account for some personal matters. The workers also emphasized their own effective role, based on their experience and awareness. Considering the ethical concerns regarding AI (Dhruvitkumar et al., 2024)¹⁶⁸, the workers were reluctant to expose private data with the digital tool in the workplace. However, they believed that over time, further exploration could build trust in AI.

Expanding on the operationalization of findings through the workshop process, we asked the service partners in the project to identify which expectations mentioned in both the Onboarding and Precession phases are applicable to their services. As shown in Figure 19, simplifying system usage, empowering staff for decision-making, reducing human intervention, and fostering continuous learning during tool deployment were the most frequently mentioned. Collaborative decision-making with AI, indicating resource availability and locations, ensuring fair treatment of employees, balancing team contributions, and enabling data and workflow review were also highlighted.

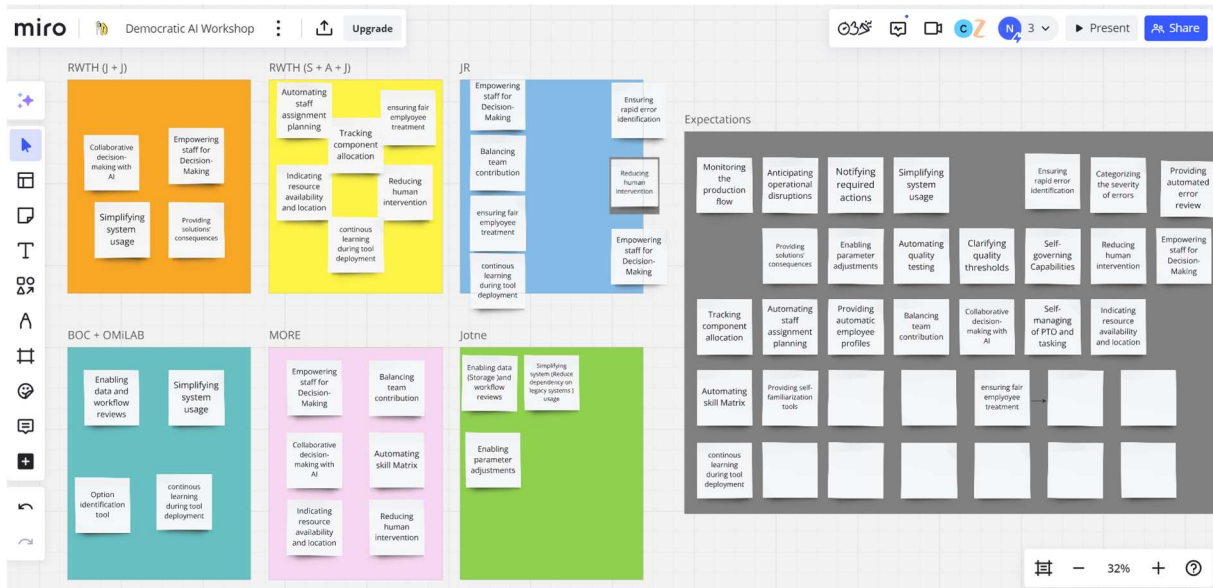


Figure 19: Second task in the democratic AI workshop.

5.2 The Question of Representation

In our particular case of DAI-DSS, the question of representation has a twofold quality. On the one hand, we have the participatory aspect of the representation of specific individual concerns in the process of collective decision-making about the introduction of such a tool. On the other hand, the implementation of such a tool is precisely about the representation of workers and their specific claims in the technical object of the MAS. From this perspective, representation is the key issue in the question of processes of democratization in companies. This is because it ensures that, in the event of undesirable developments, options for adaptation exist for the various groups of actors in the organization. Due to the aforementioned complexity and inherent complexity of the question of representation, three lines of argumentation will be pursued in the following. Firstly, we will further deepen and conclude the conceptual considerations that we had already begun in Deliverable 3.2. Secondly, we will underline and thus support these considerations with corresponding empirical results from the case studies. Thirdly, it is to be shown that the question of legitimacy can only be answered while implementing social procedures of embedding DAI-DSS.

5.2.1 Conceptual Questions

In the conceptual structure, we argued that there are basically three crucial elements. Firstly, the MAS functions as a form of indirect representation. It represents a kind of mediation instance between workers and the respective production situations in which concrete allocation decisions have to be made. Secondly, this indirect representation is to be thought of and developed as a layered model of different options and ranges of representation. This is because appropriately constituted processes of mediation are needed to maintain the ability to make decisions in the organization. Thirdly, these layers can be used to elaborate specific functionalities of the relation between technical representation in the MAS and the requirements of workers. In this sense, the following concept development can be understood as the realization of precisely such a “nested representation” (Figure 20).

This structure enables the development of democratic decision-making in the tension between wishes for representation and the form of being represented. This starts from the **zero level** of problem recognition by workers and continuing to the fifth level of DAI-DSS. Team members are typically the first who identify problems within their own division and consequently seek collectively to find possible solution. The extent of priority, comprehensiveness, and importance of the issue determines the solutions suggested by the teams. Moving to the **second level**, representatives convey the identified issues and proposed solutions to experts, engaging in negotiation to explore potential solutions. The **third level** provides an opportunity for team members to select preferred solutions from suggestions by representatives and experts. The suggested solutions have one additional step before embedding in MAS, where worker's representatives and experts play a significant role; At the **fourth level**, representatives and experts negotiate the most chosen solutions and decide which ones have the capability to coordinate with the system's goal for MAS implementation. This interaction not only assesses the precision of the solution but also enhances decision legitimacy, representation, transparency, and trust within the company. All relevant information and parameters regarding the process including the input coming from the people is modelled into the MAS aligned with the goals of the specific process. A negotiation dynamic between the agents is then designed in direction to the system's goals within the boundaries and constraints natural to each type of process deriving from the kind of task or activity developed in each scenario. The agents' interactions play a decisive role in balancing different kinds of parameters among the existent stakeholders. These digital representations of human stakeholders and their interactions following well-established goals while considering fairness aspects in the workplace allows to a democratization of decision-making.

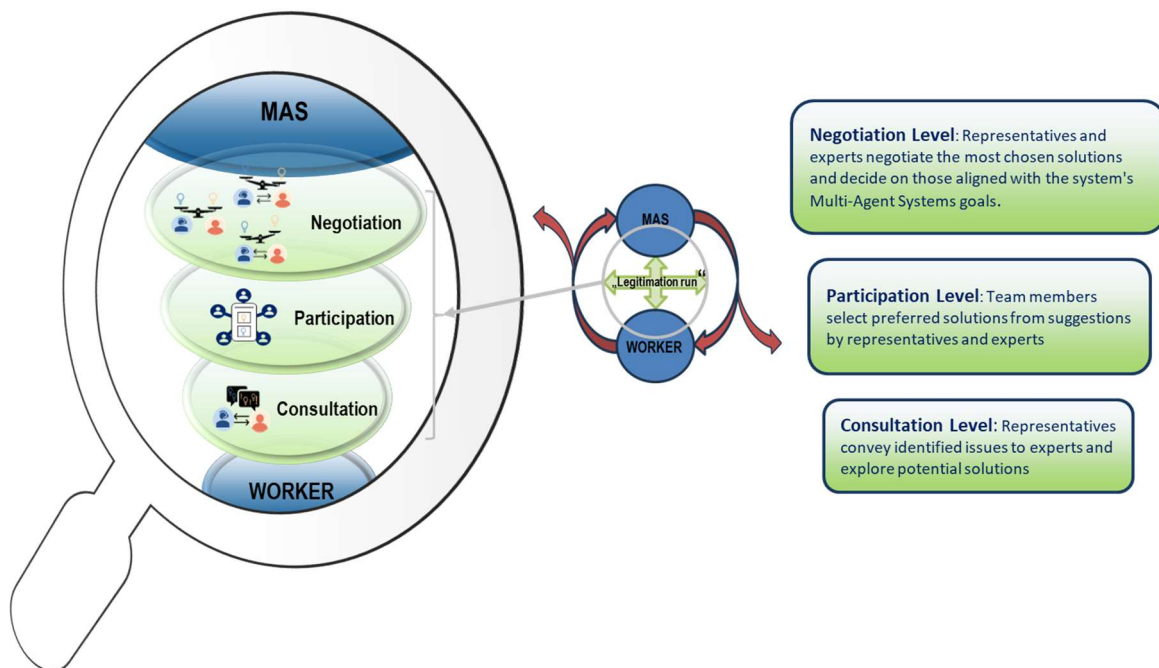


Figure 20: Levels of decision-making processes with MAS giving a socio-technical structure to DAI-DSS.

A “nested representation” like this also impressively demonstrates the socio-technical complexity of such demands for democratization of decision-making processes in companies. Although these respective levels, which must be related to each other here, can be technically mapped, they cannot be exclusively technically realized. Rather, a

socio-technical arrangement is needed in which the necessary coordination processes are always anchored institutionally and procedurally in the organization of operations in the company. This conceptual framework can also be found as an innovation item in Section 2.

5.2.2 Empirical Insights

As mentioned in the previous sections, to democratize decision-making through the implementation of MASs in companies, we conducted a case study consisting of onboarding and precision phases. By utilizing a comprehensive methodology including documentary analysis, observations, individual interviews, and a focus group interview we gained valuable insights into the current decision-making processes, the level of employee engagement in our selected use case, the functionalities employees expect from a DSS, and the challenges workers currently face within the company.

The findings underscore the need for the development of a robust DSS to initially offer production-line responsible persons a diverse array of solutions, thereby facilitating informed decision-making. The implementation of such a system, coupled with the delegation of tasks and responsibilities to the system, shows potential for alleviating existing workloads and time pressures, ultimately fostering employee productivity and innovation across all organizational levels. With regard to the exploration of employee involvement, we found that the dominant high level of trust among team members has led to a decreased tendency to raise objections or offer suggestions. Despite the presence of a strong level of trust within the team, such trust may be weakened in instances where a mismatch arises between decisions reached collectively and their subsequent implementation within the company. However, with the implementation of a DSS, such problems can be addressed, as they may arise from the involvement of several decision-makers, each with potentially conflicting perspectives and suggestions. In addition, the analysis of outcomes within the involvement dimension indicates that decisions made at the shop floor level are typically characterized by a higher degree of transparency, often resulting from regular meetings, active inquiry, and the inclusive participation of all relevant staff members. Conversely, decisions formulated at the upper management level are less frequently perceived as transparent to the entire staff. Furthermore, notably, employees view the DSS as a supplementary tool, rather than a replacement for human decision-making.

Addressing the challenges in the second step of the research reveals a significant area for discussion. In addition to the issues identified in the individual interviews, the focus group interview provided shop-floor workers with an opportunity to share their core concerns within the workflow. Notably, discussions primarily focused on personal dynamics rather than work-related concerns. Workers highlighted the importance of balanced team contributions, equitable work assignments, fair recognition and appreciation, and fostering an inclusive and respectful work environment.

However, there is a tension emerging. The tension between the needs and ambitions of the ones being represented in the decision support tool and the forms and logics of representing within the tool. This tension can be seen as the litmus test of the development of these kinds of tools. If there are no further options to get an insight or to intervene into the ways of being represented, the tool's legitimacy is at risk. In alignment to this, the workers express a firm belief that such a system should not take away their autonomy in decision-making but should instead enhance their ability to make informed, efficient, and simplified decisions. The consensus among employees is that the DSS should act as a guiding mechanism or an option identification tool, providing insights and recommendations to

facilitate improved decision-making processes, while the ultimate responsibility for decision-making remains with human actors.

Building on this perspective, we can conclude that forms of democratization via MAS can only be achieved if key democratic features are incorporated into the development and implementation process. This ensures that the use of the AI tool not only increases efficiency but also improves the company and empowers employees to engage in meaningful participation. According to our findings, these democratic features are transparency, fairness, and representation. This implementation process leads to continuous stabilization and fosters legitimacy.

5.3 Legitimacy via Social Embedding and Procedural Implementation of AI Tools

The question of the democratization of companies through digital tools, as the previous considerations and findings should show, represents a process that depends on a variety of special conditions. The particular potential of AI solutions (in our case an MAS) for such processes of democratization in companies clearly stands out. This is because not only the obvious claims can be realized, but the range of criteria for democracy-building qualities is considerably expanded. This finding alone points to the opportunity that such technologies offer for a democratically transformed socio-technical order. In particular, the opportunity to increase representativeness stands out in its significance. This is because the tool can be used to uncover a whole range of layers for democratization in the process organization of companies. However, this circumstance does not per se mean that the democratizing potential will actually be realized. This depends on further conditions.

Among other things, the fact that this potential for democracy-building also has a number of possible side effects and problems in terms of restricting democratizing dynamics must be noted. The introduction of such a tool can, quite contrary to the democratizing tendencies, also be used to strengthen hierarchical governance in companies by creating a more refined form of command structure. This option cannot be ruled out from the outset. Rather, this possible outcome of the development points to a very crucial circumstance in the introduction and implementation of such tools. These can only contribute to democratization in companies to the extent that they themselves have been democratically introduced. Thus, the legitimacy of the use of such tools in companies is not only due to their convincing functionality, but rather to the courage to make the introduction itself a democratizing event within the company.

6 SUMMARY AND CONCLUSIONS

This deliverable consists of four main parts. The first part, Section 2 is dedicated to the FAIRWork innovation shop as a key result of our work. Not only does it provide essential tools and heuristics for implementing DAI-DSS in companies. Moreover, this shop provides an overview of the central tasks involved in such an implementation. At the same time, the modular structure of this shop ensures that the respective results are always kept up to date and that items can be easily added if this proves to be helpful and relevant in the light of new findings.

The second part, Section 3, focuses on the detailed outline of the research services, methods and studies that make up a research collection. It also shows the application of sensors to capture critical information about the mental, affective, and motivational state of humans. Furthermore, it introduces a novel framework using Personas for Human Digital Twins in decision-making. The third part, Section 4, provides an overview of the different research strands which have been dedicated to the key questions of explainability and fairness in AI services, enfolded the relevant conceptual as well as empirical work and ending up in the quest of how to think and establish an “Ethical Watchdog” as tool and procedure. The fourth part, Section 5, specifically focused on the multifaceted questions of democratization at company’s workplaces via AI technologies. Especially, the representation problem was elaborated as cornerstone question in this regard.

The identification of key research factors within industrial use cases further strengthens the practical implications of future studies. By analysing these factors from both human and technical perspectives, the report offers valuable insights that can guide developers and practitioners in optimising their decision support systems. This comprehensive understanding of the challenges and requirements in real-world scenarios ease the development of tailored solutions that address demanding manufacturing needs. This report provides with relevant concepts and tools for conducting further studies of this type to increase our knowledge about a broad range of application scenarios.

Ultimately, the collective efforts of examining literature, employing research methodologies, identifying key research factors, and implementing an effective communication strategy contribute to the broader goal of advancing decision-making processes and facilitating the successful adoption of AI and MAS technologies in decision support systems.

7 ANNEX A: LIST OF ABBREVIATIONS

Abbreviation	Meaning
AI	Artificial Intelligence
BPMN	Business Process Model and Notation
CM	Conceptual Modelling
CMAI	Conceptual Modelling with AI
CP	Constraint Programming
CPS	Cyber-Physical Systems
CSV	Comma-Separated Values
DAI-DSS	Democratised AI-Decision Support System
DORA	Digital Operational Resilience Act
DHS	Digital Human Sensor
DMN	Decision Model and Notation
DSS	Decision Support System
ER	Entity-Relationship
GDPR	General Data Protection Regulation
GenAI	Generative AI
GUI	Graphical User Interface
HR	Heart Rate
HRV	Heart Rate Variability
HTTP	Hypertext Transfer Protocol
ISB	Intelligent Sensor Box
ISO	International Standards Organization
IT	Information Technology
JSP	Job Shop Problem
JSON	JavaScript Object Notation
LIME	Local Interpretable Model-agnostic Explanations
LLM	Large Language Model
MAS	Multi-Agent System
ML	Machine Learning
NFO	Non-Functional Overreaching
NLP	Natural Language Processing
OCR	Optical Character Recognition
PSS	Perceived Stress Scale
PTO	Paid Time Off
RAG	Retrieval-Augmented Generation
REM	Rapid Eye Movement
REST	Representational State Transfer
RESTQ	Recovery-Stress Questionnaire
RL	Reinforcement Learning
RRSM	Resilience Risk Stratification Model
SHAP	Shapley Additive Explanations
TRL	Technology Readiness Levels
UI	User Interface
URL	Uniform Resource Locator

UX	User Experience
WP	Work Package
XAI	Explainable AI
XML	Explainable Machine Learning

Table 3: List of abbreviations.

8 REFERENCES

- ¹ Noorman, M., & Swierstra, T. (2023). Democratizing AI from a sociotechnical perspective. *Minds and Machines*, 33, 563–586. <https://doi.org/10.1007/s11023-023-09651-z>
- ² Woitsch, R., Muck, C., Utz, W., & Zeiner, H. (2023). Towards a democratic AI-based decision support system to improve decision making in complex ecosystems. *BIR Workshops*. 209-223.
- ³ Gheibi, N., & Boesch, S. (2024). Democratization in Industry via Multi-Agent Systems, The case of a production company. In: Lucas Paletta (eds) *Cognitive Computing and Internet of Things. AHFE Open Access*, 124. AHFE International, USA. <http://doi.org/10.54941/ahfe1004709>
- ⁴ Trentesaux, D., Caillaud, E., & Rault, R. (2022). A Framework Fostering the Consideration of Ethics During the Design of Industrial Cyber-Physical Systems. In: Borangiu, T., Trentesaux, D., Leitão, P., Cardin, O., Joblot, L. (eds) *Service Oriented, Holonic and Multi-agent Manufacturing Systems for Industry of the Future. Studies in Computational Intelligence*, 1034. Springer, Cham. https://doi.org/10.1007/978-3-030-99108-1_25
- ⁵ Cervantes, J.A., López, S., Rodríguez, L.F., Cervantes, S., Cervantes, F., & Ramos, F. Artificial Moral Agents: A Survey of the Current Status. *Sci Eng Ethics* 26, 501–532 (2020). <https://doi.org/10.1007/s11948-019-00151-x>
- ⁶ Răileanu, S., & Borangiu, T. (2023). A Review of Multi-agent Systems Used in Industrial Applications. In: Borangiu, T., Trentesaux, D., Leitão, P. (eds) *Service Oriented, Holonic and Multi-Agent Manufacturing Systems for Industry of the Future. Studies in Computational Intelligence*, 1083. Springer, Cham. https://doi.org/10.1007/978-3-031-24291-5_1
- ⁷ Longo, F., Padovano, A., & Umbrello, S. (2020). Value-Oriented and Ethical Technology Engineering in Industry 5.0: A Human-Centric Perspective for the Design of the Factory of the Future. *Applied Sciences*, 10(12), 4182. <https://doi.org/10.3390/app10124182>
- ⁸ Gal K., & Grosz, B.J. (2022). Multi-Agent Systems: Technical & Ethical Challenges of Functioning in a Mixed Group. *Daedalus*, 151 (2). 114–126. https://doi.org/10.1162/daed_a_01904
- ⁹ Pacaux-Lemoine, M.P., & Trentesaux, D. (2019). Ethical risks of human-machine symbiosis in Industry 4.0: Insights from the human-machine cooperation approach. *IFAC-PapersOnLine*, 52 (19), 19-24, <https://doi.org/10.1016/j.ifacol.2019.12.077>
- ¹⁰ Gal, K., & Grosz, B.J. (2022). Multi-Agent Systems: Technical & Ethical Challenges of Functioning in a Mixed Group. *Daedalus*, 151(2), 114–126. https://doi.org/10.1162/daed_a_01904
- ¹¹ Belloni, A., Berger, A., Boissier, O., Bonnet, G., Bourgne, G., Chardel, P.-A., Cotton, J.-P., Evreux, N., Ganascia, J.-G., Jaillon, P., Mermet, B., Picard, G., Rever, B., Simon, G., Swarte, d.T., Tessier, C., Vexler, F., Voyer, R., Zimmermann, A. (2015). Dealing with ethical conflicts in autonomous agents and multi-agent systems. *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*.
- ¹² Woodgate, J., & Ajmeri, N. (2022). Macro Ethics for Governing Equitable Sociotechnical Systems. *Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems*, 1824–1828. IFAAMAS Press. <https://dl.acm.org/doi/10.5555/3535850.3536118>
- ¹³ European Commission: Directorate-General for Research and Innovation, Breque, M., De Nul, L. & Petridis, A., *Industry 5.0 – Towards a sustainable, human-centric and resilient European industry*, Publications Office of the European Union, 2021, <https://data.europa.eu/doi/10.2777/308407>
- ¹⁴ Paletta, L., Zeiner, H., Schneeberger, M., & Quadri, Y. (2023). Digital Shadows and Twins for Human Experts and Data-Driven Services in a Framework of Democratic AI-based Decision Support. In: Lucas Paletta, Hasan Ayaz and Umer Asgher (eds) *Cognitive Computing and Internet of Things. AHFE Open Access*, 73. <http://doi.org/10.54941/ahfe1003971>
- ¹⁵ Southwick, S.M., Bonanno, G.A., Masten, A.S., Panter-Brick, C., & Yehuda, R. (2014) Resilience definitions, theory, and challenges: interdisciplinary perspectives. *European Journal of Psychotraumatology*. 5. PMID: 25317257; PMCID: PMC4185134. <https://doi.org/10.3402/ejpt.v5.25338>
- ¹⁶ Smith, B.W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard J. (2008). The brief resilience scale: assessing the ability to bounce back. *International Journal of Behavioural Medicine*, 15(3), 194-200. <http://doi.org/10.1080/10705500802222972>

- ¹⁷ Schetter, C.D., & Dolbier, C. (2011). Resilience in the Context of Chronic Stress and Health in Adults. *Social and Personality Psychology Compass*, 5(9), 634–652. <http://doi.org/10.1111/j.1751-9004.2011.00379.x>
- ¹⁸ Paletta, L., Zeiner, H., Schneeberger, M., & Quadri, Y. (2023). Digital Shadows and Twins for Human Experts and Data-Driven Services in a Framework of Democratic AI-based Decision Support. In: Paletta, L., Ayaz, A. & Asgher, U. (eds.). *AHFE Open Access*, 73. <http://doi.org/10.54941/ahfe1003971>
- ¹⁹ Paletta, L., Schneeberger, M., Pszeida, M., Mosbacher, J., Haid, F., Tschuden, J., & Zeiner, H. (2024). Resilience Scores from Wearable Biosignal Sensors for Decision Support of Worker Allocation in Production. In: Lucas Paletta (eds). *AHFE Open Access*, 124. <http://doi.org/10.54941/ahfe1004713>
- ²⁰ Lazarus, R.S., & Folkman, S. (1987). Transactional theory and research on emotions and coping. *European Journal of Personality*, 1, 141–169. <http://doi.org/10.1002/per.2410010304>
- ²¹ Bakker, A.B., & Demerouti, E. (2007). The Job Demands-Resources model: state of the art. *Journal of Managerial Psychology*, 22, 309–328. <http://doi.org/10.1108/02683940710733115>
- ²² van Veldhoven, M.J.P.M. (2008). Need for recovery after work: An overview of construct, measurement and research. In Houdmont J., & Leka S. (Eds.), *Occupational health psychology*. 1–25.
- ²³ Hobfoll, S.E. (2001). The Influence of Culture, Community, and the Nested-Self in the Stress Process: Advancing Conservation of Resources Theory. *Applied Psychology*, 50, 337–421. <http://doi.org/10.1111/1464-0597.00062>
- ²⁴ de Vries H., Kamphuis W., Oldenhuis H., van der Schans C., & Sanderman R. (2019). Modelling employee resilience using wearables and apps: a conceptual framework and research design. *International Journal of Advanced Life Sciences*, 11, 110–117.
- ²⁵ Hobfoll, S.E. (2001). The Influence of Culture, Community, and the Nested-Self in the Stress Process: Advancing Conservation of Resources Theory. *Applied Psychology*, 50, 337–421. <https://doi.org/10.1111/1464-0597.00062>
- ²⁶ Paletta, L., Zeiner, H., Schneeberger, M., Pszeida, M., Mosbacher, J.A., & Tschuden, J. (2024). Resilience Scores for Decision Support Using Wearable Biosignal Data with Requirements on Fair and Transparent AI. *IEEE 29th International Conference on Emerging Technologies and Factory Automation (ETFA)*, 1-4. <https://doi.org/10.1109/ETFA61755.2024.10710860>
- ²⁷ Kellmann, M., & Kallus, K.W. (Eds.). (2024). *The Recovery-Stress Questionnaires: A User Manual* (1st ed.). <https://doi.org/10.4324/9781032643380>
- ²⁸ Heidari, J., Beckmann, J., Bertollo, M., Brink, M., Kallus, K. W., Robazza, C., & Kellmann, M. (2019). Multidimensional Monitoring of Recovery Status and Implications for Performance. *International Journal of Sports Physiology and Performance*, 14(1), 2-8. <https://doi.org/10.1123/ijspp.2017-0669>
- ²⁹ Halson, S.L. (2014). Monitoring training load to understand fatigue in athletes. *Sports Medicine*, 44, 139–147. <https://doi.org/10.1007/s40279-014-0253-z>
- ³⁰ Kellmann, M., & Kallus, K.W. (Eds.). (2024). *The Recovery-Stress Questionnaires: A User Manual* (1st ed.). <https://doi.org/10.4324/9781032643380>
- ³¹ Meeusen, R., Duclos, M., Foster, C., Fry, A., Gleeson, M., Nieman, D., Raglin, J., Rietjens, G., Steinacker, J., & Urhausen, A. (2013). Prevention, diagnosis and treatment of the overtraining syndrome: joint consensus statement of the European College of Sport Science and the American College of Sports Medicine. *Medicine and Science in Sports and Exercise*, 45, 186–205. <https://doi.org/10.1249/MSS.0b013e318279a10a>
- ³² Jiménez, P., Dunkl, A., & Kallus, K.W. (2024). Recovery–stress-questionnaire for work. In: Kallus KW, Kellmann M, eds. *The Recovery-Stress Questionnaires: User Manual*, 167–198.
- ³³ Smith, B.W., Dalen, J., Wiggins, K., Tooley, E., Christopher, P., & Bernard, J. (2008). The brief resilience scale: assessing the ability to bounce back. *International Journal of Behavioural Medicine*, 15(3), 194-200. <https://doi.org/10.1080/10705500802222972>
- ³⁴ Paletta, L., Zeiner, H., Schneeberger, M., & Quadri, Y. (2023). Digital Shadows and Twins for Human Experts and Data-Driven Services in a Framework of Democratic AI-based Decision Support. In: Paletta, L., Ayaz, A. & Asgher, U. (eds.). *AHFE Open Access*, 73. <http://doi.org/10.54941/ahfe1003971>
- ³⁵ Paletta, L., Zeiner, H., Schneeberger, M., Pszeida, M., Mosbacher, J.A., & Tschuden, J. (2024). Resilience Scores for Decision Support Using Wearable Biosignal Data with Requirements on Fair and Transparent AI. *IEEE*

29th International Conference on Emerging Technologies and Factory Automation (ETFA), 1-4. <https://doi.org/10.1109/ETFA61755.2024.10710860>

³⁶ Kellmann, M., & Kallus, K.W. (Eds.). (2024). *The Recovery-Stress Questionnaires: A User Manual* (1st ed.). <https://doi.org/10.4324/9781032643380>

³⁷ Cohen, S., Kamarck, T., & Mermelstein, R. (1983). A global measure of perceived stress. *Journal of Health and Social Behavior*, 24(4), 385–396. <https://doi.org/10.2307/2136404>

³⁸ Suresh, H., & Gutttag, J.V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint*, 2(8), 72.

³⁹ Jiang, J. & Lu, Z. (2019). Learning Fairness in Multi-Agent Systems. *Advances in Neural Information Processing Systems*, 32.

⁴⁰ Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Del Ser, J., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106. <https://doi.org/10.1016/j.inffus.2024.102301>

⁴¹ Vilone, G., & Longo, L., (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89-106. <https://doi.org/10.1016/j.inffus.2021.05.009>

⁴² Prunet, T., Absi, N., Borodin, V., & Cattaruzza, D. (2024). Optimization of human-aware logistics and manufacturing systems: A comprehensive review of modeling approaches and applications. *EURO Journal on Transportation and Logistics*, 100136. <https://doi.org/10.1016/j.ejtl.2024.100136>

⁴³ Fraga, T. B. (2015). Three hybridization models based on local search scheme for job shop scheduling problem. *IOP Conference Series: Materials Science and Engineering*, 83(1), 012001). <https://doi.org/10.1088/1757-899X/83/1/012001>

⁴⁴ Dantzig, G. B., (1951). Maximization of a linear function of variables subject to linear inequalities. In: Koopmans, T. C. (Ed.), *Activity Analysis of Production and Allocation*. Wiley, 339–347

⁴⁵ Laguna, M., Marti, R., 2013. Heuristics. In: Gass, S. I., Fu, M. C. (Eds.), *Encyclopedia of Operations Research and Management Science*. Springer, Boston, MA, 695–703.

⁴⁶ Nummiluikki, J., Saxholm, S., Kärkkäinen, A., & Koskinen, S. (2023). Digital Calibration Certificate in an industrial application. *Acta IMEKO*, 12(1), 1-6. <https://doi.org/10.21014/actaimeko.v12i1.1402>

⁴⁷ Berti, N., Finco, S., & Battini, D. (2021). A new methodological framework to schedule job assignments by considering human factors and workers' individual needs. *Proceedings of the Summer School Francesco Turco*, 7.

⁴⁸ Tropschuh, B., Cegarra, J., & Battaia, O. (2024). Integrating physiological and mental aspects in employee scheduling: an overview for practitioners in production management. *International Journal of Production Research*, 62(6), 2093-2106. <https://doi.org/10.1080/00207543.2023.2217278>

⁴⁹ Adenipekun, E. O., Limère, V., & Schmid, N. A. (2022). The impact of transportation optimisation on assembly line feeding. *Omega*, 107, 102544. <https://doi.org/10.1016/j.omega.2021.102544>

⁵⁰ Fu, J., & Ma, L. (2022). Long-haul vehicle routing and scheduling with biomathematical fatigue constraints. *Transp. Sci.* 56 (2), 404–435. <https://www.webofscience.com/wos/woscc/full-record/WOS:000731909100001>

⁵¹ Olbrych, S., Nasuta, A., Kemmerling, M., Abdelrazeq, A., & Schmitt, R. (2024). From Simple to Sophisticated: Investigating the Spectrum of Decision Support Complexity with AI Integration in Manufacturing. In: Lucas Paletta (eds) *Cognitive Computing and Internet of Things*. *AHFE Open Access*, 124. <http://doi.org/10.54941/ahfe1004711>

⁵² Nasuta, A., Kemmerling, M., Lütticke, D., & Schmitt, R.H. (2024). Reward Shaping for Job Shop Scheduling. In: Nicosia, G., Ojha, V., La Malfa, E., La Malfa, G., Pardalos, P.M., Umeton, R. (eds) *Machine Learning, Optimization, and Data Science*. *Lecture Notes in Computer Science*, 14505. https://doi.org/10.1007/978-3-031-53969-5_16

⁵³ Olbrych, S. [Github-Link] <https://github.com/Sylwia-Olbrych/FAIRWork-AHP-FuzzyLogic-Safety-Efficiency>

⁵⁴ Olbrych, S. [Github-Link] <https://github.com/Sylwia-Olbrych/FAIRWork-Genetic-Algorithm-Allocation>

⁵⁵ Olbrych, S. [Github-Link] <https://github.com/Sylwia-Olbrych/FAIRWork-RL-Inventory-Management>

⁵⁶ Biewald, L. (2020). Experiment Tracking with Weights and Biases.

- ⁵⁷ Bertolini, M., Mezzogori, D., Neroni, M., & Zammori, F. (2021). Machine Learning for industrial applications: A comprehensive literature review. *Expert Systems with Applications*, 175, 114820. <https://doi.org/10.1016/j.eswa.2021.114820>
- ⁵⁸ Ivanov, D., Tang, C. S., Dolgui, A., Battini, D., & Das, A. (2020). Researchers' perspectives on Industry 4.0: multi-disciplinary analysis and opportunities for operations management. *International Journal of Production Research*, 59(7), 2055–2078. <https://doi.org/10.1080/00207543.2020.1798035>
- ⁵⁹ Olbrych, S., Nasuta, A., Kemmerling, M., Abdelrazeq, A., & Schmitt, R. (2024). From Simple to Sophisticated: Investigating the Spectrum of Decision Support Complexity with AI Integration in Manufacturing. In: Lucas Paletta (eds) *Cognitive Computing and Internet of Things. AHFE Open Access*, 124. <http://doi.org/10.54941/ahfe1004711>
- ⁶⁰ Nasuta, A., Kemmerling, M., Lütticke, D., & Schmitt, R.H. (2024). Reward Shaping for Job Shop Scheduling. In: Nicosia, G., Ojha, V., La Malfa, E., La Malfa, G., Pardalos, P.M., & Umeton, R. (eds). *Machine Learning, Optimization, and Data Science. Lecture Notes in Computer Science*, 14505. https://doi.org/10.1007/978-3-031-53969-5_16
- ⁶¹ Jaffri, A. (2024). *Explore Beyond GenAI on the 2024 Hype Cycle for Artificial Intelligence*. URL: <https://www.gartner.com/en/articles/hype-cycle-for-artificial-intelligence> (accessed 10.2.2025)
- ⁶² AI HLEG. (2019). *Ethics guidelines for trustworthy AI*. EC. URL <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- ⁶³ AI HLEG. (2020). *The Assessment List for Trustworthy Artificial Intelligence (ALTAI)*. EC. <https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment>
- ⁶⁴ Paletta, L. (2024). *First DAI-DSS Research Collection - D3.2*. https://fairwork-project.eu/deliverables/D3.2_First-DAI-DSS-ResearchCollection_V1.0_preliminary.pdf
- ⁶⁵ Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1), 65–100. [https://doi.org/10.1016/S0364-0213\(87\)80026-5](https://doi.org/10.1016/S0364-0213(87)80026-5)
- ⁶⁶ Mayr, H. C., & Thalheim, B. (2020). The triptych of conceptual modelling: A framework for a better understanding of conceptual modeling. *Software and Systems Modeling*, 20(1), 7-24. <https://doi.org/10.1007/s10270-020-00836-Z>
- ⁶⁷ Fettke, P. (2020). Conceptual Modelling and Artificial Intelligence: Overview and research challenges from the perspective of predictive business process management. *Companion Proceedings of Modellierung*, 2542, 157-164. CEUR-WS.org. <https://ceur-ws.org/Vol-2542/MOD-KI4.pdf>
- ⁶⁸ Bork, D., Ali, S. J., & Roelens, B. (2023). Conceptual modeling and artificial intelligence: A systematic mapping study. *arXiv preprint*.
- ⁶⁹ Shlezinger, N., Whang, J., Eldar, Y., & Dimakis, A. (2020). Model-Based Deep Learning. *Proceedings of the IEEE*, 111(5), 465-499. <https://doi.org/10.1109/JPROC.2023.3247480>
- ⁷⁰ Mattioli, J., Pedroza, G., Khalfaoui, S., & Leroy, B. (2022). Combining Data-Driven and Knowledge-Based AI Paradigms for Engineering AI-Based Safety-Critical Systems. In *Workshop on Artificial Intelligence Safety (SafeAI)*.
- ⁷¹ Bee-up [Software/Platform]. OMILAB. <https://bee-up.omilab.org/activities/bee-up/> (accessed 15.01.2025)
- ⁷² Adoxx [Software/Platform]. OMILAB. <https://www.adoxx.org/> (accessed 15.01.2025)
- ⁷³ Yokoy [Software/Platform]. Yokoy Switzerland Ltd. <https://yokoy.io/> (accessed: 10.2.2025)
- ⁷⁴ Scene2Model [Software/Platform]. OMILAB. <https://scene2model.omilab.org/> (accessed: 10.2.2025)
- ⁷⁵ Woitsch, R., Muck, C., Utz, W., & Zeiner, H. (2024). Enable Flexibilisation in FAIRWork's Democratic AI-based Decision Support System by Applying Conceptual Models Using ADOxx. *Complex Systems Informatics and Modeling Quarterly*, 38, 27–53. <https://doi.org/10.7250/csimg.2024-38.02>
- ⁷⁶ Muck, C., & Palkovits-Rauter, S. (2022). Conceptualizing Design Thinking Artefacts: The Scene2Model Storyboard Approach. In D. Karagiannis, M. Lee, K. Hinkelmann, & W. Utz (Eds.), *Domain-Specific Conceptual Modeling: Concepts, Methods and ADOxx Tools*, 567–587. https://doi.org/10.1007/978-3-030-93547-4_25
- ⁷⁷ Woitsch, R., Muck, C., Utz, W., & Zeiner, H. (2024). Enable Flexibilisation in FAIRWork's Democratic AI-based Decision Support System by Applying Conceptual Models Using ADOxx. *Complex Systems Informatics and Modeling Quarterly*, 38, 27–53. <https://doi.org/https://doi.org/10.7250/csimg.2024-38.02>

- ⁷⁸ Muck, C., Tschuden, J., Zeiner, H., & Utz, W. (2024). Explainability of Industrial Decision Support System using Digital Design Thinking with Scene2Model. *Cognitive Computing and Internet of Things*, 124(124). <http://doi.org/10.54941/ahfe1004710>
- ⁷⁹ Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. *Cognitive Science*, 11(1), 65–100. [https://doi.org/10.1016/S0364-0213\(87\)80026-5](https://doi.org/10.1016/S0364-0213(87)80026-5)
- ⁸⁰ Mayr, H. C., & Thalheim, B. (2020). The triptych of conceptual modelling: A framework for a better understanding of conceptual modeling. *Software and Systems Modeling*, 20(1), 7-24. <https://doi.org/10.1007/s10270-020-00836-Z>
- ⁸¹ Bork, D., Buchmann, R., Karagiannis, D., Lee, M., & Miron, E.-T. (2018). An Open Platform for Modeling Method Conceptualization: The OMiLAB Digital Ecosystem. *Communications of the Association for Information Systems*, 34, 555–579. <http://eprints.cs.univie.ac.at/5462/>
- ⁸² Mayr, H. C., & Thalheim, B. (2020). The triptych of conceptual modelling: A framework for a better understanding of conceptual modeling. *Software and Systems Modeling*, 20(1), 7-24. <https://doi.org/10.1007/s10270-020-00836-Z>
- ⁸³ Vernadat, F. (2020). Enterprise modelling: Research review and outlook. *Computers in Industry*, 122, 103265. <https://doi.org/10.1016/j.compind.2020.103265>
- ⁸⁴ Szvetits, M., & Zdun, U. (2016). Systematic literature review of the objectives, techniques, kinds, and architectures of models at runtime. *Software & Systems Modeling*, 15(1), 31–69. <https://doi.org/10.1007/s10270-013-0394-9>
- ⁸⁵ Moody, D. (2009). The “physics” of notations: toward a scientific basis for constructing visual notations in software engineering. *IEEE Transactions on Software Engineering*, 35(6), 756–779. <https://doi.org/10.1109/TSE.2009.67>
- ⁸⁶ Miron, E.-T., Muck, C., & Karagiannis, D. (2019). Transforming Haptic Storyboards into Diagrammatic Models: The Scene2Model Tool. *Proceedings of the 52nd Hawaii International Conference on System Sciences*.
- ⁸⁷ Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction*, 36(6), 495-504. <https://doi.org/10.1080/10447318.2020.1741118>
- ⁸⁸ Shneiderman, B. (2022). *Human-centered AI*. Oxford University Press.
- ⁸⁹ Nahavandi, S. (2019). Industry 5.0—A human-centric solution. *Sustainability*, 11(16), 4371. <https://doi.org/10.3390/su11164371>
- ⁹⁰ AI HLEG. (2019a). *Policy and investment recommendations for trustworthy Artificial Intelligence*. URL <https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence>
- ⁹¹ AI HLEG. (2019b). *Ethics guidelines for trustworthy AI*. European Commission. <https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai>
- ⁹² Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- ⁹³ Felzmann, H., Villaronga, E. F., Lutz, C., & Tamò-Larrieux, A. (2019). Transparency you can trust: Transparency requirements for artificial intelligence between legal norms and contextual concerns. *Big Data & Society*, 6(1), 1–14. <https://doi.org/10.1177/2053951719860542>
- ⁹⁴ Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 24:1-24:45. <https://doi.org/10.1145/3387166>
- ⁹⁵ van Nuenen, T., Ferrer, X., Such, J. M., & Cote, M. (2020). Transparency for whom? Assessing discriminatory artificial intelligence. *Computer*, 53(11), 36–44. <https://doi.org/10.1109/MC.2020.3002181>
- ⁹⁶ Venkatesh, V., Thong, J. Y., & Xu, X. (2016). Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the association for Information Systems*, 17(5), 328-376. Available at SSRN: <https://ssrn.com/abstract=2800121>

- ⁹⁷ Siau, K., & Wang, W. (2018). Building trust in artificial intelligence, machine learning, and robotics. *Cutter business technology journal*, 31(2), 47-53.
- ⁹⁸ Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- ⁹⁹ Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science*, 48(1), 137–141. <https://doi.org/10.1007/s11747-019-00710-5>
- ¹⁰⁰ Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., & Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080. <https://doi.org/10.1073/pnas.1900654116>
- ¹⁰¹ Páez, A. (2019). The Pragmatic Turn in Explainable Artificial Intelligence (XAI). *Minds and Machines*, 29(3), 441–459. <https://doi.org/10.1007/s11023-019-09502-w>
- ¹⁰² Miller, A. P. (2018). Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review*. <https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms>
- ¹⁰³ Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, 58, 82–115. <https://doi.org/10.1016/j.inffus.2019.12.012>
- ¹⁰⁴ Kaplan, A. D., Kessler, T. T., Brill, J. C., & Hancock, P. A. (2023). Trust in artificial intelligence: Meta-analytic findings. *Human Factors*, 65(2), 337-359. <https://doi.org/10.1177/001872082110139>
- ¹⁰⁵ Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 24:1-24:45. <https://doi.org/10.1145/3387166>
- ¹⁰⁶ Werz, J. M., Borowski, E., & Isenhardt, I. (2020). When imprecision improves advice: Disclosing algorithmic error probability to increase advice taking from algorithms. In C. Stephanidis & M. Antona (Eds.), *HCI International 2020—Posters*, 504–511. https://doi.org/10.1007/978-3-030-50726-8_66
- ¹⁰⁷ Werz, J. M., Borowski, E., & Isenhardt, I. (2024). Explainability as a means for transparency? Lay users' requirements towards transparent AI. *Cognitive Computing and Internet of Things*, 124. <https://doi.org/10.54941/ahfe1004712>
- ¹⁰⁸ Daschner, S., & Obermaier, R. (2022). Algorithm aversion? On the influence of advice accuracy on trust in algorithmic advice. *Journal of Decision Systems*, 31(sup1), 77–97. <https://doi.org/10.1080/12460125.2022.2070951>
- ¹⁰⁹ Springer, A., & Whittaker, S. (2018). What are you hiding? Algorithmic transparency and user perceptions. *AAAI Spring Symposium Series*, 1–4.
- ¹¹⁰ Ali, S., Abuhmed, T., El-Sappagh, S., Muhammad, K., Alonso-Moral, J. M., Confalonieri, R., Guidotti, R., Del Ser, J., Díaz-Rodríguez, N., & Herrera, F. (2023). Explainable Artificial Intelligence (XAI): What we know and what is left to attain Trustworthy Artificial Intelligence. *Information Fusion*, 99, 101805. <https://doi.org/10.1016/j.inffus.2023.101805>
- ¹¹¹ Carvalho, D. V., Pereira, E. M., & Cardoso, J. S. (2019). Machine Learning Interpretability: A Survey on Methods and Metrics. *Electronics*, 8(8), Article 8. <https://doi.org/10.3390/electronics8080832>
- ¹¹² Molnar, C. (2019). *Interpretable Machine Learning (1st edition)*. Christoph Molnar (CC Attribution 2.0). <https://christophm.github.io/interpretable-ml-book/index.html>
- ¹¹³ Mohseni, S., Zarei, N., & Ragan, E. D. (2021). A Multidisciplinary Survey and Framework for Design and Evaluation of Explainable AI Systems. *ACM Transactions on Interactive Intelligent Systems*, 11(3–4), 24:1-24:45. <https://doi.org/10.1145/3387166>
- ¹¹⁴ van Nuenen, T., Ferrer, X., Such, J. M., & Cote, M. (2020). Transparency for whom? Assessing discriminatory artificial intelligence. *Computer*, 53(11), 36–44. <https://doi.org/10.1109/MC.2020.3002181>
- ¹¹⁵ Héder, M. (2023). *Explainable AI: A Brief History of the Concept*. ERCIM News (134): 9–10.

- ¹¹⁶ Longo, L., Brcic, M., Cabitza, F., Choi, J., Confalonieri, R., Ser, J. D., Guidotti, R., Hayashi, Y., Herrera, F., Holzinger, A., Jiang, R., Khosravi, H., Lecue, F., Malgieri, G., Páez, A., Samek, W., Schneider, J., Speith, T., & Stumpf, S. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*, 106, 102301. <https://doi.org/10.1016/j.inffus.2024.102301>
- ¹¹⁷ Vilone, G. & Longo, L. (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*, 76, 89-106. <https://doi.org/10.1016/j.inffus.2021.05.009>
- ¹¹⁸ Castelveccchi, D (2016). Can we open the black box of AI? *Nature*. 538 (7623): 20-23. <https://doi.org/10.1038/538020a>
- ¹¹⁹ Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*. 54, 6 (2021), 1-35.
- ¹²⁰ Ribeiro, M.T., Singh, S., & Guestrin, C. (2016). "Why should I trust you?" explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, 1135-1144. 2016.
- ¹²¹ Lundberg, S. M., & Lee, S.-I. (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. <https://papers.nips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html>
- ¹²² Aas, K., Jullum, M., & Løland, A. (2021). Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. *Artificial Intelligence* 298.
- ¹²³ Wexler, J., Pushkarna, M., Bolukbasi, T., Wattenberg, M., Viégas, F., & Wilson, J. (2020). The What-If Tool: Interactive Probing of Machine Learning Models. *IEEE Transactions on Visualization and Computer Graphics*, 26(1), 56–65. *IEEE Transactions on Visualization and Computer Graphics*. <https://doi.org/10.1109/TVCG.2019.2934619>
- ¹²⁴ Sen, A. (1986). *Social choice theory*. Handbook of mathematical economics, 3, 1073-1181.
- ¹²⁵ De Jong, S., Tuyls, K., & Verbeeck, K. (2008). Fairness in multi-agent systems. *The Knowledge Engineering Review*, 23(2), 153-180. <https://doi.org/10.1017/S026988890800132X>
- ¹²⁶ Jiang, J., & Lu, Z. (2019). Learning fairness in multi-agent systems. *Advances in Neural Information Processing Systems*, 32.
- ¹²⁷ McGarraghy S, Olafsdottir G, Kazakov R, Huber É, Loveluck W, Gudbrandsdottir IY, Čechura L, Esposito G, Samoggia A, Aubert P-M, et al. Conceptual System Dynamics and Agent-Based Modelling Simulation of Interorganisational Fairness in Food Value Chains: Research Agenda and Case Studies. *Agriculture*. 2022; 12(2):280. <https://doi.org/10.3390/agriculture12020280>
- ¹²⁸ Tolan, S. (2019). Fair and unbiased algorithmic decision making: Current state and future challenges. *arXiv preprint arXiv:1901.04730*.
- ¹²⁹ de Jong, S., Tuyls, K., & Verbeeck, K. (2008). Artificial agents learning human fairness. Proceedings of the 7th International Joint Conference on Autonomous Agents and Multiagent Systems - Volume 2, 863–870. *International Foundation for Autonomous Agents and Multiagent Systems*.
- ¹³⁰ Grupen, N. A., Selman, B., & Lee, D. D. (2021). Fairness for Cooperative Multi-Agent Learning with Equivariant Policies. CoRR, abs/2106.05727. Retrieved from <https://arxiv.org/abs/2106.05727>
- ¹³¹ Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and Explanation in AI-Informed Decision Making. *Machine Learning and Knowledge Extraction*, 4(2), 556–579. <https://doi.org/10.3390/make4020026>
- ¹³² Jiang, J., & Lu, Z. (2019). Learning Fairness in Multi-Agent Systems. CoRR, abs/1910.14472. Retrieved from <http://arxiv.org/abs/1910.14472>
- ¹³³ Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and explanation in AI-informed decision making. *Machine Learning and Knowledge Extraction*, 4(2), 556-579.
- ¹³⁴ Saleiro, P., Kuester, B., Stevens, A., Anisfeld, A., Hinkson, L., London, J., & Ghani, R. (2018). Aequitas: A Bias and Fairness Audit Toolkit. *arXiv preprint arXiv:1811.05577*.

- ¹³⁵ Bellamy, R.K.E., Dey, K., Hind, M., Hoffman, S.C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., et al. (2018). AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*.
- ¹³⁶ Suresh, H., & Gutttag, J.V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002*.
- ¹³⁷ Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Front. Big Data*, 2, 13. <https://doi.org/10.3389/fdata.2019.00013>
- ¹³⁸ Suresh, H., & Gutttag, J.V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002*.
- ¹³⁹ Olteanu, A., Castillo, C., Diaz, F., & Kiciman, E. (2019). Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Frontiers in Big Data*, 2, 13. <https://doi.org/10.3389/fdata.2019.00013>
- ¹⁴⁰ Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35. <https://doi.org/10.1145/345760>
- ¹⁴¹ Suresh, H., & Gutttag, J.V. (2019). A Framework for Understanding Unintended Consequences of Machine Learning. *arXiv preprint arXiv:1901.10002*.
- ¹⁴² Blyth, C.R. (1972). On Simpson's paradox and the sure-thing principle. *Journal of the American Statistical Association*, 67, 338, 364–366.
- ¹⁴³ Binns, R.D.P. (2018). Fairness in machine learning: Lessons from political philosophy. *Journal of Machine Learning Research*.
- ¹⁴⁴ Hutchinson, B., & Mitchell, M. (2019). 50 Years of Test (Un) fairness: Lessons for Machine Learning. *Proceedings of the Conference on Fairness, Accountability, and Transparency*. ACM, 49–58. <https://doi.org/10.1145/3287560.3287600>
- ¹⁴⁵ Saxena, N.A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D.C., & Liu, Y. (2019). How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM, 99–106.
- ¹⁴⁶ Verma, S., & Rubin, J. (2018). Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.
- ¹⁴⁷ Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35.
- ¹⁴⁸ Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Computing Surveys*, 54(6), 1–35.
- ¹⁴⁹ Chandrasekaran, A. [Gartner]. 2024. Actionable Predictions for the Future of GenAI. <https://www.gartner.com/en/articles/3-bold-and-actionable-predictions-for-the-future-of-genai>
- ¹⁵⁰ Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., ... & Zhou, B. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys*, 55(9), 1-46. <https://doi.org/10.1145/3555803>
- ¹⁵¹ AI HLEG. (2019). Ethics guidelines for trustworthy AI. EC. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- ¹⁵² Szczuko P. (2024). Dos and Don'ts of LLMs. *Keynote on 32 International Conference on Information Systems Development*
- ¹⁵³ Akhai, S. (2023). *From black boxes to transparent machines: the quest for explainable AI*. Available at SSRN 4390887.
- ¹⁵⁴ Mattioli, J., Pedroza, G., Khalfaoui, S., & Leroy, B. (2022, February). Combining Data-Driven and Knowledge-Based AI Paradigms for Engineering AI-Based Safety-Critical Systems. *Workshop on Artificial Intelligence Safety (SafeAI)*.
- ¹⁵⁵ Bork, D., Ali, S. J., & Roelens, B. (2023). Conceptual Modeling and Artificial Intelligence: A Systematic Mapping Study.
- ¹⁵⁶ Karagiannis, D., & Kühn, H. (2002). Metamodelling platforms. *EC-Web*, 2455, 182.
- ¹⁵⁷ Karagiannis, D. (2015). Agile modeling method engineering. *Proceedings of the 19th Panhellenic Conference on Informatics*, 5–10.

- ¹⁵⁸ Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.*, 51(5). <https://doi.org/10.1145/3236009>
- ¹⁵⁹ Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.*, 54(6). <https://doi.org/10.1145/3457607>
- ¹⁶⁰ Binns, R. (2021). Fairness in Machine Learning: Lessons from Political Philosophy. *arXiv [Cs.CY]*. Retrieved from <http://arxiv.org/abs/1712.03586>
- ¹⁶¹ Floridi, L., & Taddeo, M. (2016). What is data ethics? *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2083), 20160360. <https://doi.org/10.1098/rsta.2016.0360>
- ¹⁶² Noorman, M., & Swierstra, T. (2023). Democratizing AI from a sociotechnical perspective. *Minds and Machines*, 1-24
- ¹⁶³ Woitsch, R., Muck, C., Utz, W., & Zeiner, H. (2023). Towards a democratic AI-based decision support system to improve decision making in complex ecosystems.
- ¹⁶⁴ Charles, L., Xia, S., & Coutts, A. P. (2022). Digitalization and Employment. *International Labour Organization Review*, 1-53.
- ¹⁶⁵ Hilton, S. K., Arkorful, H., & Martins, A. (2021). Democratic leadership and organizational performance: the moderating effect of contingent reward. *Management Research Review*, 44(7), 1042-1058.
- ¹⁶⁶ Dingwerth, K., Schmidtke, H., & Weise, T. (2020). The rise of democratic legitimation: why international organizations speak the language of democracy. *European Journal of International Relations*, 26(3), 714-741.
- ¹⁶⁷ Council of Europe (2020): E-Democracy Handbook. Available online at <https://rm.coe.int/11th-cddg-session-10-11-september-2020-e-democracy-handbook/16809f5a72> (accessed on 1/2/2024)
- ¹⁶⁸ Dhruvitkumar, T., Blessing, J., Smart, G., & Samuel, A. (2024). AI (Artificial Intelligence) in daily life. *Journal Name*, 600o, 35.