FIRST DAI-DSS RESEARCH COLLECTION

D3.2

Editor Name	Lucas Paletta (JR)
Submission Date	30.4.2024
Version	1.0
State	Final
Confidentially Level	PU

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

EXECUTIVE SUMMARY

This report focuses on deliverable "**First DAI-DSS Research Collection**", which is part of the Horizon Europe project FAIRWork. The deliverable aims to describe a first iteration on guidelines, methods and tools for democratising the production process in the light of their flexibilisation while using Artificial Intelligence, Optimisation, Human Factors Analytics, and Multi Agent Systems (MAS) as mediators in form of prototypes, physical experiments in laboratories, implemented questionnaires, modelling tools or semantic model of criteria catalogues. It presents a collection of concepts, methods, studies, and services of a research framework within the Democratised AI Decision Support System (DAI-DSS). The DAI-DSS research collection is based on the fundamental principles related to the research intended within the frame of this project that was described in Deliverable "D3.1 DAI-DSS Research Specification" and incorporates cross-connections with Deliverable "D4.1.1 DAI-DSS Architecture and initial Documentation and Test Report" on the basis of functional components of the DAI-DSS system architecture.

The first part of the report provides an **overview of the individual research tracks** related to the various research strategies that incrementally shape and extend the research collection within the frame of this project and in the context of the scientific as well as industrial communities. It covers the most significant research domains such as democratisation of decision-making as well as digital shadows for resilience risk stratification or human experts. Additionally, it explores technical approaches like Artificial Intelligence (AI) and MAS crucial for improving Decision Support Systems (DSS). This section also presents the state-of-the-art in key aspects of today's technology, particularly reliability and trustworthiness in AI. The output of this research overview leads to the outline of research activities in multiple domains that will be addressed within the FAIRWork project.

The second part of the report focuses on the research collection in terms of the concrete **research methods and services** employed to investigate the technical aspects of decision-making processes, human aspects in the process, and digital Human Factors measurements. It presents research approaches 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. Additionally, the report outlines the use of wearable sensors to capture critical information about the physiological, cognitive-emotional, and resilience state of humans, including implementation details of the Intelligent Sensor Box (ISB). Furthermore, the novel framework using Personas as Human Digital Twins for Decision Making in the context of Industry 5.0 is described in detail.

The third part covers initial observations on **explainability and fairness** in FAIRWork from an **algorithmic point of view** and summarises some reviews and surveys in the field.

The report also provides results of the strategy for **scientific dissemination** in the context of the research methodology of the FAIRWork project. The objective is to continuously disseminate project achievements, raise awareness about the project, and gather feedback to improve the created research artefacts.

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

Contributors and Reviewers

Contributors	Reviewers
Lucas Paletta, Herwig Zeiner, Michael Schneeberger, Julia Tschuden, Martin Pszeida, Andreas A. Mosbacher (JR) Gustavo Vieira (MORE) Sylwia Olbrych, Alexander Nasuta, Johanna Werz, Noushin Gheibi, Stefan Böschen (RWTH) Magdalena Dienstl, Marlene Mayer (BOC) Christian Muck (OMILAB)	Roland Perko (JR) Anas Abdelrazeq (RWTH) Knut Hinkelmann (OMILAB)

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

Version History

Version	Date	Authors	Sections Affected
1.0	April 30, 2024	Lucas Paletta 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 13
1.1 Purpose of the Document 13
1.2 Document Structure 13
2 Overview of Research Tracks 14
2.1 Research Tracks as Context of the Research Collection 14
2.2 Democratisation of Decision-Making in Socio-Technical Settings 15
2.3 Digital Human Factors Analytics 18
2.4 AI supported Optimisation in Decision Support Systems 21
2.5 AI-Enriched Decision Support Systems 22
2.6 Decision-Making Using Multi Agent Systems 23
2.7 Model-based Knowledge Engineering for Decision Support 24
2.8 Reliable and Trustworthy AI 26
3 Research Methods and Services 29
3.1 Overview of methods and services 29
3.2 Methods on Democratization of Decision-Making in Socio-Technical Settings 29
3.3 Methods and Services for Digital Human Factors Analytics 31
3.3.1 Overview 31
3.3.2 Service: AI-based Physiological Strain Estimation 32
3.3.2.1 Motivation and Reference to FAIRWork Use Case 32
3.3.2.2 Innovation beyond the State-of-the-art 33
3.3.2.3 Description of Functionality 34
3.3.2.4 Interface 34
3.3.2.5 Experiments 35
3.3.2.6 Results 36
3.3.2.7 Integration into the DAI-DSS architecture 37
3.3.3 Service: Heuristic Cognitive-emotional Stress Estimation 37
3.3.3.1 Motivation and Reference to FAIRWork Use Case 37
3.3.3.2 Innovation beyond the State-of-the-art 37
3.3.3.3 Description of functionality 38
3.3.3.4 Experiments 38
3.3.3.5 Integration into the DAI-DSS Architecture 39
3.3.4 Service: Resilience Score 39
3.3.4.1 Motivation and Reference to FAIRWork Use Case 39
3.3.4.2 Innovation beyond the State-of-the-art 41

3.3.4.3 Description of Functionality 43
3.3.4.4 Interfaces 44
3.3.4.5 Study Plan 44
3.3.4.6 Integration into the DAI-DSS Architecture 44
3.3.5 Concept: Persona-based Representation of Human Digital Twin 45
3.3.6 Outlook 45
3.4 Methods and Services for Optimisation in Decision Support Systems 46
3.4.1 Overview 46
3.4.2 Method: Automated Test Building Support with Hybrid Filtering 46
3.4.2.1 Motivation and Reference to FAIRWork Use Case 46
3.4.2.2 Innovation beyond the State-of-the-art 47
3.4.2.3 Description of Functionality 48
3.4.2.4 Interfaces 48
3.4.2.5 Experiments 48
3.4.2.6 Results 48
3.4.2.7 Integration into the DAI-DSS architecture 48
3.4.3 Method: Mathematical Optimisation/Heuristic for Worker Assigment 48
3.4.3.1 Motivation and Reference to FAIRWork Use Case 48
3.4.3.2 Innovation beyond the State-of-the-art 48
3.4.3.3 Description of Functionality 48
3.4.3.4 Interfaces 49
3.4.3.5 Experiments 49
3.4.3.6 Results 49
3.4.3.7 Integration into the DAI-DSS Architecture 49
3.5 Methods and Services for AI-Enriched Decision Support Systems 49
3.5.1 Overview 49
3.5.2 AI Catalogue – A Systematic Literature Review 49
3.5.2.1 Motivation and Reference to FAIRWork Use Case 49
3.5.2.2 Innovation beyond the State-of-the-art 49
3.5.2.3 Description of Functionality 50
3.5.3 Guidelines and Recommendations for AI Developers 50
3.5.3.1 Motivation and Reference to FAIRWork Use case 50
3.5.3.2 Innovation beyond the State-of-the-art 50
3.5.3.3 Description of Functionality 50
3.5.3.4 Results 50
3.5.4 Industrial Scheduling Optimisation 51

3.5.4.1 Motivation and Reference to FAIRWork Use Case 51
3.5.4.2 Innovation beyond the State-of-the-art 51
3.5.4.3 Description of functionality 52
3.5.4.4 Experiments 52
3.5.4.5 Results 53
3.6 Methods and Services for Decision-Making Using Multi Agent Systems 53
3.6.1 Overview 53
3.6.2 Multi-Agent Resource Allocation Service 53
3.6.2.1 Motivation and Reference to FAIRWork Use Case 53
3.6.2.2 Innovation beyond the state-of-the-art 53
3.6.2.3 Description of functionality 54
3.6.2.4 Interfaces 54
3.6.2.5 Experiments 55
3.7 Model-based Knowledge Engineering for Decision Support 55
3.7.1 Overview 55
3.7.1 Method: Conceptual Modelling for Knowledge Engineering – a three-layered approach 55
3.7.1.1 Motivation and Reference to FAIRWork Use Case 55
3.7.1.2 Innovation beyond the State-of-the-art 56
3.7.1.3 Interfaces 56
3.7.1.4 Experiments 56
3.7.1.5 Results - Fuzzy logic, Rules 57
3.7.1.6 Integration into the DAI-DSS Architecture 60
3.7.2 Service: Model-based configuration of Rule-Based Decision Services 61
3.7.2.1 Motivation and Reference to FAIRWork Use Case 61
3.7.2.2 Innovation beyond the State-of-the-art 61
3.7.2.3 Description of Functionality 61
3.7.2.4 Interfaces 62
3.7.2.5 Experiments 62
3.7.2.6 Results 64
3.7.2.7 Integration into the DAI-DSS architecture 64
3.7.2.8 Ethical issues 64
3.7.3 Service: Supporting FAIRWork's Design Methodology Through Supported Knowledge Transfer 64
3.7.3.1 Motivation within FAIRWork 64
3.7.3.2 Innovation beyond the state-of-the-art 65
3.7.3.3 Description of functionality 66
3.7.3.4 First Experiment Prototype 66

3.7.3.5 Interfaces 67
3.7.3.6 Integration into the DAI-DSS architecture 67
3.7.4 Outlook 67
3.8 Methods and Services for Reliable and Trustworthy AI 67
3.8.1 Overview 67
3.8.2 Method: Qualitative Focus Groups about AI Transparency 68
3.8.2.1 Motivation and Reference to FAIRWork Use Case 68
3.8.2.2 Innovation beyond the State-of-the-art 68
3.8.2.3 Description of Functionality 69
3.8.2.4 Experimental method 69
3.8.2.5 Results 69
3.8.2.6 Integration into the DAI-DSS Architecture 70
3.8.3 Method: Quantitative Experiment comparing AI Transparency Methods (completed) 70
3.8.3.1 Motivation and Reference to FAIRWork Use Case 70
3.8.3.2 Innovation beyond the State-of-the-art 71
3.8.3.3 Experimental Method 71
3.8.3.4 Results 72
3.8.3.5 Integration into the DAI-DSS Architecture 73
3.8.4 Method: Practical Application of Transparency in Different DAI-DSS Services (ongoing) 74
3.8.4.1 Motivation and Reference to FAIRWork Use Case 74
3.8.4.2 Innovation beyond the State-of-the-art 74
3.8.4.3 Description of Functionality 74
3.8.4.4 Experiments 74
3.8.4.5 Results 74
3.8.4.6 Integration into the DAI-DSS Architecture 75
3.8.5 Method: Evaluation of different DAI DSS services concerning Trustworthiness (ongoing) 75
3.8.5.1 Motivation and Reference to FAIRWork Use Case 75
3.8.5.2 Innovation beyond the State-of-the-art 75
3.8.5.3 Description of Functionality 76
3.8.5.4 Experiments 76
3.8.5.5 Results 76
3.8.5.6 Integration into the DAI-DSS Architecture 77
3.8.6 Outlook 77
3.9 Ethical Watchdog 78
3.9.1 Overview 78
3.9.2 Model-based assessment of ethical criteria to ensure compliance 78

3.9.2.1 Motivation and Reference to FAIRWork Use Case 78
3.9.2.2 Initial Experiments 78
3.9.3 Concept: Features of Human-centred Machine Learning Model 81
3.9.3.1 Motivation and Reference to FAIRWork Use Case 81
3.9.4 Concept: Supporting Decision Explanations through Conceptual Modelling 81
4 Explainability and Fairness in AI Services 83
4.1 Overview 83
4.2 Explainability of AI Services 84
4.3 Fairness in AI Services 84
4.3.1 Observations on Fairness in FAIRWork 84
4.3.2 Measures of Fairness 86
5 Summary and Conclusions 90
6 Scientific Dissemination 91
6.1 Publications Developed in the Context of the Research Collection 91
6.2 Organisation of Scientific Events 92
7 Annex A: List of Abbreviations 93

LIST OF FIGURES

Figure 1: Research tracks underlying the outline of the research collection. 14
Figure 2: Democratic Exploration Space: The Relation Question. 17
Figure 3: Levels of Decision-Making Processes with MAS giving a socio-technical structure to DAI-DSS. 18
Figure 4. One of the key dimensions of the Industry 5.0 initiative of the European Commission is an inherently social 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. 19
Figure 5: Wearable biosignal sensor-based assessment in the context of objective functions and optimisation. 20
Figure 6: Overview of three-layered approach. 25
Figure 7: Results of the RMSE in the regression-based estimation of the core body temperature. (a) Performance results in terms of error between actual core body temperature and estimated temperature using various AI-based and statistical models. The estimation using the Gaussian Progress Regression (top) provided the minimum RMSE value. (b) Actual course of core body temperature (blue line) and learned temperature course by the Gaussian Process Regression Model (red line) based on the data sets of the first five test subjects in the field tests of first responders. 35
Figure 8: Scatter plots of PSI and fitted/estimated PSI* comparing linear model (left column) with GPR model (right column) on artefact-adjusted data. These plots show results of PSI calculation method using fixed min/max heart rate and skin temperature values; the x=y diagonal represents a theoretically perfect match. 36
Figure 9: Example of a session with physiological strain on the treadmill with clearly specified step-wise load program and the Development Monitor with the course of raw data and generated PSI* in real-time visualisation. 37
Figure 10: First explorative studies with cognitive-emotional strain we applied a sequence of three tasks. a baseline session without substantial activity, a stimulus reaction choice task called „determination test“, and a task that is known to challenge cognitive load, that is, the n-back task. 38
Figure 11: Resulting data from first explorative studies with cognitive-emotional strain: operator video demonstrating the course of raw and processed data about cognitive-emotional stress, with cognitive activities of the operator synchronised with the measurement results. 39
Figure 12: The resilience risk stratification model (RRSM) as proposed to provide resilience scores for the decision support for the manager to decide on worker allocation. The specific contributions to this model is the accumulation of physiological and cognitive-emotional strain (PSI*, CES) to determine an integrated score representing the need for recovery and some degree of mental exhaustion. We model the relation of strain to determine a resilience score that both represents, resources and coping capacity to be able to master upcoming stressful challenges in the work environment. 40
Figure 13: Stages of the computation of the resilience score that underlies the risk stratification model. 42
Figure 14: FAIRWork Resilience Monitor. 43
Figure 15: Generation of PSI* and CES scores as basic data for the computation of the resilience score (JR Human Factors Lab, Graz, Austria). The resulting strain scores of limited-time experimental sessions are finally mapped to a Daily Score Score (i.e., *DSC(n))* of a specific day *n*. 44
Figure 16: Sketch of the adaptive DAI-DSS persona framework (Paletta et al., 2023). 45
Figure 17: Disjunctive graph scheduling (Nasuta et al., 2024). 52
Figure 18: Current approaches and outlook. 57
Figure 19: Rule-Based Allocation Service assigned to different layers. 58
Figure 20: Fuzzy Logic approach assigned to different layers. 58
Figure 21: Identification for Fuzzy Logic application. 59
Figure 22: Fuzzy Logic abstract logic. 60

Figure 23: Visualization of the Experiment for the Model-based Designing and Configuring of Decision Services. 63
Figure 24: Effect of different transparency conditions on usage of the algorithm. 73
Figure 25: Methodology extended with certification aspects. 79
Figure 26: Example of model-based signing service of “Specification Layer”. 79
Figure 27: Exemplary questionnaire. 80
Figure 28: Evaluation of the Fuzzy Logic with the Questionnaire. 80

LIST OF TABLES

Table 1: Comparison of models to predict body temperature during exercise in the heat (from Belval, 2016)..... 33
Table 2: Various error measures for the comparison between the application of linear and non-linear (i.e., using Gaussian Process Regression) regression models. 36
Table 3: Decision matrix for DSS selection (Olbrych et al., 2024). 51
Table 4: Four transparency conditions of the study. 72
Table 5: Abbreviations and long version explanation. 93

1 INTRODUCTION

1.1 Purpose of the Document

The goal of this document is to outline a first version of comprehensive DAI-DSS grounded research collection based on the fundamental principles embodied in the research tracks within the frame of this project that were schematically sketched in Deliverable “D3.1 DAI-DSS Research Specification”, to further proof the appropriateness of Multi Agent System (MAS) and Artificial Intelligence (AI) in Decision Support Systems (DSS).

The purpose of gathering this research collection consisting of research studies, principles, methods and services is to provide a detailed account about the most recent developments of the individual 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.

1.2 Document Structure

The document is structured as follows. The next section presents an overview of the principles of the individual research tracks relevant to the DAI-DSS framework conducted in this project. This overview includes several research tracks that outline the central themes and discuss topics such as democratization of decision-making using MAS, digital shadows and twins for human experts, AI-enriched DSS, and reliable and trustworthy AI.

The central section on the research methods and services embodies and outlines the research collection including detailed methodology that the authors implemented to investigate the human and technical factors in decision-making and the use of AI and MAS to enhance it.

Following the authors present how they disseminated their research findings and how they plan to progress further.

Finally, the report concludes with a summary section, where the authors summarize the key points debated in the report and emphasize the importance of incorporating the human perspective into decision-making processes and the need for reliable and trustworthy AI.

2 OVERVIEW OF RESEARCH TRACKS

2.1 Research Tracks as Context of the Research Collection

The research tracks presented in Figure 1 provide the scientific context in which the complete scenario of the numerous research activities within the FAIRWork project evolves.

The human is in the centre of the decision support system, as a decision maker, as a worker, and, in particular, as a means for the democratisation process in the industrial socio-technical settings. The **Democratisation of the Decision-Making** is a first, central and highly innovative research theme that is tackled at the very beginning. Fundamental to the human-centred socio-technical settings are the various aspects of the Human Factors in 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.

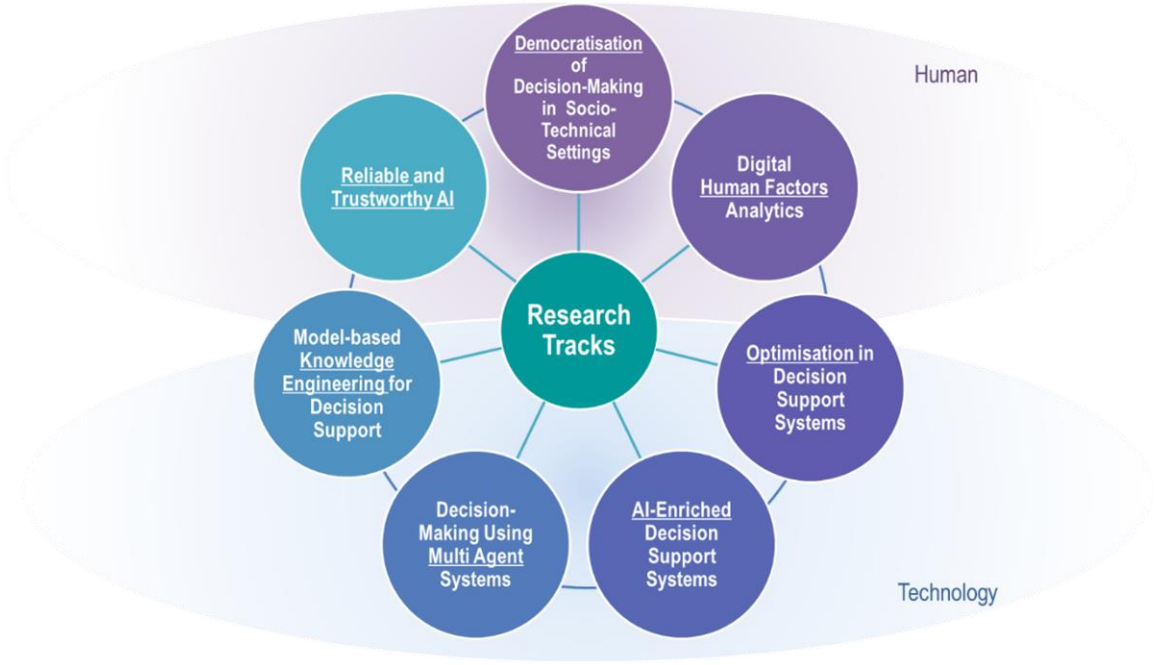


Figure 1: Research tracks underlying the outline of the research collection.

The more technological aspect of the project is firstly represented by the **Optimisation in Decision Support Systems** as a further central research area that relates relevant input data about the human behavioural status as well as the various system data with objective functions in order to provide meaningful orientation to the decision-making and the system processes. The **AI-enrichment of the Decision Support Systems** provides intelligence, such as, adaptiveness, reasoning and temporal context to the decision-making process. Another framework for the development of intelligent systems is represented by the research track provided by the **Decision-Making using Multi-Agent Systems**. In addition, the project FAIRWork focuses on **Model-based Knowledge Engineering for Decision Support** and with this strategy enables to complement AI-based and Multi-Agent-based approaches for a fully flexible treatment of challenges in decision-making within industrial process management.

Finally, there is a strong focus in the project FAIRWork on the **Reliability and Trustworthy AI.** This research track investigates and identifies requirements for an ethically correct integration of technologies in the socio-technical settings.

2.2 Democratisation of Decision-Making in Socio-Technical Settings

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 et al., 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 optimize the decision-support system for democratic decision-making, it is essential to closely examine employees' daily work routines and individual decision-making processes within the organization. This examination should focus on three key dimensions:

- Exploring the Decision-Making Dimension: This dimension includes identifying whether decision-making follows a traditional or flat structure, evaluating the types of decisions made, assessing the effectiveness of the current process, understanding the challenges faced during decision-making, and recognizing the factors employees should consider in their decision-making approach.
- Examining the Involvement Dimension: Involving employees in decision-making not only significantly enhance organizational efficiency but also fosters their creativity and commitment (Charles et al., 2021)[3]. Consequently, involvement is evaluated based on three key components: the opportunities for employee's objection and suggestion, the methods of transparency regarding decisions made, and the importance of trust among team members, which are interrelated concepts according to research findings (Rawlins, 2008[4], Grates, 2007[5], Chene and Chr, 2011[6]).
- Understanding the Expectation Dimension: The essential purpose of research in FAIRWork is exploring optimization prospects for a decision support system while considering the experiential insights of employees engaged in the workflow. To achieve this goal, the third dimension investigates participants' expectations of this system with the aim of simplifying, accelerating and ensuring the work process.

Expanding on the discourse surrounding the democratization of decision-making within socio-technical contexts, it is imperative to delve into the broader concept, specifically focusing on Democratic Decision-Making with Multi-Agent Systems (MAS). From a viewpoint of political systems, democracy is the most ambitious form to perform power by, for and with the people. Therefore, the quest of institutions is decisive to allow such a performance. With regard to our task here, democratic decision-making can be seen as an effort to involve all individuals within a group in the decision-making process and to prevent illegitimate centralization and concentration of power in this process. Typically, hierarchical organizations are not the place of democratic decision-making. In the meanwhile, questions of participation, representation or transparency are getting decisive. Thus, stakeholders and managers

[1] Noorman, M., & Swierstra, T. (2023). Democratizing AI from a sociotechnical perspective. Minds and Machines, 1-24
[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.
[3] Charles, J., Francis, F., & Zirra, C. T. O. P. (2021). Effect of employee involvement in decision making and organization productivity. Archives of Business Research (ABR), 9(3), 28-34.
[4] Rawlins, B. R. (2008). Measuring the relationship between organizational transparency and employee trust.
[5] GRATES, G. (2007). Edelman Trust Barometer 2007. Edelman.
[6] Chene, M. (2011). Good practice in strengthening transparency, participation, accountability and integrity. Transparency International and Chr, Michelsen Institute.

in companies are also striving for fair decision-making models (Charles et al., 2022[7], Hilton et al., 2021[8], Dingwerth et al., 2020[9]). Nowadays, the use of new technologies to enhance democratic decision-making models opens up new possibilities in this context, demanding additional examination. Under the FAIRWork project, we aim to develop a democratic model to enable fair decision-making in a production company using MAS for DAI-DSS.

The **Multi-Agent Systems (MAS)** MAS has to be interpreted as a technical form of representation. On one hand, this enhance opportunities of representation as the MAS can be supported with a broad range of parameters representing the present status of workers and production lines. On the other hand, this might go along with specific shortcomings of representativeness as the quality of the technical representation depends on the quality of the parameters set up so far. Moreover, although the quest of representation might be answered sufficiently on the technical level, there might problems of legitimacy emerge in the specific socio-organizational setting. Thus, the question for analysing the socio-technical setting becomes decisive. Thereby, the key question of representation has to be set up as part of a socio-technical practice and has to be part of the socio-technical modelling with and through MAS. Against this background, the specific method presented here is a reflexive method of analysing the embedding of the MAS tool into the respective socio-technical settings.

Developing a Democratic Decision-Making the balancing of the tension between social aims of representation and the technical form of representation is decisive. Doing so, it is recognised that shop-floor workers are embedded in the socio-process practices of their specific work environment. Consequently, there arises a distinct necessity for tailored representation to address their unique needs. Representation within this context may take shape through two primary channels: either workers autonomously articulate their needs, or experts are tasked with identifying and expressing these requirements (see Figure 2).

Given the potential variability in worker fitness and alertness, digital tools may offer advantages over workers' self-assessment in certain situations, while personal experiences may provide deeper insights in others. Our approach integrates both methods, employing digital tools alongside workers' self-assessment to comprehensively address the organizational issues. This integration extends to the operation of digital tools relying on specific inputs such as parameters, factors, and indicators representing the workers. These elements are integral components of the MAS representation, crucial for enhancing its functionality and informing decision-making processes. MAS creates specific realities for workers by making decisions using its internal support system. This prompts questions about the reliability of outcomes for the workers and whether they are adequately represented by the MAS. Consequently, tensions arise when confronting decisions made by the system. To address how these tensions are resolved, we propose considering the concept of legitimation. This involves examining why such a system is introduced as a social process, or why monitoring such a process is necessary. These questions revolve around the legitimacy of implementing such a system within the company's organizational framework. Hence, it is imperative to address the tensions indicated by the arrows. This raises the fundamental research question: How could a democratic character of such system be asserted?

[7] Charles, L., Xia, S., & Coutts, A. P. (2022). Digitalization and Employment. International Labour Organization Review, 1-53.
[8] 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.
[9] 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.

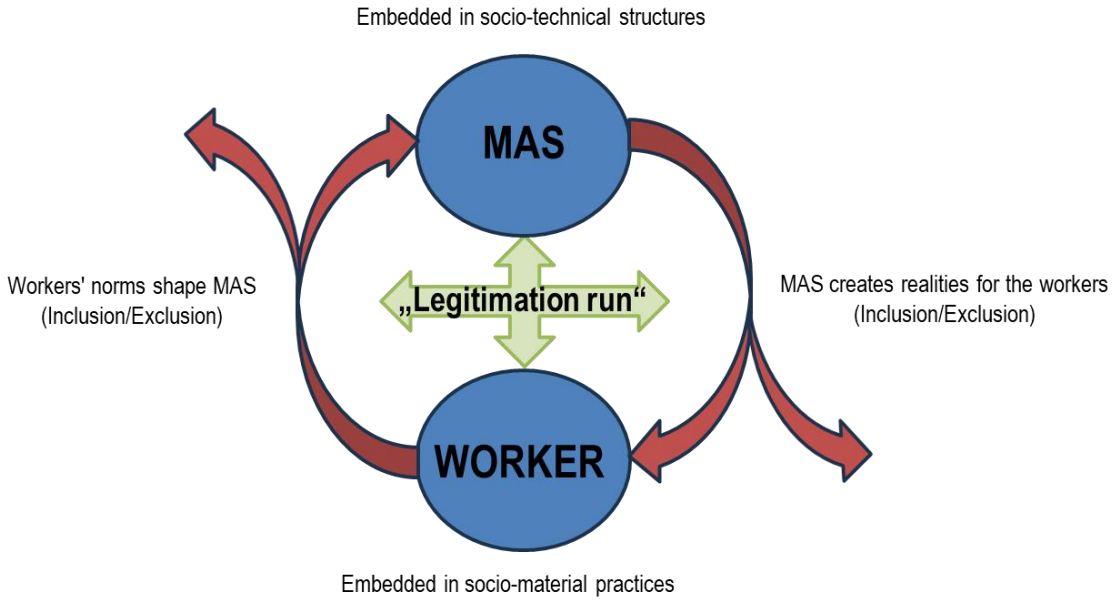


Figure 2: Democratic Exploration Space: The Relation Question.

To find answers to the above questions, we need to focus on the legitimation process, which aims to address the tension between social calls for representation and technical forms of representation. Therefore, we define the legitimation process by exploring levels of Decision-Making Processes within a company with the goal of modelling the socio-technical setting of decision-making, the relatedness of these levels and the feedback-loops possibly therein. With this regard, our model is represented as a five-layer funnel diagram, facilitating reflexive cycles of opening-up and closing-down of representations (see Figure 3).

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.

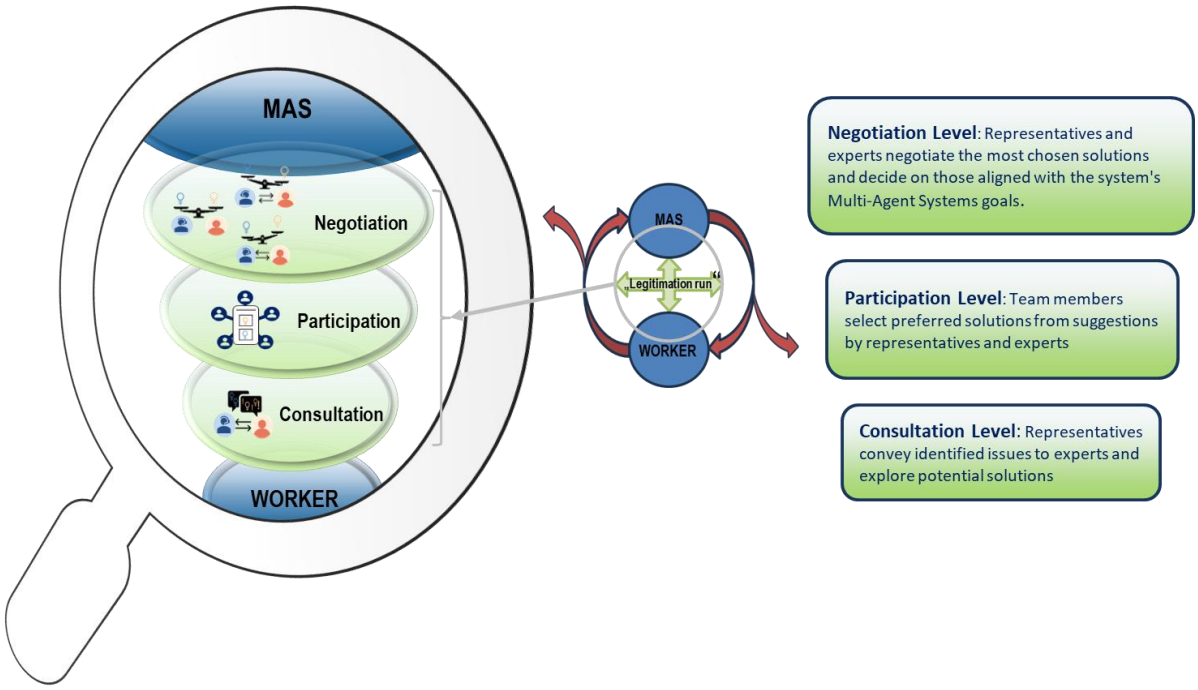


Figure 3: Levels of Decision-Making Processes with MAS giving a socio-technical structure to DAI-DSS.

2.3 Digital Human Factors Analytics

One of the key design and performance dimensions of the **Industry 5.0** initiative of the European Commission[10] is an inherently social 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 (see Figure 4). Building upon the digitalisation of the industrial processes we focus on the workers' well-being, empowering workers through the use of digital devices, endorsing a human-centric approach to technology.

[10] European Commission, Directorate-General for Research and Innovation, Renda, A., Schwaag Serger, S., Tataj, D. et al., Industry 5.0, a transformative vision for Europe – Governing systemic transformations towards a sustainable industry, Publications Office of the European Union, 2021, https://data.europa.eu/doi/10.2777/17322.

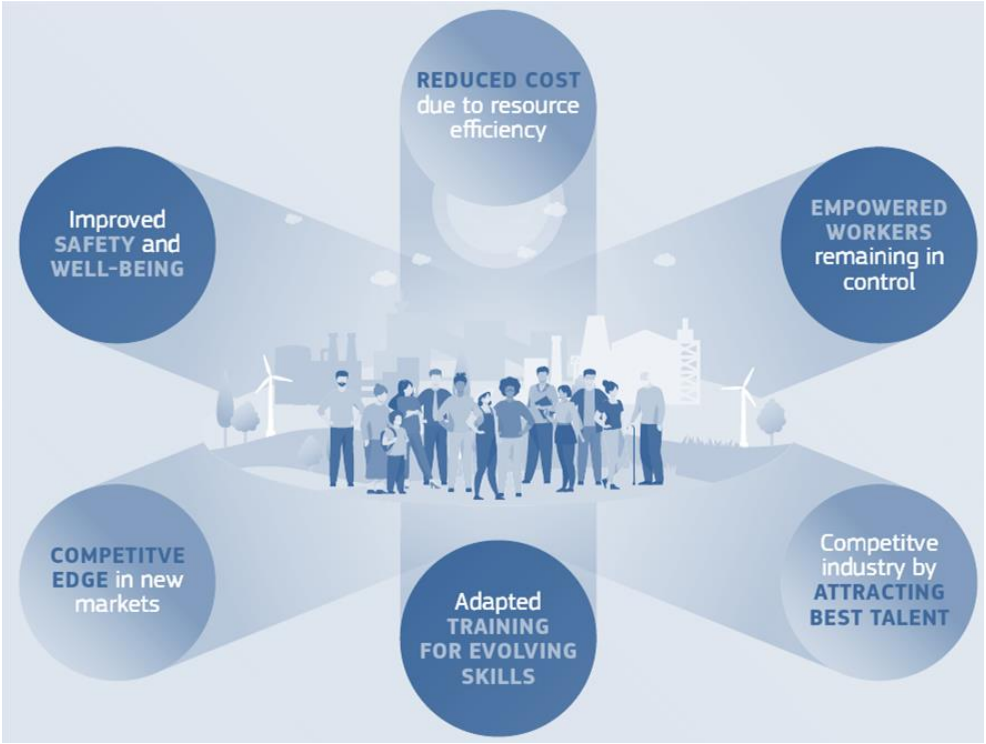


Figure 4. One of the key dimensions of the Industry 5.0 initiative of the European Commission is an inherently social dimension, demanding attention to the wellbeing of workers, the need for social inclusion and the adoption of technologies that do not substitute, but rather complement human capabilities[11].

FAIRWork provides a clear redirection and orientation of the digital transformation to enable more human, respectful people's **wellbeing** and sense of self-reliance on the digital economy. The new industrial system will have to reflect the growing post-COVID shift in human consciousness that survival in the face of crisis and access to essential goods and services is more important than the ownership of most material goods, all of which are useless when confined to home living and virtual working. The vision for 'Industry 5.0' moves past a narrow and traditional focus on technology-or economic enabled growth of the existing extractive, production and consumption driven economic model to a more transformative view of growth that is focused on human progress and well-being based on reducing and shifting consumption to new forms of sustainable, circular and regenerative economic value creation. Industry 5.0 aims to nest the Industry 4.0 approach in a broader context, providing directionality to the technological transformation of industrial production for the prosperity of a global socio-technical system in a holistic manner.

Stress overload can impact work and organisational success. However, with planning and human-centred responses, organizations can help build resilience among the workforce and enable them to adapt positively with the business. Indeed, the toxic stress overload caused by a crisis can diminish individual and broader human capital[12]. Beyond the visible impact of crises on personal health, family, and financial stability, sustained toxic stress can impact the part of the brain responsible for executive function[13]. This negative impact can weaken working memory, attention control, cognitive flexibility, and problem-solving—the cognitive processes that make people capable and productive both in their personal and professional lives[14]. Further, when a crisis is widespread, as with

[11] European Commission, Directorate-General for Research and Innovation, Industry 5.0 – Human-centric, sustainable and resilient, Publications Office, 2020, https://data.europa.eu/doi/10.2777/073781
[12] David L. Shern, Andrea K. Blanch, and Sarah M. Steverman, Impact of toxic stress on individuals and communities: A review of the literature, *Mental Health America*, September 16, 2014.
[13] Isham A, Mair S and T Jackson 2020. *Wellbeing and productivity: a review of the literature*. CUSP Working Paper No 22. Guildford: University of Surrey.
[14] Sylvia R. Karasu, "The obliterative, dislocating effects of stress," Psychology Today, January 24, 2019; *Psychology Today*, "Executive function," accessed July 25, 2020

COVID-19 and other recent events, stress can diminish the broader well-being and productivity of a company's workforce, leading to significant implications for the business as a whole.

When discussing responses to workplace stress it is important to mention the individual characteristic of **resilience**. Individuals possessing high resilience are said to display a greater capacity to cope with stressful work demands in comparison to other employees[15]. Resilience has been negatively related to burnout[16], positively related to job satisfaction[17] and negatively related to productivity losses and likelihood of absence from work[18]. Resilience can be conceptualised as a trait, it is also commonly considered as a process or capacity that can be developed in a temporal context, i.e., over time[19]. Under this process conceptualisation, some workplaces have started to implement interventions to increase the resilience of their workforce. A systematic review[20] revealed that the length of these interventions could vary from a single session of 90 minutes up to a 12-week programme and employ a variety of techniques such as skills-based coaching, mindfulness and compassion-based practices, cognitive behavioural techniques, e.g., energy management and relaxation training.

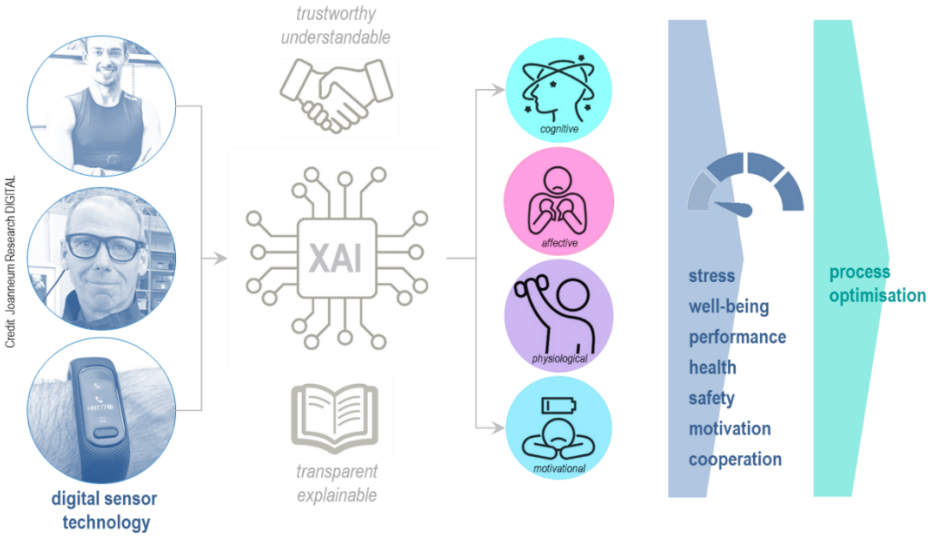


Figure 5: Wearable biosignal sensor-based assessment in the context of objective functions and optimisation.

The innovation trajectory of Digital Human Factors Analytics is substantially oriented to monitor and enable to manage the fundamental effects of toxic stress, resilience, and well-being of the worker. AI plays here a major role in developing an adjustable association of raw sensor data with relevant cognitive, affective, physiological and motivational constructs. Explainable AI (XAI) approaches make this association transparent for the expert and understandable for the worker as well as the decision maker and in this sense lays the foundation for a trustworthy intelligent service. In this context Figure 5 visualises the overall schema about how the biosignal sensor-based assessment is applied in FAIRWork in the context of objective functions.

[15] Winwood, P. C., Colon, R., & McEwen, K. (2013). A Practical Measure of Workplace Resilience. *Journal of Occupational and Environmental Medicine*, 55(10), 1205–1212. https://doi.org/10.1097/JOM.0b013e3182a2a60a

[16] Cooke, G. P., Doust, J. A., & Steele, M. C. (2013). A survey of resilience, burnout, and tolerance of uncertainty in Australian general practice registrars. *BMC Medical Education*, 13(1), 2. https://doi.org/10.1186/1472-6920-13-2

[17] Zheng, Z., Gangaram, P., Xie, H., Chua, S., Ong, S. B. C., & Koh, S. E. (2017). Job satisfaction and resilience in psychiatric nurses: A study at theInstitute of Mental Health, Singapore. *International Journal of Mental Health Nursing*, 26(6), 612–619. https://doi.org/10.1111/inm.12286

[18] Shatté, A., Perlman, A., Smith, B., & Lynch, W. D. (2017). The Positive Effect of Resilience on Stress and Business Outcomes in Difficult Work Environments. *Journal of Occupational and Environmental Medicine*, 59(2), 135–140. https://doi.org/10.1097/JOM.0000000000000914

[19] Howe, A., Smajdor, A., & Stöckl, A. (2012). Towards an understanding of resilience and its relevance to medical training. *Medical Education*, 46(4), 349–356. https://doi.org/10.1111/j.1365-2923.2011.04188.x

[20] Robertson, I. T., Cooper, C. L., Sarkar, M., & Curran, T. (2015). Resilience training in the workplace from 2003 to 2014: A systematic review. *Journal of Occupational and Organizational Psychology*, 88(3), 533–562. https://doi.org/10.1111/joop.12120

The "**Intelligent Sensor Box**" provides an architecture and intelligent sensor- and AI-based software components for the computing of the individual worker's well-being data in terms of the estimation of risk in the context of stress, resilience, and well-being. It includes components that estimate the daily quantities of stress via "**AI-based Physiological Stress Estimation**" as well "**AI-based Cognitive-emotional Stress Estimation**" and finally integrates long-term measurements into the "**Resilience Estimation**" component in order to receive a final resilience score for further processing. The "**Persona Clustering**" component finally provides a service for the determination of clusters in human-centered data. For this purpose it requires input of pseudonymised human data (socio-biographical data, measured stress data, etc.) from the Knowledge Base.

2.4 AI supported Optimisation in Decision Support Systems

Decision-making in production involves various challenges, from the allocation of resources to the optimisation of processes and supply chains. 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[21]). 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 different ways to support decision making with optimisation techniques:

- Mathematical programming is a central methodology in operations research. The simplex method, first published by Dantzig[22], 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[23] 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 FAIRWorks 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

[21] Tatiana Balbi Fraga 2015 *IOP Conf. Ser.: Mater. Sci. Eng.* **83** 012001**DOI** 10.1088/1757-899X/83/1/012001
[22] 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, pp. 339–347
[23] Laguna, M., Mart´ı, R., 2013. Heuristics. In: Gass, S. I., Fu, M. C. (Eds.), Encyclopedia of Operations Research and Management Science. Springer, Boston, MA, pp. 695–703.

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.

Approaches from mathematical optimisation can be combined with data-driven methods such as reinforcement learning.

2.5 AI-Enriched Decision Support Systems

During the allotted time frame for this periodic report, our efforts in WP3 were dedicated to developing guidelines tailored for developers specialising in creating recommendation tools. The research area of implementing AI methodologies in the DSS is wide-ranging. However, the extended research presented in Deliverable 3.1 provided a broad overview of the current state of the art and thus led us to define the following research questions:

- How can the decision-making process and architecture in the manufacturing environment be accelerated through the use of AI-based methodologies?
- How can existing AI methodologies enrich classical decision support systems in order to be applied in complex and dynamic manufacturing processes?
- How do different optimisation metrics or constraints affect a schedule in a manufacturing context?
- How and to what extent can AI techniques be utilised for optimisation in industrial scheduling?

To address the first question, we are conducting a systematic literature review. So far, we have identified over 200 scientific articles published within the past 14 years. These articles specifically focus on the implementation of machine learning, deep learning, or reinforcement learning methods within the scope of resource and production planning. The primary inquiry guiding this review is: *"What machine learning techniques have been applied within decision support systems in the manufacturing environment to facilitate resource and production planning?"*

Through this review process, we aim to gain insights into utilising machine learning methodologies in DSSs tailored for the manufacturing sector. This study enables us to discern patterns, trends, and best practices in the integration of advanced technologies to enhance the implementation of ML techniques for resource and production planning.

In the scope of the second research question, we recognised a gap in DSS classifications, where the need for adequate resources to guide developers in selecting an appropriate method is evident. In response to this gap, our research aims to improve the clarity and understanding of integrating existing methods, including AI, into DSSs. 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, aiding them in selecting the most suitable DSS for specific contexts and, consequently, enhancing its applicability and effectiveness in decision-making. This strategic approach aligns seamlessly with the dynamic needs of users and the evolving manufacturing landscape by incorporating AI techniques, ensuring that DSSs remain adaptive, effective, and finely tuned to the user requirements in the modern manufacturing domain. This research's outcome was summarised as a publication at the 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) (Olbrych et al., 2024[24]).

Additionally, we explore ML methodologies for optimising industrial scheduling, focusing particularly on Reinforcement Learning (RL), which has emerged as a promising approach in this domain and offers innovative solutions to complex optimisation challenges. As scheduling is critical in industrial settings for efficient resource allocation, downtime reduction, and productivity maximisation, integrating advanced methodologies like RL can significantly enhance decision-making processes and overall operational efficiency. In particular, the job shop

[24] Olbrych S., Nasuta A., Kemmerling M., Abdelrazeq A., Schmitt, R. H. "From Simple to Sophisticated: Investigating the Spectrum of Decision Support Complexity with AI Integration in Manufacturing". (2024). http://doi.org/inprint

problem (JSP) is highly relevant for flexible production scheduling in the modern era. Addressing the third research question, we examined various reward functions using a novel flexible RL environment for the JSP based on the disjunctive graph approach. Our experiments show that a formulation of the reward function based on machine utilisation is most appropriate for minimising the make-span of a JSP among the investigated reward functions. The results have been published at the 9th International Conference on Machine Learning, Optimization, and Data Science (LOD 2023) (Nasuta et al., 2024[25]).

2.6 Decision-Making Using Multi Agent Systems

Research on democratization of decision-making with Multi-Agent Systems (MAS) has been developed in the last months focusing on how MAS can enable a participatory environment for stakeholders in decision-making. We have been considering an approach where worker-related parameters play an essential role in the decision processes. Taking Deliverable 3.1 as basis, we delve deeper in exploration of multi-agent modelling, emphasizing the integration of worker preferences, skills, and interactions within the system. The incorporation of human-centric parameters enables the conception of a strengthened democratic decision-making. The consideration of humanly relevant factors such as human preference together with multi-agent modelling allows the exploration of balancing multiple stakeholders in a dynamic multi-parameter environment.

The following questions raised in the first stage of the project begin to be clarified:

- How can MAS be designed to ensure a democratic decision-making process in the industry?
- How can MAS contribute to enhancing worker participation in decision-making in the industry?
- How to transmit outcome and decision confidence from MAS to humans?
- How to model the human information that enables the agent to concretely and definitely, behave on behalf of the human?

In decision-making process, conflicts are inevitable due to the diversity of stakeholders' interests and also the high number of actors that might reach equal numbers. MAS allow for the incorporation of conflict resolution mechanisms that allow agents to negotiate, compromise, and reach consensus. Algorithms designed for conflict detection and management ensure that when agents have conflicting goals or actions, there are predefined pathways to resolution. Conflict resolution mechanisms are being considered in the current study in order to provide a robust support to decision. It is also important that a single agent or a group of agents don't dominate the decision-making process. Weight balancing plays an important role in ensuring that all stakeholders have a proportionate influence on the outcome through assigning appropriate weights to different agents' inputs based on given contexts. We are exploring how weights can improve decision-making fairness in industry.

MAS can significantly enhance worker participation in industrial decision-making by incorporating highly relevant data that encompasses a wide range of worker inputs, from human factors to individual preferences integrating valuable knowledge into decision-making in a bottom-up perspective. This approach not only leads to more informed and holistic decisions but also allows for worker satisfaction enhancement and engagement by valuing their input and making them active participants in shaping decisions.

Transmitting outcome and decision confidence from Multi-Agent Systems (MAS) to humans effectively involves MAS explainability. The ability of the system to clearly communicate its processes, decisions, and the confidence level of its outcomes in a manner that is understandable to humans is crucial for MAS solutions adoption. This involves breaking down complex decisions into simpler, understandable components, allowing users to see how and why specific decisions were made, how data and inputs were considered, and how they influenced the

[25] 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. LOD 2023. Lecture Notes in Computer Science, vol 14505. Springer, Cham. https://doi.org/10.1007/978-3-031-53969-5_16

outcome. Transparency through explainability in MAS has been approached in order to allow an evaluation of how MAS can engage human adoption of multi-agent solutions in supporting decision-making.

Human data modelling is responsible for the agent system to be socially aligned with human values and expectations, facilitating harmonious and effective human-agent symbiosis. Since humans are not static entities, their preferences and priorities change over time due to various factors. Agents have to incorporate these characteristics to achieve seamless integration in a human centered system. The social-technical analysis is fundamental to provide an understanding of the systemic interrelations that significantly influence the concrete agent behaviour on human behalf.

2.7 Model-based Knowledge Engineering for Decision Support

Modelling is an approach that creates simplified representations of reality (Talheim & Nissen, 2015)[26], facilitating interaction with complex systems, whereby they are created using a comprehensible representation, allowing to share them with other stakeholders or machines (Mayr & Talheim, 2020)[27]. Models can be created inductively if the model is based on observations or deductively if the model is created out of theories. Numerous variants can be distinguished regarding the types of models, but formal and conceptual (semi-formal) models are the most common in business informatics (Wilde & Hess, 2006)[28]. In our research, we focus on conceptual modelling, as it enables one to examine a system from a conceptual perspective and to capture its most important structural, behavioural, or semantic features. This is beneficial when a current system should be examined and when creating a new one (Karagiannis et. al., 2016)[29]. The focus of this research track is set on using conceptual modelling for encoding this knowledge and increase the model value (Bork et. al., 2018)[30] by offering functionality based on processing the models and supporting the humans.

For conducting our research regarding the use and integration of AI algorithms and multi-agent systems (MAS) in FAIRWork's DAI-DSS, modelling is used to externalize the knowledge of decision-makers and experts familiar with the use case scenario and depict the corresponding decision paths. Such models have the advantage that humans and machines can interpret them. So, these models can further be used to support the mapping process between defined requirements and suitable AI solutions. Furthermore, models can assist in the configuration of the DAI-DSS on one hand and on the other use the knowledge encoded in the models in a comprehensible way to support the explanation of the decisions to involved stakeholders.

By applying a systematic three-layered approach starting with the problem setting in the use case resulting in an executable AI algorithm for decision support, the goal of BOC's research in this research track is to close the semantic gap in between the layers. The overall methodology of the requirement and capability mapping description is illustrated Figure 6 and includes:

1) Identification: The identification and design of decision processes, corresponding success factors, and relevant KPIs through harvesting, and modeling domain knowledge can be supported with approaches such as co-creative workshops, or interviews.

2) Specification: In the second layer the initially defined properties are described on abstract decision logic or methods for knowledge-based mechanisms e.g. input and output data types and underlying expectations for

[26] Thalheim, B., & Nissen, I. (2015). Wissenschaft und Kunst der Modellierung: Kieler Zugang zur Definition, Nutzung und Zukunft (Vol. 64). Walter de Gruyter GmbH & Co KG.
27 Mayr, H. C., & Thalheim, B. (2020). The triptych of conceptual modeling. Software and Systems Modeling, 1–18.
[28] Wilde, T., & Hess, T. (2006). *Methodenspektrum der Wirtschaftsinformatik: Überblick und Portfoliobildung* (Arbeitsbericht No. 2).
[29] Karagiannis, D., Buchmann, R., Burzynski, P., Reimer, U., & Walch, M. (2016). Fundamental Conceptual Modeling Languages in OMiLAB. *Domain-Specific Conceptual Modeling: Concepts, Methods and Tools*, 3–30. https://doi.org/10.1007/978-3-319-39417-6_1
[30] *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/*

mechanisms as well as desired solution output are defined (e.g. training and test data that enable the training of AI as well as certification or approval of AI results).

3) Configuration: Concrete decision algorithms for execution or calculation (e.g. concrete rules that are executed by rule engines, fuzzy logic that can be interpreted or semantic that can be inferenced) are configured.

This methodology can be applied to different scenarios in FAIRWork such as worker allocation or automated test building. By following a bottom-up approach starting from the OMiLAB experiment in D4.2, the three layers are investigated through creating prototypes for different AI approaches (e.g. rules, fuzzy rules, agents, artificial neural networks etc.). The chosen AI approaches are analysed with the goal to identify underlying rules, conditions and requirements for the three layers.

Relevant research questions that should be analysed with our research approach include:

- What are the capabilities of the available services, and what are their characteristics?
- How can decision models and AI solutions be mapped depending on their strength in a specific situation?
- How can suitable services and offerings be integrated into a legacy system?
- Is a model-based approach suitable to support the continuous management of a decision support system?
- How do we distinguish/define the three layers?
- How can modelling support knowledge engineering (speed up, improved understandability & communication)? (Modelling for AI)
- How do the three layers look for different AI approaches and how can they be connected through meta modelling?

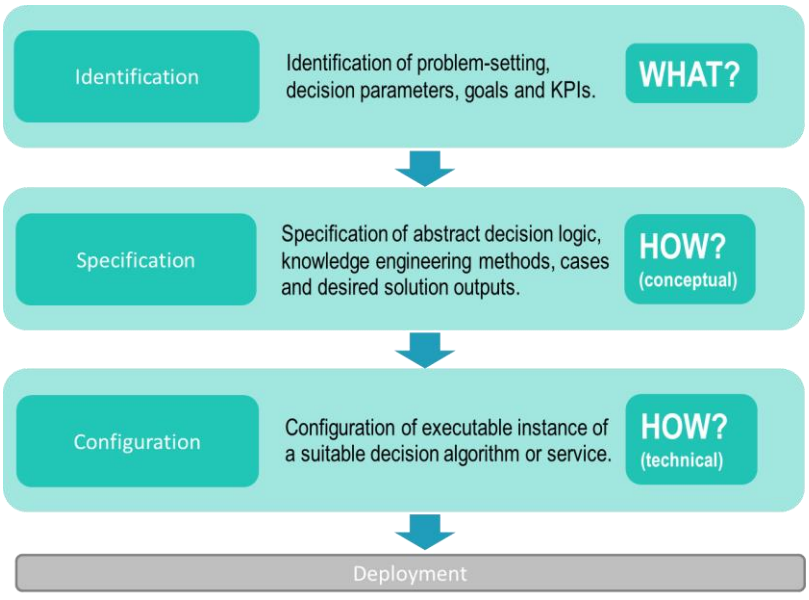


Figure 6: Overview of three-layered approach.

Within this research track, OMiLAB focuses on researching how modelling can be used to support the design of decision support within the DAI-DSS and how information can be exchanged between modelling environments and the DAI-DSS. The goal is to facilitate the reusing of modelled information in a flexible way, by finding a generic and adaptable way to exchange this information, allowing to use it with various modelling methods and decision services. This increases the value of models decreases the manual tasks. In addition, by allowing the flexibly adapt the used modelling methods and available services, the ones can be chosen that best fits the users' needs. Therefore, a semantic rich and flexible way to enable such information exchange is needed.

Additionally, the information exchange must be bi-directionally, meaning that information from the models can be used in the DAI-DSS, for example as input for the configuration. But also, the information exchange in the other direction should be possible, information from the decisions can be used to adapt or create models. In this way, knowledge about concrete decisions within the system can be transferred to models, which can be used to support the representation of made decisions.

The information exchange will be used during FAIRWork's design procedure, introduced in deliverable 2.1(Zeiner, 2023)[31], where knowledge about the decision is collected using different modelling methods and correspond to the three layers introduced in Figure 6. This knowledge is used in the configuration of the DAI-DSS, where the generic information exchange should support.

The research from OMiLAB will use its *Digital Innovation Environment (DIEn)* (Karagiannis et. al., 2022)[32] and its experiment-based approach. This approach itself should also be improved, whereby the focus will be set on investigating how the experiment approach can be adapted to FAIRWork, to allow quick results that can be evaluated with involved stakeholders. Such experiments need a high-level representation of the scenario, an environment where the can be executed and finally a comprehensible representation of the decision knowledge. The research in this context will focus on how such experiments must be structured, to support the exploration of decision problems.

2.8 Reliable and Trustworthy AI

One further component in the concept of **Industry 5.0** concerns a **human-centric approach** to technologies such as AI[33]. That is, AI systems should be designed in a reliable and trustworthy manner, adapted to the human rather than requiring the human to adapt to the system[34,35].

To build trust in AI, two main questions arise for the user: Is the system working well enough? And is it working in a correct way? To shed light onto these two aspects, transparency about performance as well as the inner workings of AI are needed. This is in line with the **Key Requirements towards AI** that have to be fulfilled to be trustworthy, according to the European Commission's high-level expert group on artificial intelligence (AI HLEG 2019a[36], 2019b[37]): They comprise human agency as well as technical robustness, privacy, fairness, and transparency. While most aspects have to be granted from a technological side, communicating their successful execution to the users is a matter of transparency (Arrieta et al., 2020[38]; Felzmann et al., 2019[39]; Mohseni et al., 2021[40]; van Nuenen et al., 2020[41]).

[31] Zeiner, H. (2023). Specification of FAIRWork Use Case and DAI-DSS Prototype Report 2.1.https://fairwork-project.eu/deliverables/D2.1_Specification%20of%20FAIRWork-v1.0a-preliminary.pdf (accessed: 15.02.2024)

32 Karagiannis, D., Buchmann, R. A., & Utz, W. (2022). The OMiLAB Digital Innovation environment: Agile conceptual models to bridge business value with Digital and Physical Twins for Product-Service Systems development. Computers in Industry, 138, 103631. https://doi.org/https://doi.org/10.1016/j.compind.2022.103631

[33]Nahavandi, S. (2019). Industry 5.0—A human-centric solution. Sustainability, 11(16), 4371.

[34]Shneiderman, B. (2020). Human-centered artificial intelligence: Reliable, safe & trustworthy. International Journal of Human–Computer Interaction, 36(6), 495-504.

[35]Shneiderman, B. (2022). Human-centered AI. Oxford University Press.

[36] AI HLEG. (2019a). *Policy and investment recommendations for trustworthy Artificial Intelligence*. https://digital-strategy.ec.europa.eu/en/library/policy-and-investment-recommendations-trustworthy-artificial-intelligence

[37] AI HLEG. (2019b). *Ethics guidelines for trustworthy AI*. European Commission. https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

[38] 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

[39] 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

[40] 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

[41] 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

Therefore, in order to find ways to build trust in AI, we especially focused on the effect of **transparency** on users trust in an AI system. In Deliverable 3.1, we provided a broad overview of the current state of the art for trust in AI systems. Since trust functions as an important prerequisite of technology acceptance, adoption, and use in general (Venkatesh et al., 2016) and for AI in particular (Siau & Wang, 2018).

While developers and AI experts have increased their work on interpretability and explainability, they have been focusing on a technological viewpoint (Arrieta et al., 2020[42]; Rai, 2020[43]; Murdoch et al., 2019[44]). However, the solutions gathered under the term of explainability often are not understandable by the lay end users. That is, researchers such as Paéz (2019)[45] and Miller (2018)[46] requested to turn towards end users. Páez called for research on understandability. According to him, more important than detailed information about AI processes is a holistic investigation on how to make AI really understandable to increase trust. Such requests have since led to a rising number of research on transparency for end users.

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

- How can an AI decision support system be designed in a way to foster trust?
- What does transparency comprise for lay users beyond the technical approach of Explainability?
- What is the effect of transparency on trust, acceptance and usage in the application of production lines and Multi-Agent Systems, i.e. the FAIRWork project?
- How do different AI system factors influence the effect of transparency on trust, acceptance and usage in more general terms?
- Which system factors influence the effect of transparency on trust in the use cases?
- How can transparency be provided in an understandable way for the stakeholders in FAIRWork?

To answer the questions, we work in close collaboration with the FAIRWork use-case partners. During our visit at flex in Althofen, Austria, we gathered insights on factors of trust and transparency, acceptance and usage of an AI-based decision support system. To do this, we conducted **interviews with operators and managers** from FLEX Althofen. In addition, we collected **fears and hopes** regarding an democratic AI decision support system during a workshop that took place at FLEX Timisoara.

The aim is to build a **transparency matrix** that relates system factors with required transparency measures that have to be taken into account to foster trust. With the help of this matrix, the technical developments of FAIRWork can be improved and implemented in a way that puts the user in the center. In the context of FAIRWork, a workshop took place to identify the best ways to add transparency to the different AI-services. This work will continue in the future by closely collaborating with each AI-service, to consult towards further transparency implementations. Eventually, the workshop concept together with respective **instructive material** will be improved to enable computer scientists developing AI services to adhere to transparency measures. Overall, the results of all studies can be applied to systems beyond the FAIRWork project to improve the design and implementation of AI systems in a working context.

To accompany the implementation of FAIRWork services, participants from the FLEX Althofen plant answered several questionnaires on the status quo about different working conditions such as workload, opinion on the current decision making, as well as attitude towards automated systems. This questionnaire study will be continued at the

[42] 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

[43] 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

[44] 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

[45] 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

[46] Miller, A. P. (2018, July 26). Want Less-Biased Decisions? Use Algorithms. *Harvard Business Review.* https://hbr.org/2018/07/want-less-biased-decisions-use-algorithms

CRF plant in Turin. Based on these questionnaires, future technological implementations can be measured in their success in fostering trust.

3 RESEARCH METHODS AND SERVICES

3.1 Overview of methods and services

This Chapter is dedicated to the documentation of the various services that have been developed in the context of the WP3 research collection. For each service, a motivation and a reference to the use cases of the project is given, as well as a state-of-the-art Section, functional description, interfaces, and finally, the outlook on next developments towards the last Deliverable in WP3, i.e., D3.3 (“Final DAI-DSS Research Collection”). In addition, we provide an early account on methods and concepts that currently are in a state of low Technology Readiness Level. Finally, studies are described as well with the relation to its socio-technical settings.

3.2 Methods on Democratization of Decision-Making in Socio-Technical Settings

In order to explore the dynamics of democratization in industry and the contextual situation for implementing a Decision-Support Systems (DSS) within a company, our investigation led us to employ a case study approach (Yin, 2014). The case study focused on a comprehensive analysis of our use case partner FLEX, a manufacturing company, which initially involved verifying operational documentation describing workflow procedures through exploration of specified scenarios. The first scenario, 'Automated test building,' begins with a new product order or the need for a production process update, facilitated by the program manager and forwarded to the automation engineer during the design phase. Another significant scenario, 'Worker allocation,' occurs prior to or during shifts, based on the assessment of workers' abilities and the production plan. Additionally, in the 'Machine maintenance' scenario, triggered by breakdowns or failures during the execution phase, known problems are addressed through a ticketing system.

Our research encompassed first a document analysis, second a visit of various departments within the company, beginning with an examination of company products and production halls, where ongoing tasks such as quality checks, labelling, identification, and evaluation were looked at with the method of participatory observation. Subsequently, attention shifted to the operation of collaborative robots with human involvement. Third, we conducted interviews with employees across different hierarchical positions and levels of experience (N=8); However, this represents only the first round of interviews, with plans for a second wave later this year. Moreover, workshop formats will be used to deepen the insights for the alignment of the results of the sociological analysis towards the development of the Multi-Agent Systems.

It should be acknowledged that the following results are based on the initial phase of our case study. The method of analysing socio-technical practices allows thereby, first the check of the adequateness of the technical representation of the workers’ status, second it focus on the quest of legitimacy: Which aligning social procedures are needed to safe the legitimacy of the application of such a system in a hierarchical organisation.

3.2.1 Insights from Interviews: Qualitative Results

As mentioned above, the evaluation of interview results has been conducted, introducing three dimensions. The first dimension, Decision-Making, has revealed that, decision-making situations can vary, ranging from decisions regarding the production line, such as assessing the authenticity of machine failure, determining when to report recurring errors to the supervisor, and potentially shutting down the line, to addressing stock differences or product shortages affecting production. These situations may also involve attributing fault, whether to internal parties, customers, or transport companies, in cases such as product failures and transport damage, and deciding whether to dismantle material returns. Moreover, decisions may involve personnel matters like overtime management, task

allocation, and determining roles and timing for redundancy. Additionally, managerial and works council decisions may include adjusting budget allocations, managing staffing levels, handling dismissals, and addressing employee issues. Based on statements from interviews, in the decision-making process, employees may face various challenges, with one of the most crucial being feelings of uncertainty regarding decision accuracy. This uncertainty may arise due to a lack of experience or insufficient information about the system, components, customer, employees, and timestamps. Consequently, the uncertainty presents itself differently: experienced individuals tend to accept the consequences, while amateurs often experience significant stress. Uncertain decisions can also jeopardize production line stability, particularly when made amidst disparities. These situations can become even more challenging, especially when superiors or experts are unavailable for consultation or conversely when multiple decision-makers hold differing ideas, resulting in slower progress in the production line. Other issues that may affect the decision-making process include work and time pressure, planning weekend work shifts, decisions with partial fairness, and budget constraints, which can lead to project cancellations. In light of the mentioned situations and challenges, certain qualities and characteristics define the decision-making process within the SME. These include carefulness in decision-making and, notably, consideration of risks, and subsequently accepting their responsibility. In addition to utilizing the skills matrix, which outlines employee task assignments, supervisors should also take into account shift scheduling and workload when making decisions. Considering the priority of tasks is also crucial, as it determines the allocation of responsibilities, whether to highly qualified individuals or to those with lower experience. Based on the interviews, it was also found that most tasks in the company are repetitive, with rare occurrences of new tasks. This could explain why decision-making relies more on experiential knowledge, prompting employees with less experience to frequently seek guidance from their supervisors to ensure the accuracy of their decisions. While decisions are primarily reached collectively through discussions and regular meetings, hierarchical decision-making, particularly at the managerial level, is also observed.

The second dimension, Involvement, has uncovered that suggestions and objections receive primary consideration when they contribute to enhancing production or when failures or problems are identified in the production process. Additionally, while the ideas are conveyed to the chef team through supervisors, annual employee discussions also provide an opportunity for individuals to express their views directly. Moreover, opinions regarding staff assignment planning can be widely acknowledged. Surprisingly, some employees may refrain from offering suggestions or objections, arguing that intervening in superiors' decisions is not expedient, and that everyone should focus solely on their own work, believing this approach leads to a stress-free environment. The responses to transparency questions can be divided into two categories: On one hand, decision-making transparency can be achieved through explanations from experts, addressing uncertainties raised by employees, clarifying new themes and decisions during weekly and monthly meetings, involving all relevant staff in decision-making, and fostering expertise and knowledge among the team over time. On the other hand, the transparency of decisions is context-dependent, and in certain cases, some decisions, particularly those at managerial levels, may not require understanding by others.

The participants emphasised the high level of trust among the team, highlighting that without trust in their team members, progressing in the work process would be impossible. They remarked that this existing trust minimizes the need for negotiation on every subject, thereby accelerating the work process. Direct interaction between team members was cited as a key factor in increasing trust, leading employees to have complete trust in their superiors due to awareness of their knowledge and experiences. Additionally, system functioning contributes to building trust between superiors and employees. However, mismatches between decisions reached and their implementation in the workplace were noted to harm the trust employees have in their superiors.

The final dimension that emerged from the results, Expectation, has revealed the importance of interviewees having access to multiple proposed solutions for various scenarios. Recognizing work and time pressures as key challenges in decision-making, employees express a preference for reduced human intervention and alleviated time pressure through the digitalization of time-consuming and effortful tasks; For instance, the systematic categorization of the severity of various incoming errors would allow employees to easily identify the appropriate

course of action when encountering specific types of errors. This approach would bypass the necessity for extensive time allocation to initially determine the error type, thereby reducing uncertainty in decision-making processes. Additionally, there is a desire for data within the system to be organized in a comprehensible manner, especially considering the vast amount of data across the entire company. Providing self-familiarization tools for both new hires and existing staff with the workflow is also seen as essential for facilitating the work process.

Another highlighted point focuses on relieving employees and reducing workload by delegating risk assessment tasks to the system. Within this context, decision-makers also express a desire to automate fault attribution processes, potentially by establishing specific criteria to streamline the determination of responsibility for particular damages or errors, thereby simplifying decision-making procedures. The expressed expectations encompass points aimed at accelerating the work process. Recognizing staff assignment planning as a particularly time-consuming task, especially on weekends, underscores the necessity of implementing an automated system for assignment planning to assist supervisors, thereby significantly enhancing workflow efficiency and alleviating work-related pressure. Moreover, considering the aforementioned work pressure within the company and the potential for oversight regarding certain tasks and deadlines, employees have proposed the implementation of a signalling reminder system for upcoming plans. Additionally, access to a well-defined procedural framework would reduce challenges in finding experts and supervisors for consultations, thus furthering the acceleration of the work process. Other strategies, such as personally managing vacations and sick leave arrangements, along with task assignments tailored to individual preferences and competencies, serve to increase workflow efficiency. In addition, the implementation of automated customer service functionalities, such as email drafting, has the potential to mitigate time pressure. From a more technical perspective, it is advantageous to establish a structured inspection sequence for prioritizing error, enabling employees to accelerate the identification of the primary issue and address other errors effectively. The fulfilment of these requisites not only fosters punctuality but also enhances staff productivity and creativity. The expectations articulated by the interviewees regarding their anticipated functionalities of this system can significantly contribute to the decision support system's development and ensure the effectiveness of its workflow processes. The suggestions provided underline the critical need for accessing precise and reliable solutions, alongside secure access to all production and employee data. Therefore, establishing a comprehensive database containing all relevant company information is imperative.

Additionally, enabling data and workflow reviews can enhance the accuracy of system data and contribute to the efficiency of the decision-making process. Moreover, it is essential to provide insights into the consequences of proposed solutions, ensuring a thorough understanding of potential outcomes. Indicating resource availability and their precise whereabouts through system would ensure the progression of workflow processes, covering aspects such as product, machinery and even substitute employee availability. From another perspective, employees anticipate that the integration of a decision support system would ensure equitable treatment for all staff members. However, the comprehensiveness of this system should not hinder critical thinking, as creativity and innovation must remain essential.

Finally, there is an emphasis on protecting human discretion, recognising the value of human judgment and autonomy within decision-making frameworks. Particularly, the participants envision the system as a tool that provides guidance, with ultimate decisions resting in human hands.

3.3 Methods and Services for Digital Human Factors Analytics

3.3.1 Overview

The vision is to implement innovative models that can recognise intra-individual changes in these outcomes and relate them to worker allocation so that an automated meaningful attribution of resilient workers to immediate stressful work can be reasonably realised. This Section presents the conceptual framework and an appropriate

method in the context of the FAIRWork project, which aims to develop such models for the benefit of the industrial community. We present a cyclical conceptual framework based on existing theories of stress and resilience and provide adequate novelties. The initial stage of the FAIRWork-based operationalisation of the concepts and the daily measurement cycle are described, including the use of wearable sensor technology (e.g., heart rate, skin temperature, heart rate variability measurements) in the frame of the Intelligent Sensor Box concept (see Deliverable D3.1). Analyses target the development of within-subject and between-subject models – e.g., within a persona-oriented frame and represent a first innovative step within the research track presented in Section 2.3. Future work will focus on further developing these models and eventually explore the effectiveness of the envisioned personalised resilience system.

3.3.2 Service: AI-based Physiological Strain Estimation

3.3.2.1 Motivation and Reference to FAIRWork Use Case

Motivation. One central aspect of the Digital Human Factors Analytics in the FAIRWork project is the estimation of physiological strain and resilience risk from the wearable sensors. This estimation of the risk of a worker's health is highly dependent on the measurement of the heart rate as well as, in particular, on the as precise as possible estimation of the core temperature of the worker. While heart rate measurements are already very precise, the estimation of the core temperature is still a matter of ambitious research. Today the estimation of the core temperature is applied by the values derived from a sensor on the skin of a user and implicitly scaled by means of a final estimate of the physiological strain index PSI[47], i.e., modified physiological heat strain index PSI* (Buller et al., 2008[48]; Cuddy et al., 2013[49]; Seeberg et al., 2013[50]). Previous work has provided nice results based on linear regression-based estimators and investigated in detail the results with respect to various conditions in several laboratory-based and field-based trials.

With this novel service we document the development of the application of nonlinear estimators to estimate the core temperature from skin temperature sensor values. These estimates are then fed into the PSI* functional in order to finally compare errors of linear PSI* results with the errors of the newly attained nonlinear PSI* results. The objective of this task is to get as minimum as possible numerical deviation from the "true" PSI results that are obtained from the core temperature sensors – exclusively collected from digital data of the capsules that are swallowed and provide digital temperature values from within the body – and the rather precise heart rate measurements from Golden Standard wearable sensor equipment.

Reference to FAIRWork use case. The stress estimation of the worker is of central importance in the context of resilience risk stratification. The physiological strain index is particularly relevant for the estimation of stress that is accumulating with respect to the workplace where physiologically intensive tasks have to be accomplished. The risk stratification is less relevant to prevent physiological collapse at the workplace since this should usually not be the use case in an ergonomically standardised work environment. However, persistent physiological strain over longer periods of time can have an impact on the fatigue state, i.e., leading to mental exhaustion, and in this is an important parameter for the overall resilience risk stratification as a key objective in the work of Digital Human Factors Analytics.

[47] Moran DS, Shitzer A, and Pandolf KB. (1998). A physiological strain index to evaluate heat stress. *Am. J. Physiol.* 275, R129-34. Available at: http://ajpregu.physiology.org/content/275/1/R129.abstract.
[48] Buller MJ, Latzka WA, Yokota M, Tharion WJ, and Moran DS. (2008). A real-time heat strain risk classifier using heart rate and skin temperature. *Physiol. Meas*. 29,doi:10.1088/0967-3334/29/12/N01.
[49] Cuddy J S, Buller M, Hailes WS, and Ruby BC. (2013). Skin Temperature and Heart Rate Can Be Used to Estimate Physiological Strain During Exercise in the Heat in a Cohort of Fit and Unfit Males. *Mil. Med.* 178
[50] Seeberg TM, Vardøy ASB, Visser Taklo MM, and Austad HO. (2013). Decision Support for Subjects Exposed to Heat Stress. *IEEE J. Biomed. Heal. Informatics* 17, 402–410. doi:10.1109/JBHI.2013.2245141.

3.3.2.2 Innovation beyond the State-of-the-art

The complexity of the human thermoregulatory models suggests that the responses of core body temperature and physiological measures are part of a dynamical system (Buller et al., 2018[51]). With the use of current physiological monitoring techniques, certain variables (e.g., heart rate and skin temperature) can be readily observed, whereas others, the body core temperature, in particular, can only be readily observed directly in a laboratory setting.

Exploiting knowledge of physiological relationships between variables has led to successful estimation of hidden variables. This type of problem can be represented as a Hidden Markov Model (discrete) or a Kalman filter (continuous; Buller et al., 2013[52]; Buller et al., 2018). These models accommodate the complex time-based relationships of the human thermoregulatory system. Both a Hidden Markov Model and a Kalman filter take the form of recursive algorithms based on Bayesian inference, estimating the state of the system and repeatedly updating that state from the next observation. A Kalman filter is used when the data can be modelled as continuous linear Gaussian distributions, where a Hidden Markov Model is used when the data are modelled as discrete states, each with its own likelihood of occurring.

Table 1: Comparison of models to predict body temperature during exercise in the heat (from Belval, 2016[53]).

Study	Environmental Conditions	Exercise Protocol	Variables	Modeling Technique	Diagnostic Outcome
Xu et al., 2013	25°C, 50%RH; 35°C, 70%RH; 42°C, 25%RH; Army combat uniform with body armor; Laboratory	2h treadmill walking at 350W and 540W	Sternum T_{sk}, Sternum Heat Flux	Linear Regression	R^2=0.75
Niedermann et al., 2013	10°C, 30°C; Laboratory	Treadmill running 40% and 60% VO_{2peak}	HR, Chest Heat Flux, Back Heat Flux, Upper Arm T_{sk}, Lower Arm T_{sk}, Thigh T_{sk}	Principle Component and Linear Regression	RMSE=0.28-0.34°C
Buller et al., 2013	24-35°C, 42-97%RH; Army combat uniform with body armor; Outdoors	24h military field exercise	Heart Rate	Kalman Filter	Bias= -0.003±0.32, RMSE= 0.30±0.13
Kim et al., 2015	29.5 to 25.5°C; firefighter PPE; Laboratory	60 minutes of treadmill walking	Chest and Forehead T_{sk}	Linear Regression	T_{chest}, R^2= 0.826; $T_{forehead}$, R^2= 0.824
Richmond et al., 2015	25°C, 50%RH; 35°C, 35%RH; 40°C, 25%RH; variety of clothing conditions; Laboratory	40 minutes of walking with 20 minutes of rest	Insulated 11-site T_{sk}, microclimate T_{sk}, HR, and work	Bootstrap Regression	R^2= 0.86, SEE=0.27°C
Belval et al., 2016	39.8±1.7°C, 33.4±10.7%RH; Laboratory	Treadmill walking and jogging	T_{neck}, T_{arm}, T_{thigh}, T_{calf}, HR, Incline, RH, BF, BM, Height	Multivariate Adaptive Linear Regression Splines	R^2=0.776 RMSE= 0.428°C

Belval (2016) presented several different Machine Learning methods to utilize in the development of prediction models for internal body temperature during exercise in the heat. For a regression model, he found a multivariate adaptive regression splines model performed best. Furthermore, an overview on related research results was presented (see Table 1). Dolson et al. (2022)[54] provided a systematic review and identified 20 studies representing a total of 25 distinct algorithms to predict the core body temperature using wearable technology. Most of these

[51] Buller, M.J, Welles, A.P., and Friedl, K.E. (2018). Wearable physiological monitoring for human thermal-work strain optimization, *Appl Physiol* 124: 432–441, 2018. doi:10.1152/japplphysiol.00353.2017.

[52] Buller, M.J., Tharion, W.J., et al. (2013). Estimation of human core temperature from sequential heart rate observations. *Physiol Meas* 34: 781–798, 2013. doi:10.1088/0967-3334/34/7/781.

[53] Belval, L.N. (2016). *Prediction of Internal Body Temperature using Machine Learning Models*, Master's Thesis, 902, University of Connecticut, 2016. https://digitalcommons.lib.uconn.edu/gs_theses/902

[54] Dolson, C.M.; Harlow, E.R.; Phelan, D.M.; Gabbett, T.J.; Gaal, B.; McMellen, C.; Geletka, B.J.; Calcei, J.G.; Voos, J.E.; Seshadri, D.R. (2022). Wearable Sensor Technology to Predict Core Body Temperature: A Systematic Review. *Sensors* 2022, 22, 7639. https://doi.org/10.3390/s22197639

algorithms provided Kalman filters for the prediction, few algorithms incorporated individual and environmental data into their core body temperature prediction algorithms, despite the known impact of individual health and situational and environmental factors on the core body temperature (CBT). The RMSE error was found to be between 0.13° Celsius and 1.97° Celsius. Some companies provide information about the performance figures of estimating CBT from wearable sensors. For example, Greenteg AG provides a validated temperature sensor, and reports that it can reach a mean absolute deviation of 0.21° Celsius for daily life and sports.

The presented Machine Learning framework provides a comparison between a large set of Artificial Intelligence methods, including neural network approaches, and finally has determined the minimum RMSE to be found by the Gaussian Process Regression method.

3.3.2.3 Description of Functionality

Rationale of Machine Learning framework. The Physiological Strain Index (PSI) is a well-recognised indicator of physical exertion. In addition to the heart rate also the core body temperature has a significant influence on PSI. However, the "gold standard"[55] measurement method of core body temperature by swallowing a temperature measuring capsule by the test person is very cumbersome (the pill must be swallowed a few hours before the measurements) and cost-intensive. Therefore, for daily measurements we introduce the PSI* indicator, which is based on the skin temperature instead of the core body temperature. The skin temperature values can be measured much more easily laterally at the chest of a subject. To increase the accuracy of the models, we have extended the input parameter set and used the following readily available input features:

- skin temperature [°C]
- heart rate [bpm]
- heart rate variability (HRV, SDNN) [ms] –not yet used in this version
- age [years]
- weight [kg]
- height [m]
- sex [f/m] – not yet used in this version

The output variable represents the objective of the estimation, as follows,

- core temperature [°C] (capsule-based Gold Standard measurement for Ground Truth, i.e., supervised values).

The category sex was currently included in the models but only for possible future use. At present, the parameter has no influence, since only training data from male firefighters were available for training the models. Furthermore, we will apply heart rate variability as a further input feature dimension in future work. The model for the development of the new AI-based estimator for physiological strain indexing was based on the knowledge about human physiology, in particular, physiological strain provided by the University of Léon, Spain, and the expertise on digital technologies including sensor-based AI methodology provided by FAIRWork partner JR. In the future, we plan to incorporate further input parameters into the model, such as, ambient temperature as well as humidity and validate if this could actually improve the results.

3.3.2.4 Interface

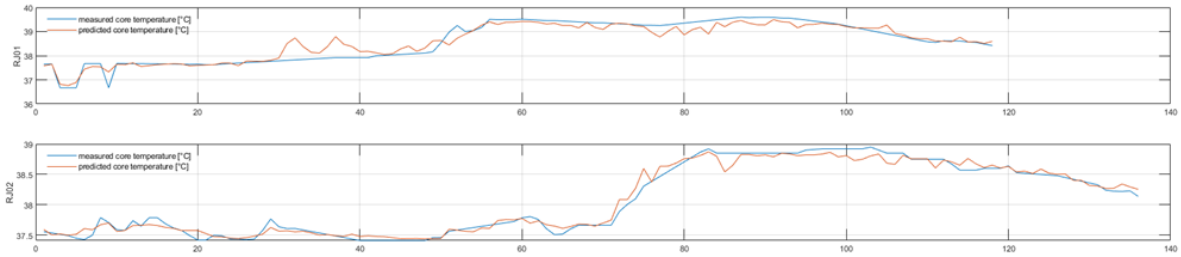
The AI-based Physiological Strain Estimation Service takes all necessary biosignal information of a worker collected by the local wearable sensor network of the Intelligent Sensor Box from its internal data lake. The additional personal metadata of a worker which are required for strain estimation are read from the internal Health Profile.

[55] https://en.wikipedia.org/wiki/Gold_standard_(test)

The final strain estimation is written back to the internal data lake. All interfaces are internal to the Intelligent Sensor Box and are based on a REST API.

Model Type (Kernel/Method)	RMSE (Validation)	MSE (Validation)	RSquared (Validation)	MAE (Validation)
Gauss. Process Regr. (exp.)	**0.279**	**0.078**	**0.812**	**0.181**
Gauss. Process Regr. (rat. Quadric)	0.280	0.079	0.811	0.182
Gauss. Process Regr. (matern 5/2)	0.285	0.081	0.805	0.189
Gauss. Process Regr. (squared exp.)	0.296	0.088	0.789	0.201
SVM (RBF fine)	0.297	0.088	0.788	0.186
Ensemble (bagged trees)	0.310	0.096	0.769	0.222
Tree (fine tree)	0.331	0.110	0.735	0.211
Tree (medium tree)	0.332	0.110	0.735	0.227
SVM (cubic)	0.352	0.124	0.701	0.247
SVM (RBF medium)	0.362	0.131	0.685	0.249
Tree (coarse tree)	0.365	0.133	0.679	0.276
Neural Network (tri 10-10-10 ReLU)	0.377	0.142	0.658	0.274
Neural Network (bi 10-10 ReLU)	0.380	0.144	0.652	0.280
Neural Network (WNN ReLU)	0.391	0.153	0.632	0.291
Least Sqares Kernel Regression	0.394	0.155	0.626	0.297
Neural Network (MNN ReLU)	0.410	0.168	0.595	0.307
Kernel Regression (SVM)	0.412	0.170	0.591	0.283
SVM (quadric)	0.428	0.183	0.559	0.303
Linear Regression	0.433	0.188	0.548	0.322
Stepwise Linear Regression	0.436	0.190	0.542	0.324
Neural Network (NNN ReLU)	0.444	0.197	0.525	0.343
SVM (RBF coarse)	0.455	0.207	0.501	0.345
Linear Regression	0.488	0.238	0.427	0.394
Linear Regression	0.489	0.239	0.425	0.394
SVM (linear)	0.496	0.246	0.408	0.393
Ensemble (boosted trees)	1.655	2.739	-5.596	1.622

(a)



(b)

Figure 7: Results of the RMSE in the regression-based estimation of the core body temperature. (a) Performance results in terms of error between actual core body temperature and estimated temperature using various AI-based and statistical models. The estimation using the Gaussian Progress Regression (top) provided the minimum RMSE value. (b) Actual course of core body temperature (blue line) and learned temperature course by the Gaussian Process Regression Model (red line) based on the data sets of the first five test subjects in the field tests of first responders.

3.3.2.5 Experiments

For the training of the non-linear models we used gold standard data recorded by the University of León during field tests of first responders (Polar H7 Bluetooth chest). We used 12 out of the 13 available data sets from first responders that were recorded during the field tests. One dataset (subject RJ11) was discarded due to many artefacts in the heart rate (HR) and inter-beat interval (IBI)-related signals. In addition, from the available datasets, only the time intervals covered by all necessary bio-signals were used. In some cases, artefacts or non-valid intervals of values were also excluded from the training. The chosen methodology was to evaluated the nonlinear models on two different ways: once with and once without these non-valid signal sections to assess their impact on the overall performance. After data cleaning 971 data sample vectors remained from the 1736 original ones of the 12 subjects to train the nonlinear models. We trained and validated several non-linear models on the database. These models included Linear Regressions, Multilayer Perceptron Networks ("Neuronal Networks"), Gaussian Progress Regression (Rasmussen and Williams, 2006[56]), Support Vector Regression, Kernel Regressions and

[56] Rasmussen, C.E. &, and. Williams, K.I. (2006). *Gaussian Processes for Machine Learning*, MIT Press, 2006, ISBN 026218253X.

Regression Trees. For the training and validation, we resampled the data on a basis of 1-minute time intervals and applied 5-fold cross-validation. Note that we learned the core body temperature and used the predicted value for the subsequent PSI calculation.

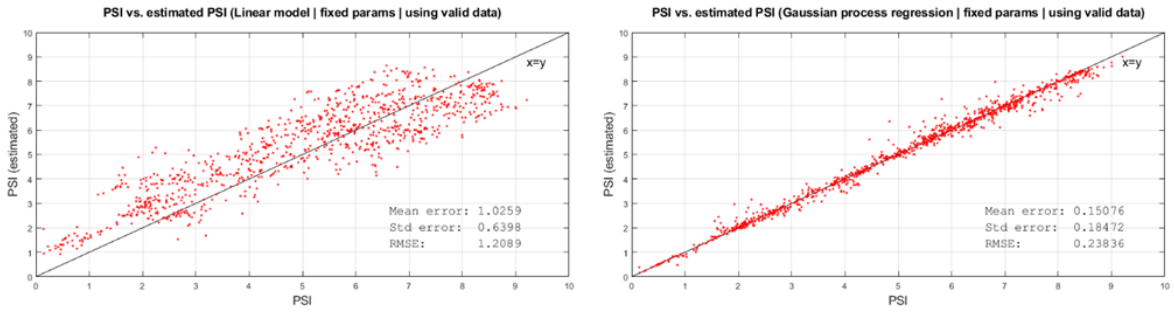


Figure 8: Scatter plots of PSI and fitted/estimated PSI* comparing linear model (left column) with GPR model (right column) on artefact-adjusted data. These plots show results of PSI calculation method using fixed min/max heart rate and skin temperature values; the x=y diagonal represents a theoretically perfect match.

3.3.2.6 Results

From the – in total, 26 different - evaluated machine learning models for nonlinear function approximation we have found that Gaussian Process Regression (Rassmussen and Williams, 2006) with exponential kernel performed best on the used data (see Figure 7a). The validation of the model resulted in a root-mean-square error (RMSE) of 0.279 °C, a mean-squared error (MSE) of 0.078 °C, an R-squared value of 0.812 °C, and a mean-absolute error (MAE) of 0.181 °C. A Gaussian Process Regression (GPR) model is based a nonparametric kernel-based probabilistic attempt. We used the MATLAB implementation of this and all other models that are mentioned in Figure 7a. The best performing kernel on the clean data was the exponential one (SigmaL = 2.4037, SigmaF = 0.7235). We used a constant basis function, an “exact” fitting method and a “random” active set method. The remaining model parameters are Beta = 38.017, Sigma = 0.195 and LogLikelihood = -242.504.

Figure 7 visualises the actual course of core body temperature (blue line) as collected from the field trial as well as in parallel and synchronised the learned temperature course by the Gaussian Process Regression Model (red line) based on the data sets of the first five test subjects in the field tests of first responders. The estimated core body temperature is very close to the ground truth information.

Since we wanted to improve the previous linearly estimated Physiological Strain Index model PSI*, we also compared the PSI* using the predicted core temperature of the Gaussian Process Regression Model (GPR) with the PSI* of the linear model. We uses fixed min/max values for the input parameters heart rate (HR0 = 70 bpm, HRmax = 208 - 0.7 x age [years]) and skin temperature Tchest_0 = 33.5 °C, Tchest_max = 40 °C). For consistency, we also applied both calculation methods for PSI* and compared the performance of the linear model with the learned GPR model for both methods (see Table 2).

Table 2: Various error measures for the comparison between the application of linear and non-linear (i.e., using Gaussian Process Regression) regression models.

method	error on whole data			error on valid data		
	M	SD	RMSE	M	SD	RMSE
PSI vs. PSI*	1.764	2.303	2.900	0.883	0.575	1.054
PSI vs. estimated PSI (linear model)	1.587	1.476	2.167	1.026	0.640	1.209
PSI vs. estimated PSI (GPR model)	0.525	0.755	0.920	0.151	0.185	0.238

Figure 8 shows scatter plots of PSI* and fitted/estimated PSI comparing the linear model (left column) with the GPR model (right column) on artefact-adjusted data samples. These plots show the results of the PSI calculation method using fixed min/max heart rate and skin temperature values, and the x=y diagonal represents a theoretically perfect match.

3.3.2.7 Integration into the DAI-DSS architecture

The integration of this service into the DAI-DSS architecture is applied within the Component of the **Intelligent Sensor Box** (see Deliverable D3.1).

Figure 9 displays a sample session with physiological strain on the treadmill with clearly specified step-wise load program (left) and the Development Monitor (right) with the course of raw data and generated PSI* in real-time visualisation. The PSI* can be generated in real-time on the basis of an AI-based estimate of the core body temperature.

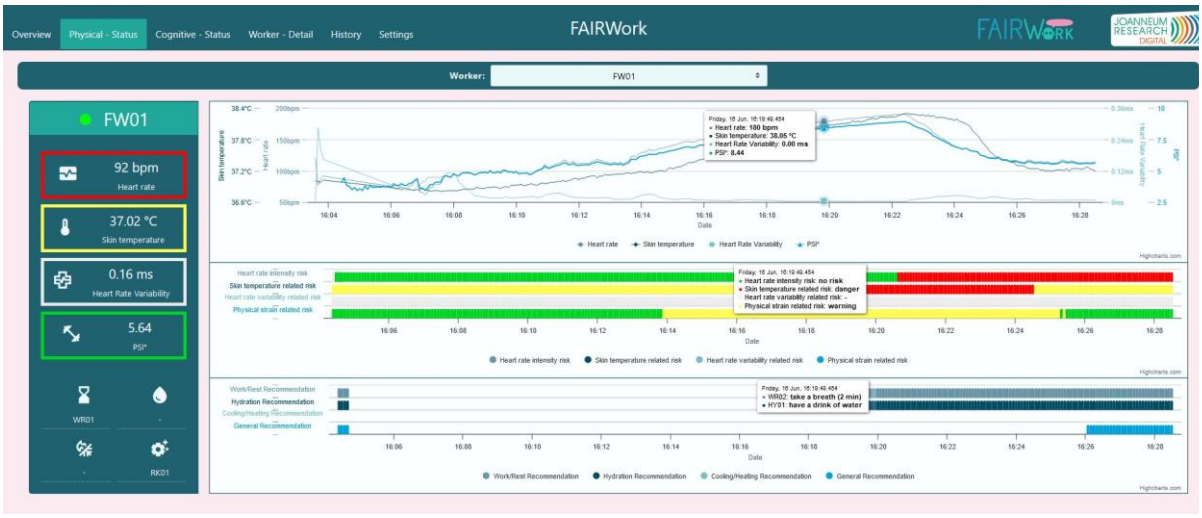


Figure 9: Example of a session with physiological strain on the treadmill with clearly specified step-wise load program and the Development Monitor with the course of raw data and generated PSI* in real-time visualisation.

3.3.3 Service: Heuristic Cognitive-emotional Stress Estimation

3.3.3.1 Motivation and Reference to FAIRWork Use Case

In the FAIRWork project, we need a service for the estimation of a worker's individual cognitive-emotional stress. It should be based on wearable sensors that provide data for stress-related information, such as, the heart rate variability, based on sensors for heart rate and temperature.

3.3.3.2 Innovation beyond the State-of-the-art

The current work is based on a first model provided by Paletta et al. (2022[57]).

[57] Paletta, L., Pszeida, M., Schneeberger, M., Dini, A., Reim, L., Kallus, K.W. (2022). Cognitive-emotional Stress and Risk Stratification of Situational Awareness in Immersive First Responder Training, 2022 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Ioannina, Greece, 2022, pp. 1-4, doi: 10.1109/BHI56158.2022.9926805.

3.3.3.3 Description of functionality

At the current point of development in the project we are using a heuristically defined measure for the estimation of cognitive readiness, i.e., the cognitive-emotional stress score CES_{score}, with

$$CES_{score,t} = \eta * \left\{1 - \frac{HRV_t - HRV_{min}}{HRV_{max} - HRV_{min}}\right\} + \frac{T_{skin,t} - T_{skin,min}}{T_{skin,max} - T_{skin,min}} + \frac{HR_t - HR_{min}}{HR_{max} - HR_{min}}$$

with a pre-defined heuristically selected η=8 according to previous experience.

The CES score exhibits a steep increase exactly where the HRV signal was dramatically increased, reflecting the predominant role of HRV in stress conditions. It is planned to extend the CES score to other input parameters, such as, eye movement features and respiration rate as soon as these are provided and stably delivered by the FAIRWork demonstrator.

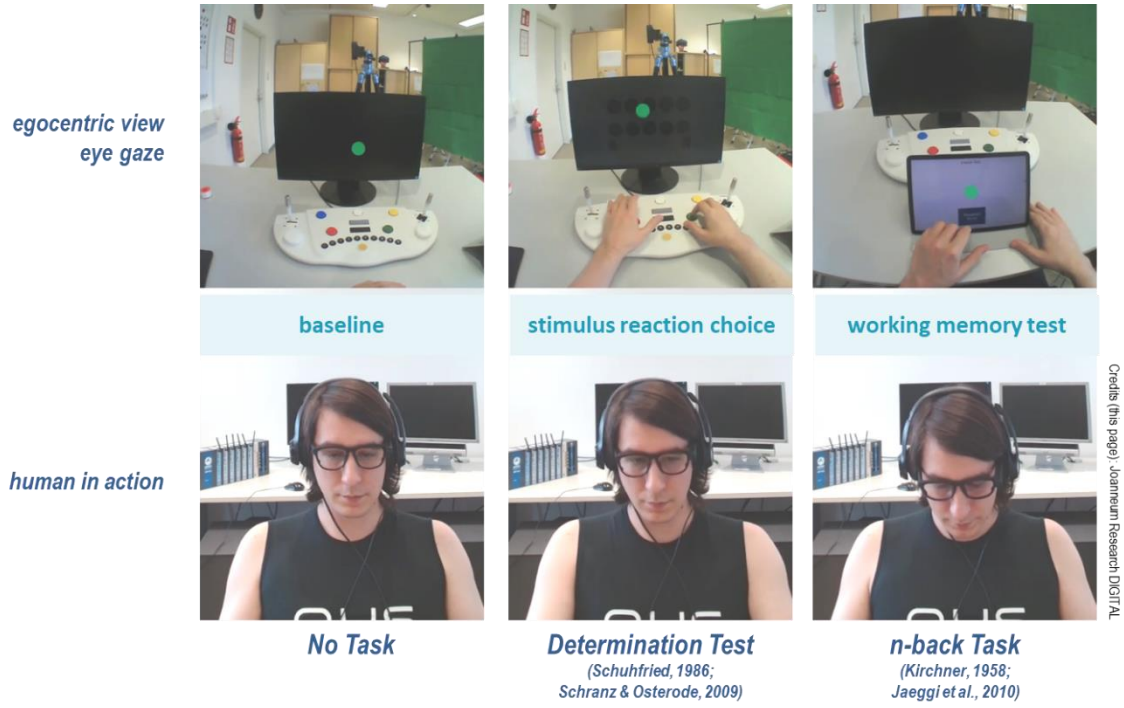


Figure 10: First explorative studies with cognitive-emotional strain we applied a sequence of three tasks. a baseline session without substantial activity, a stimulus reaction choice task called „determination test“, and a task that is known to challenge cognitive load, that is, the n-back task.

3.3.3.4 Experiments

In first explorative studies with cognitive-emotional strain we applied a sequence of three tasks. a baseline session without substantial activity, a stimulus reaction choice task called „determination test“, and a task that is known to challenge cognitive load, that is, the n-back task that requires excellent short-term memory to appropriately react in time to a sequence of images presented to the operator. This working memory task is analysed with an operator video (see Figure 10), being synchronised with the video of the egocentric camera oriented towards the screen with gaze visualisation in real time, and, to the right, real-time raw data output of skin temperature, heart rate, eye tracking based cognitive load score, with a resulting metadata stream represented by a so-far heuristic index for cognitive-emotional strain. below again the synchronised risk stratification in traffic light representation.

Figure 11 shows resulting data from first explorative studies with cognitive-emotional strain: operator video demonstrating the course of raw and processed data about cognitive-emotional stress, with cognitive activities of the operator synchronised with the measurement results.

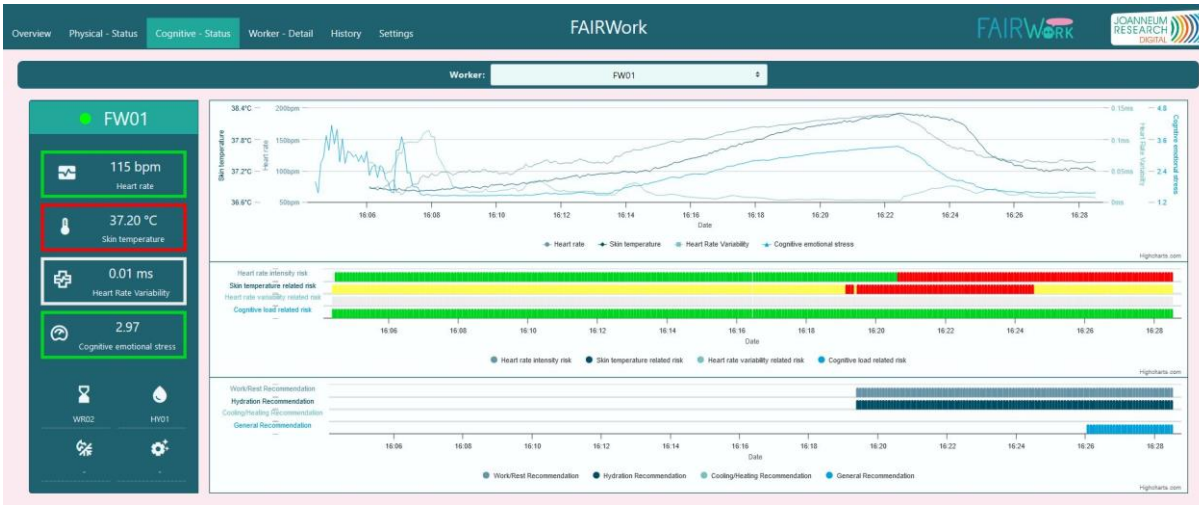


Figure 11: Resulting data from first explorative studies with cognitive-emotional strain: operator video demonstrating the course of raw and processed data about cognitive-emotional stress, with cognitive activities of the operator synchronised with the measurement results**.**

3.3.3.5 Integration into the DAI-DSS Architecture

The integration of this service into the DAI-DSS architecture is applied within the Component of the **Intelligent Sensor Box** (see Deliverable D3.1).

3.3.4 Service: Resilience Score

3.3.4.1 Motivation and Reference to FAIRWork Use Case

Motivation. Estimation of the resilience risk from the wearable sensors is a major aspect of the Digital Human Factors Analytics in the FAIRWork project since it refers to the dynamics that are at the same time crucial for the health of the worker but also substantial for economical quantities of the manufacturing company. Occupational stress can cause health problems, productivity loss or absenteeism. A substantial level of resilience capacity can enable the worker to positively encounter some adversities and to prevent the negative consequences of occupational stress. Due to advances in wearable sensor technology, relatively unobtrusive self-monitoring of resilience-related risk stratification should become possible.

The conceptual framework of the resilience risk stratification model (RRSM) is presented in Figure 12. 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, 2014[58]), the Job Demands-Resources Model of Burnout (Bakker & Demerouti, 2007[59]), the Effort-

[58] Lazarus RS, Folkman S. Transactional theory and research on emotions and coping. *Eur J Pers*. 1987;1:141–169. doi:10.1002/per.2410010304
[59] Bakker AB, Demerouti E. The Job Demands-Resources model: state of the art. *Journal of Managerial Psychology*. 2007; 22:309–328. doi:10.1108/02683940710733115.

Recovery Model (van Veldhoven, 2008[60]) and the Conservation of Resources Theory (Hobfoll, 2001[61]), as well as the WearMe project (deVries et al., 2019[62]).

Strain accumulates when (job) demands, such as time pressure or physical workload, are appraised as threats due to inefficient available resources to adaptively cope with them (Lazarus & Folkman, 2014). In our work on RRSM, we are estimating the physiological strain index PSI* as well as the cognitive-emotional strain (CES). Based on the threat of fundamental strain, an individual's **need for recovery**, characterised by feelings of exhaustion and reduced vigor to undertake new activities, depends on the individual's ability to utilise the available resources to adaptively cope with the demands (Lazarus & Folkman, 2014; Bakker & Demerouti, 2007). A high need for recovery (i.e., little vigor to undertake activities), has a negative impact on an individual's resources to appraise and cope with new demands, such as, a demanding work that should be allocated to workers. However, if there is sufficient recovery to counteract and alleviate this effect (van Veldhoven, 2008).

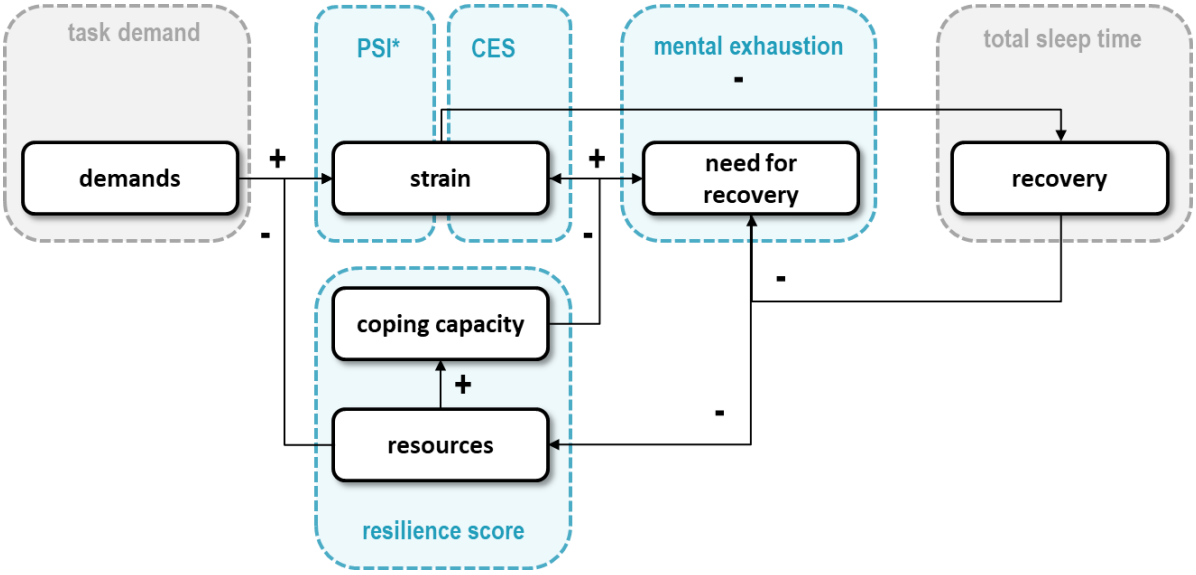


Figure 12: The resilience risk stratification model (RRSM) as proposed to provide resilience scores for the decision support for the manager to decide on worker allocation. The specific contributions to this model is the accumulation of physiological and cognitive-emotional strain (PSI*, CES) to determine an integrated score representing the need for recovery and some degree of mental exhaustion. We model the relation of strain to determine a resilience score that both represents, resources and coping capacity to be able to master upcoming stressful challenges in the work environment.

In our specific RRSM model, we model a unity of mental exhaustion in terms of the daily total strain as a function of both the PSI* and the CES data collection (see Figure 13). The accumulating effect of mental exhaustion is then represented by another functional that integrates these daily contributions over a recent time window. The resilience score that would represent the risk stratification as it is modelled at this stage is then further outlined by an inverse function of the mental exhaustion and should represent merely an orientation of the long-term resilience dynamics than a short-term sample.

The RRSM framework includes also a cyclical nature that is supported by the Conservation of Resources theory (Hobfoll, 2001), which states that initial loss of resources increases one's vulnerability to stress. Since additional resources are necessary to battle stress, this may lead to a depletion of resources or a loss spiral. The motivation

[60] van Veldhoven MJPM (2008). *Need for recovery after work: An overview of construct, measurement and research* [Internet]. Houdmont J, Leka S, editors. 3. Nottingham University Press; 2008. Available from: https://research.tilburguniversity.edu/en/publications/bbdeef64-c338-4d48-b93f-152afe2ac5ef
[61] Hobfoll SE (2001). The Influence of Culture, Community, and the Nested-Self in the Stress Process: Advancing Conservation of Resources Theory. *Applied Psychology*. 2001;50:337–421. doi:10.1111/1464-0597.00062.
[62] de Vries H, Kamphuis W, Oldenhuis H, van der Schans C, Sanderman R. Modelling employee resilience using wearables and apps: a conceptual framework and research design. *International Journal on Advances in Life Sciences*. 2019;11:110–117.

of the development of this RRSM framework is to enable to prevent this loss spiral for the benefit of the worker as well as the economic impact of the manufacturing company.

Reference to FAIRWork use case. The resilience risk stratification 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 in 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 complete economic situation of the manufacturing company. Finally, these scores can provide a relevant input to optimisation routines that would provide higher long-term benefits to the worker, to the company and ecologically relevant aspects.

3.3.4.2 Innovation beyond the State-of-the-art

Work-related stress usually occurs when the demand exceeds the worker's capacity to perform (Wegner, 1988[63]). Exposure to stress has been shown to be related to adverse effects in the way people feel, think, and behave (Griffiths, 1995[64]), and generally, it is demonstrated to have psychological consequences on workers, such as a negative emotional state of anxiety and frustration (Brunzini et al., 2021[65]). At the physiological level, it can alter unconscious vital processes, such as heart and breathing activity, whereas from the physical point of view, it affects natural posture and body activity (Brunzini et al., 2021). Industry 5.0, as a new human-centred perspective, focuses on the role of workers in the current revolution, examining the new industrial paradigm by putting human workers at the centre of production processes and ensuring that technology adapts to their requirements (Yeow et al., 2014[66]). However, stress has further consequences on production activity due to the positive correlation with errors and periods of distraction at work, reducing the quality and performance of the worker (Zizic et al., 2022[67]) and leading to new costs and losses for companies. Given the several consequences of stress on human health and companies' efficiency, the necessity of studies that focus on the stress phenomenon related to smart and intelligent manufacturing systems emerges from the literature, suggesting appropriate indicators for stress evaluation in order to support the advancement of research in this field.

Blandino (2023[68]) provides a review on the measurement technologies on stress in smart and intelligent manufacturing systems. This review identifies and summarises a growing body of literature that recognises the importance of human-centred manufacturing systems (Wang et al., 2020[69]; Nguyen Ngoc et al., 2022[70]) and the consequent human factors, especially workload, physical and mental fatigue (Villani et al.,2019[71]), and ergonomics (e.g., Stefana et al. 2022[72]) and related indicators (Argyle et al., 2023[73]; Digiesi et al., 2020[74]). From the psychological perspective, studies review traditional standard questionnaires in order to adapt them to new

[63] Wegner, D.M. Stress and mental control. In Handbook of Life Stress, Cognition and Health; Fisher, S., Reason, J., Eds.; JohnWiley & Sons Ltd.: Hoboken, NJ, USA, 1988; pp. 683–697.
[64] Cox, T.; Griffiths, A. Work-related stress: Nature and assessment. In IEE Colloquium on Stress and Mistake-Making in the Operational Workplace; IET: London, UK, 1995.
[65] Brunzini, A.; Peruzzini, M.; Grandi, F.; Khamaisi, R.K.; Pellicciari, M. A preliminary experimental study on the workers' workload assessment to design industrial products and processes. Appl. Sci. 2021, 11, 12066.
[66] Yeow, J.A.; Ng, P.K.; Tan, K.S.; Chin, T.S.; Lim, W.Y. Effects of stress, repetition, fatigue and work environment on human error in manufacturing industries. J. Appl. Sci. 2014, 14, 3464–3471.
[67] Zizic, M.C.; Mladineo, M.; Gjeldum, N.; Celent, L. From Industry 4.0 towards Industry 5.0: A Review and Analysis of Paradigm Shift for the People, Organization and Technology. Energies 2022, 15, 5221.
[68] Blandino G. How to Measure Stress in Smart and Intelligent Manufacturing Systems: A Systematic Review. *Systems*. 2023; 11(4):167. https://doi.org/10.3390/systems11040167
[69] Wang, B.; Xue, Y.; Yan, J.; Yang, X.; Zhou, Y. Human-Centered Intelligent Manufacturing: Overview and Perspectives. Chin. J. Eng. Sci. 2020, 22, 139.
[70] Nguyen Ngoc, H.; Lasa, G.; Iriarte, I. Human-centred design in industry 4.0: Case study review and opportunities for future research. J. Intell. Manuf. 2022, 33, 35–76.
[71] Villani, V.; Gabbi, M.; Sabattini, L. Promoting operator's wellbeing in Industry 5.0: Detecting mental and physical fatigue. In Proceedings of the 2022 IEEE International Conference on Systems, Man, and Cybernetics (SMC), Prague, Czech Republic, 9–12 October 2022; pp. 2030–2036.
[72] Stefana, E.; Marciano, F.; Rossi, D.; Cocca, P.; Tomasoni, G.Wearable Devices for Ergonomics: A Systematic Literature Review. Sensors 2021, 21, 777.
[73] Argyle, E.M.; Marinescu, A.; Wilson, M.L.; Lawson, G.; Sharples, S. Physiological indicators of task demand, fatigue, and cognition in future digital manufacturing environments. Int. J. Hum. Comput. Stud. 2021, 145, 102522.
[74] Digiesi, S.; Manghisi, V.M.; Facchini, F.; Klose, E.M.; Foglia, M.M.; Mummolo, C. Heart rate variability based assessment of cognitive workload in smart operators. Manag. Prod. Eng. Rev. 2020, 11, 56–64.

manufacturing contexts. For example, Lesage et al. (2012[75]) focused on the properties of the Perceived Stress Scale. On the physiological perspective, the literature includes significant studies, such as that of Leone et al. (2020[76]), who proposed a multi-sensor platform for monitoring stress in manufacturing contexts; that of Han et al. (2017[77]), who designed a wearable device for the detection of work-related stress; and that of Setz et al. (2009[78], who described a wearable device for discriminating the phenomenon of stress from the cognitive load. On the other hand, Khamaisi et al. (2022[79]) proposed strategies for identifying potential causes of stress for workers, which may be induced by collaboration with robots, as explored by Arai et al. (2010[80]). The main gaps in the literature are due to the limited investigation of the stress phenomenon with respect to the other human factors investigated and to the variety of stress-evaluation methods that lack homogeneity. In addition, the dynamics of innovative technologies in working contexts lead to changes in the production tasks that, in combination with other factors, such as environmental factors and workers' demographic profile, affect the potential sources of stress. These need to be analysed and evaluated by comparing different potential stress-measurement methods in order to develop solutions that reduce stress sources and, at the same time, increase companies' productivity and efficiency.

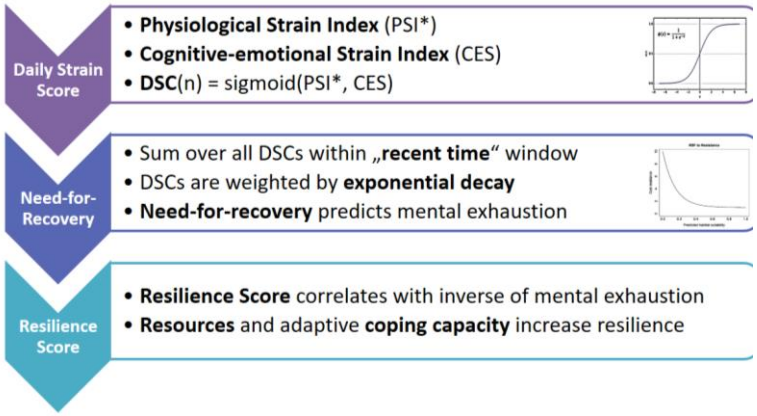


Figure 13: Stages of the computation of the resilience score that underlies the risk stratification model.

One of the comparably rare research work on wearable sensing of stress and resilience was provided by Adler et al. (2021[81]) in which a system was created to find indicators of resilience using passive wearable sensors (Fitbit armband) and smartphone-delivered ecological momentary assessment (EMA). This system that was specialised on the workplace of care professionals (resident physicians) identified resilience indicators associated with physical activity (step count), sleeping behaviour, reduced heart rate, increased mood, and reduced mood variability. deVries et al. (2019[82]) presented a framework for the integration of stress and resilience of employees that was initially based on questionnaires, EMA as well as wearable monitoring. In this wider context, Dunghana et al. (2021[83])

[75] Lesage, F.X.; Berjot, S.; Deschamps, F. Psychometric properties of the French versions of the perceived stress scale. Int. J. Occup. Med. Environ. Health 2012, 25, 178–184.
[76] Leone, A.; Rescio, G.; Siciliano, P.; Papetti, A.; Brunzini, A.; Germani, M. Multi sensors platform for stress monitoring of workers in smart manufacturing context. In Proceedings of the 2020 IEEE International Instrumentation and Measurement Technology Conference (I2MTC), Dubrovnik, Croatia, 25–28 May 2020; pp. 1–5.
[77] Han, L.; Zhang, Q.; Chen, X.; Zhan, Q.; Yang, T.; Zhao, Z. Detecting work-related stress with a wearable device. Comput. Ind. 2017, 90, 42–49.
[78] Setz, C.; Arnrich, B.; Schumm, J.; La Marca, R.; Tröster, G.; Ehlert, U. Discriminating stress from cognitive load using a wearable EDA device. IEEE Trans. Inf. Technol. Biomed. 2009, 14, 410–417.
[79] Khamaisi, R.K.; Brunzini, A.; Grandi, F.; Peruzzini, M.; Pellicciari, M. UX assessment strategy to identify potential stressful conditions for workers. Robot. Comput. -Integr. Manuf. 2022, 78, 102403.
[80] Arai, T.; Kato, R.; Fujita, M. Assessment of operator stress induced by robot collaboration in assembly. CIRP Ann. Manuf. Technol. 2010, 59, 5–8.
[81] Adler DA, Tseng VW, Qi G, Scarpa J, Sen S, Choudhury T. Identifying Mobile Sensing Indicators of Stress-Resilience. Proc ACM Interact Mob Wearable Ubiquitous Technol. 2021 Jun;5(2):51. doi: 10.1145/3463528. Epub 2021 Jun 24. PMID: 35445162; PMCID: PMC9017954.
[82] de Vries H, Kamphuis W, Oldenhuis H, van der Schans C, Sanderman R. Modelling employee resilience using wearables and apps: a conceptual framework and research design. *International Journal on Advances in Life Sciences*. 2019;11:110–117.
[83] Dhungana, D., Haselböck, A., Schmidbauer, C., Taupe, R., & Wallner, S. (2021). Enabling Resilient Production Through Adaptive Human-Machine Task Sharing. In A.-L. Andersen, Andersen Rasmus, T. D. Brunoe, M. S. S. Larsen, K. Hansen, A. Napoleone, & S. Kjeldgaard (Eds.), *Towards Sustainable Customization: Bridging Smart Products and Manufacturing Systems* (pp. 198–206). Springer Nature Switzerland AG. https://doi.org/10.1007/978-3-030-90700-6_22

presented a concept for flexible production planning that incorporates human workers and investigates different scenarios of task allocation between humans and machines and their impact on production workflows.

The innovative contribution in the FAIRWork project focusses in particular on the estimation of human resilience as a functional of stress monitoring, especially in the industrial environment of the specified use cases (e.g., worker allocation). In this context, we present an initial stage of a model that would be extended on the basis of further research by means of digital Human Factors analytics. Furthermore, this model will include further sensors, such as, the smart-shirt as well as the eye tracking glasses, for further refinement on the basis of multi-sensor based assessment of resilience.

3.3.4.3 Description of Functionality

Figure 13 represents the major processing stages of the resilience risk stratification model. In the first stage the **Daily Strain Score** (DSC) is produced. This score integrates contributions from the Physiological Strain Index PSI* as well as from the Cognitive-Emotional Stress (CES) score into DSC(n) of day n, squashed by the Sigmoid function to always be within the interval [0,1].

In a further step, the DSC(n) components of a pre-defined time window – currently, 20 working days – are integrated in a further formula that downscales more distant DSC(n) with an exponential decay function with τ being another time window for having larger weights for the recent 10 working days. A weighted and normalised sum would then represent an equivalent of a score for potential aspects of mental exhaustion, or, the need for recovery (i.e., NFR).

Finally, the **resilience score RS** is computed, in a first degree of estimation, as RS = 1.0-NFR, i.e., representing the resources that would be reduced by the size of the NFR outcome by a linear scale. Resilience risk stratification will be provided by means of specific thresholds that will be determined in a later stage of the project, concerning input of experts from the industry and from health psychology of JR.

Figure 14 represents the **Resilience Monitor** as the component that provides a visualisation of the development of strain as well as resilience scores over time, i.e., the chosen time window of working days. In this first stage of the development, we refer to a time window of 20 working days, without consideration of recovery from weekends and vacation, however, this will be complemented in future work. In the top sub-window of Figure 14 the sample collection of PSI and CES scores is represented, for the example of decreasing stress figures. Conversely, the resilience score is increasing over time with a certain inertia and delay.

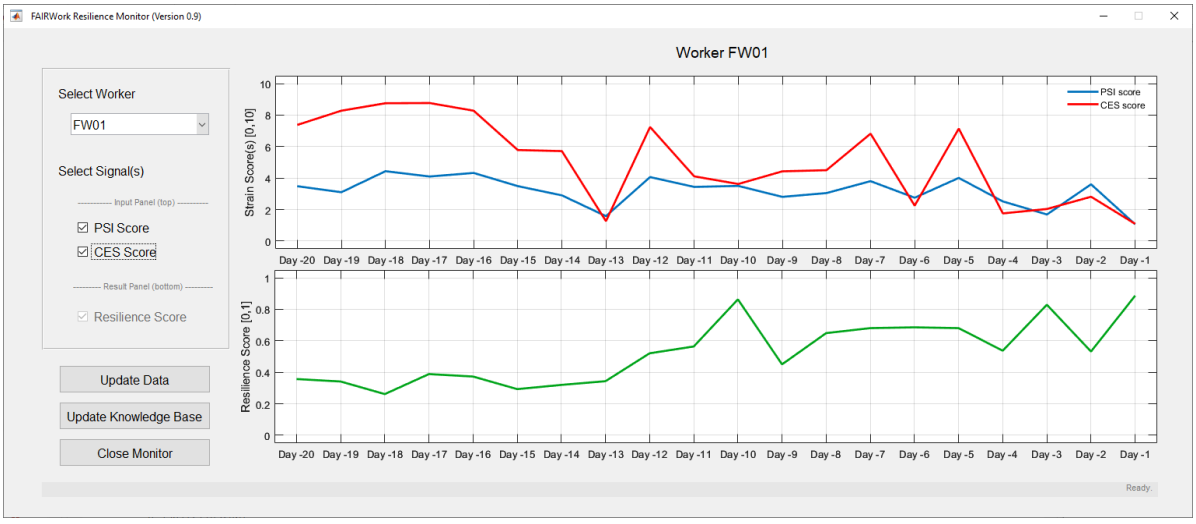


Figure 14: FAIRWork Resilience Monitor.

The individual quantities that are associated with each day are computed from the daily strain score of PSI and CES. Currently, we are using mean descriptors of the PSI as well as the CES score, respectively, from measurements of experimental work over time intervals of 10-30 minutes. Figure 15 provides a characteristic example of the generation of PSI* and CES scores as basic data for the computation of the resilience score in the JR Human Factors Lab, Graz, Austria. The resulting strain scores of limited-time experimental sessions are finally mapped to a Daily Score Score (i.e., DSC(n)) of a specific day n.

Figure 15: Generation of PSI* and CES scores as basic data for the computation of the resilience score (JR Human Factors Lab, Graz, Austria). The resulting strain scores of limited-time experimental sessions are finally mapped to a Daily Score Score (i.e., *DSC(n)*) of a specific day *n*.

3.3.4.4 Interfaces

The Resilience Risk Stratification Service reads the necessary PSI* and CES values from the Intelligent Sensor Box internal data lake (CDW). The data lake is based on MongoDB and is accessed via a REST interface. After the calculation, the processed resilience data is written back to the dedicated resilience table in the data lake. Each time the table is changed, it is automatically synchronized with the corresponding resilience table in the FAIRWork knowledge base. The resilience calculation is currently triggered manually in the FAIRWork Resilience Monitor GUI (see Figure 14) and will also be performed automatically at the end of each (working) day in the future.

3.3.4.5 Study Plan

A study in the factory of CRF (Stellantis) is planned (project month 20-21) to monitor the stress of workers directly during their tasks in the factory. Furthermore, we intend to apply a focused study on cognitive-emotional strain assessment in the Human Factors Laboratory at JR (month 22-23).

3.3.4.6 Integration into the DAI-DSS Architecture

The integration of this service into the DAI-DSS architecture is applied within the Component of the **Intelligent Sensor Box** (see Deliverable D3.1).

3.3.5 Concept: Persona-based Representation of Human Digital Twin

A manager as well as a worker generates Human Factors relevant data that are then converted into cognitive, affective and motivational features, and related to data that are relevant to performance optimisation, in the context of specific activities, be it tasks with a specific machine, or tasks during human-robot interaction.

These experiences are then clustered with an unsupervised strategy in order to develop evidence in terms of data clusters, populations, we call them personas that provide specific characteristics – for example: old, experienced people that should not be stressed too much, or young people that could provide maximum performance on the short term - in the overall distribution.

The Human Digital Twin finally is represented by a decision-making framework that is based on that persona information and the results can finally retroact to these persons in a feedback loop (see Figure 16). Digital Human Factors data are captured and brought into context with the workplace environment. The accumulation of individual behaviour profiles is then clustered into characteristic schemata that represent certain typical "persona" behaviours. Decision-making structures on these representations is then applied to provide output data of the Persona Digital Twin and hence feeds back to all individual actors through overall human-system interaction.

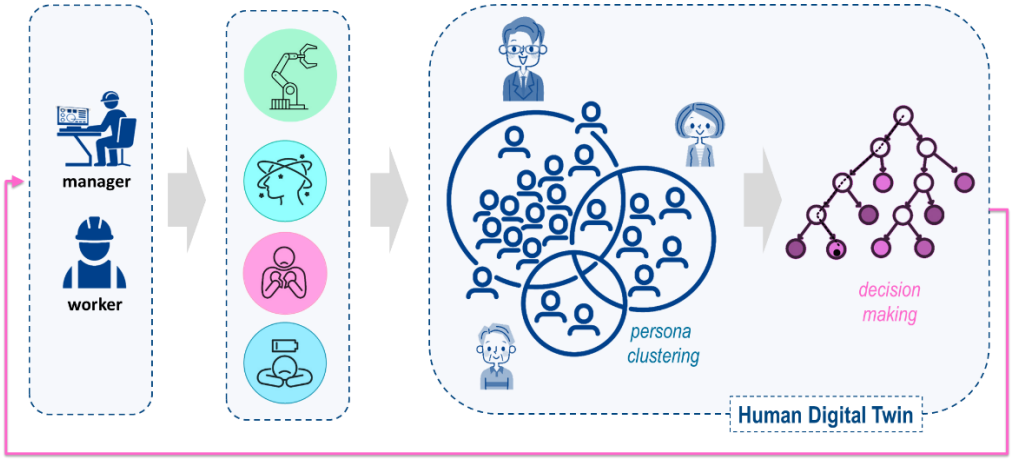


Figure 16: Sketch of the adaptive DAI-DSS persona framework (Paletta et al., 2023[84]).

3.3.6 Outlook

As a basis for the outlook, we have laid down several substantial components of the Intelligent Sensor Box. In a next step, we will have a wearable biosignal study at the production site at Stellantis, Mirafiori, Torino. A refinement of the cognitive-emotional strain model based on a study at the Human Factors Lab will follow. Based on these studies we will be able to refine and finalise the FAIRWork resilience risk Stratification Service. Together with MORE we will outline the persona-based model for digital twin application and also cooperate on the optimisation of workforce assignment and this all should eventually lead to the reporting of analytics and conclusions within the final Deliverable version D3.3.

[84] 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 (2023) International Conference. AHFE Open Access, vol 73. AHFE International, USA. http://doi.org/10.54941/ahfe1003971

3.4 Methods and Services for Optimisation in Decision Support Systems

3.4.1 Overview

The selected services have a clear optimisation goal and are based on a use case scenario, which has already been described in detailed in WP2 and WP5.

The problem definition identified from the practical application contains various individual problems, all of which belong to the area of optimisation. This deals with different decision support problems. The scientific treatment of the optimisation problems in this project can be divided into three stages:

Modelling: This involves generalising and abstracting a practical decision situation and representing it in a model using suitable variables and constraints. In most cases, different approaches with very different variables are possible. However, the choice of model has a decisive influence on the practical solvability, so that modelling is a central scientific achievement in discrete optimisation. For example, binary assignment variables between orders (e.g. material) are suitable for the disposition of the individual orders.

Algorithmic solution of the model: A distinction must be made here between exact methods, which provide an optimal solution under all environmental conditions, and approximate methods, which usually find good but unprovable optimal solutions. The optimisation problems in our project are all in the complexity class of NP-hard problems, which is why exact methods are hardly suitable for practical use. The development of approximate methods opens up a wide range of possibilities for constructive algorithms or a combination of several search strategies.

A trade-off arises as to whether the main focus (=computing time) should be placed on the individual construction steps of the constructive algorithm, or whether simple assignments should be constructed quickly and more effort (= iterations) should be put into improving them. Many approaches in the literature favour the latter method, in the hope of obtaining good solutions from simple sub-steps by performing a large number of iterations quickly. We favour the former approach and will try to achieve this through the targeted development of individual steps.

Validation: In order to check the quality and efficiency of the strategies (e.g. algorithms), it is necessary to create an automated test environment that enables the generation of representative test data and their simulated execution using the developed service or algorithm.

As an example, we describe the "Automated Test Building Support with Hybrid Filtering" service. In this service, we want to identify comparable projects. During the identification process, order details, keywords and assembly plans are taken into account. The similar projects are identified on the basis of contextual nearest neighbours. In the second service "Mathematical Optimisation/Heuristics for Worker Assignment" we find similar workers and assign them to the process.

3.4.2 Method: Automated Test Building Support with Hybrid Filtering

3.4.2.1 Motivation and Reference to FAIRWork Use Case

Originally developed with the Automated Test Build Use Case Scenario as its foundation, this service can be extended to accommodate other scenarios where filtering among pre-existing items is necessary. By identifying comparable and successfully completed product orders, this service supports automation engineers in the creation or adaption of product processes, serving as a guiding framework during the design phase for the robot cell. Comparable projects are identified using order details (e.g. required cycle time, quantity, customer details), keywords, work stages or assembly plans.

3.4.2.2 Innovation beyond the State-of-the-art

Information filtering recommender systems are widely used in various industries. Currently the most used methods include collaborative filtering, content-based filtering and hybrid filtering[85].

- *Collaborative filtering* recommends items based on the preferences of similar users.
- *Content-based filtering* recommends items based on the similarity of previously approved items.
- *Hybrid filtering* combines one or more (sub) approaches to recommend items.[86]

Since we want to identify similar products based on successfully completed product orders, content-based filtering is the pursued approach. Content-based filtering is usually implemented via heuristic methods or classification algorithms, such as rule induction, nearest neighbours, linear classifiers or probabilistic methods.[87] Items are typically represented as vectors. Their similarity can be calculated using either cosine similarity, Euclidian distance or Pearson's correlation.[88]

Another challenge is the non-homogenous nature of the input data. The search may contain textual and visual information, wherefore textual and visual features need to be extracted, similarly to Zhou et al[89]'s fake news detector. Information on textual data, such as keywords or text, can be extracted in multiple ways. Term Frequency-Inverse Document Frequency (TF-IDF) is used to measure the importance of words in descriptions, which are then used to compare the similarity of texts. Natural Language Processing (NLP) techniques, such as sentiment analysis, topic modelling or keyword extraction, are used to understand natural language and analyse it through the extraction of concepts, entities and keywords[90]. A way to discover hidden semantic structures in descriptions is by using semantic analysis[9192].

Besides deep learning- and reinforcement learning-based approaches another method that gains popularity is context-based filtering. Context-based filtering aims to include contextual data into the recommendation, such as time and location[93].

The service is deployed as a recommender system that follows a hybrid filtering approach. We combine content-based and context-based filtering to produce recommendations that are more accurate by considering the currently available machinery, equipment and resources. A downsizing of for example machinery, equipment or personnel previously available may invalidate the recommendation, wherefore a situational context is needed during the decision making process.

[85] Ko, H., Lee, S., Park, Y. and Choi, A. (2022) A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. Electronics 2022: 141. doi:https://doi.org/10.3390/electronics11010141
[86] Poonam, B.T., Goudar, R. and Barve, S. (2015). Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System, International Journal of Computer Applications 110: 31-36. doi:10.5120/19308-0760.
[87] Poonam, B.T., Goudar, R. and Barve, S. (2015). Survey on Collaborative Filtering, Content-based Filtering and Hybrid Recommendation System, International Journal of Computer Applications 110: 31-36. doi:10.5120/19308-0760.
[88] Singh, R., Maurya, S., Tripathi, T., Narula, T. and Srivastav, G. (2020). Movie Recommendation System using Cosine Similarity and KNN. International Journal of Engineering and Advanced Technology. 9. 2249-8958. doi:10.35940/ijeat.E9666.069520.
[89] Zhou, X., Wu, J. and Zafarani, R. (2020). Similarity-Aware Multi-modal Fake News Detection. In: Advances in Knowledge Discovery and Data Mining. PAKDD 2020. Lecture Notes in Computer Science, Vol 12085. Springer, Cham. https://doi.org/10.1007/978-3-030-47436-2_27
[90] Khurana, D., Koli, A., Khatter, K. et al. (2023). Natural language processing: state of the art, current trends and challenges. Multimed Tools Appl 82, 3713–3744 2023. doi: https://doi.org/10.1007/s11042-022-13428-4
91 Ko, H., Lee, S., Park, Y. and Choi, A. (2022). A Survey of Recommendation Systems: Recommendation Models, Techniques, and Application Fields. Electronics 2022: 141. doi:https://doi.org/10.3390/electronics11010141
92 Polčicová, G. and Návrat, P. (2002). Semantic Similarity in Content-Based Filtering. In: Advances in Databases and Information Systems. ADBIS 2002. Lecture Notes in Computer Science, vol 2435. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-45710-0_7
[93] Javed, U., Shaukat, K., Hameed, I., Iqbal, F., Mahboob Alam, T. and Luo, S. (2021). A Review of Content-Based and Context-Based Recommendation Systems. International Journal of Emerging Technologies in Learning (iJET), 16(3): 274-306. Kassel, Germany: International Journal of Emerging Technology in Learning.

3.4.2.3 Description of Functionality

The service retrieves data from the Knowledge Base and compares available entries to the user-defined input parameters, such as order details, keywords, work stages or assembly plans. This reduces the number of references by only presenting relevant projects. The algorithm is run during the planning and design phases of the product processes or, if required, to adapt existing processes. This pre-selection aids the automation engineer by providing possible configurations and guidance on feasibility.

3.4.2.4 Interfaces

The interface for this service will allow engineers to define several types of search parameters, such as order details, keywords, descriptions, as well as the upload of work stages or assembly plans. These inputs will be provided via text fields, dropdown menus or checkboxes. Optionally date ranges could also be included, to disregard older projects. The results will be displayed in a ranked list, sorted by similarity. Each entry includes a link that directs the engineer to the information available on the corresponding project. Since the development is currently ongoing, these details may change.

3.4.2.5 Experiments

None.

3.4.2.6 Results

Initial results available.

3.4.2.7 Integration into the DAI-DSS architecture

This service is part of the AI catalogue.

3.4.3 Method: Mathematical Optimisation/Heuristic for Worker Assigment

3.4.3.1 Motivation and Reference to FAIRWork Use Case

In this service, suitable workers are assigned to the tasks. Part of the service creation is the creation of detailed technical and strategy implementation, considering the framework conditions required for the problem, e.g., in terms of human resources, profiles of the workers, the tasks to be performed in the scenario etc. In most cases, different approaches with very different abstractions are possible. However, the choice of the model for optimal resource allocation by considering the time constraints, constraints on the worker profiles (e.g., stress levels, etc.), and available time for making the decisions have a relevant impact on the conceptual solvability of the final resource optimization problem, so that the flexible model building is a key contribution in service implementation.

3.4.3.2 Innovation beyond the State-of-the-art

Heuristics make use of previous experience and intuition to solve a problem. A heuristic algorithm is designed to solve a problem in a shorter time than exact methods, by using different techniques ranging from simple greedy rules to complex structures, which could be dependent on the problem characteristics; however, it does not guarantee to find the optimal solution. Heuristics have been used in the operational research area extensively with respect to the applications.

3.4.3.3 Description of Functionality

The sole and only function of the workers assignment to the production line. Allocation Service is to allocate one worker resource to one task. The worker resources allocation need depends on the use case and the resource allocation strategy. It will perform a resource allocation for a specific set of use case group and not for a generic

resource allocation problem. However, the algorithms can be used a suitable configuration for the doing the proper resource allocation. In the future, the service retrieves data from the Knowledge Base. Furthermore, it will use the service configuration data.

3.4.3.4 Interfaces

The interface for this service will allow engineers to define several types of optimisation parameters.

3.4.3.5 Experiments

None.

3.4.3.6 Results

Initial results are available.

3.4.3.7 Integration into the DAI-DSS Architecture

This service is part of the AI catalogue.

3.5 Methods and Services for AI-Enriched Decision Support Systems

3.5.1 Overview

The focus of AI-enrichment studies is on developing guidelines for developers who specialise in creating recommendation tools. Data-driven modelling is also explored, which includes the application of machine learning or deep learning methods to identify patterns in data. The prototypes developed during these studies can be further scaled up and tested in experiment laboratories available in the FAIRWork project, such as OMiLAB – Open Models Initiative Laboratory, Human Factors Lab (HFL) at Joanneum Research, Robotic Lab of Flex and CRF Lab.

3.5.2 AI Catalogue – A Systematic Literature Review

3.5.2.1 Motivation and Reference to FAIRWork Use Case

The results of our investigation on the first research question (see Section 2.5) serve as a catalogue of various ML, DL, and RL techniques used for production and resource planning. By analysing case studies identified through systematic literature review across the manufacturing sector, we aim to uncover the true potential of these innovative methodologies and identify new research gaps that can be explored in future studies.

3.5.2.2 Innovation beyond the State-of-the-art

A range of studies has explored the application of AI in production and resource planning. Where Zhou et al. (2021)[94] and Nti et al. (2021)[95] focus on the role of AI in production and operations management, with Zhou et al. (2021) emphasising the need for integration, flexibility, and autonomous decision-making. On the other hand, through a broad systematic literature review, Nti et al. (2021) identify opportunities for future research in AI applications in engineering and manufacturing. Nehzati and Ismail (2011)[96] and Raihan (2023)[97] evaluate AI approaches in production scheduling, energy consumption and production, respectively. These studies collectively

[94] Zhou, L., Jiang, Z., Geng, N., Niu, Y., Cui, F., Liu, K., & Qi, N. (2021). Production and operations management for intelligent manufacturing: a systematic literature review. International Journal of Production Research, 60, 808 - 846.
[95] Nti, I.K., Adekoya, A.F., Weyori, B.A., & Nyarko-Boateng, O. (2021). Applications of artificial intelligence in engineering and manufacturing: a systematic review. Journal of Intelligent Manufacturing, 33, 1581 - 1601.
[96] Nehzati, T., & Ismail, N. (2011). Application of Artificial Intelligent in Production Scheduling: a critical evaluation and comparison of key approaches.
[97] Raihan, A. (2023). A comprehensive review of artificial intelligence and machine learning applications in energy consumption and production. Journal of Technology Innovations and Energy.

show AI techniques' potential to enhance manufacturing efficiency. However, none focuses strictly on ML techniques applied for production and resource allocation for decision-making in manufacturing.

3.5.2.3 Description of Functionality

The outcome of this study serves as a catalogue of validated methods that is further used as baseline for the development of AI services. Additionally, it fosters the creation of open-access repositories with demonstrators housing codes and pre-trained models, thereby facilitating knowledge-sharing and collaboration within the AI community.

3.5.3 Guidelines and Recommendations for AI Developers

3.5.3.1 Motivation and Reference to FAIRWork Use case

While AI methodologies offer powerful solutions, they often operate as "black box" systems, where the underlying mechanisms generating outputs remain opaque and difficult to elucidate. Consequently, developers encounter challenges in effectively communicating development requirements and processes to end users, delaying the seamless integration of advanced methodologies into real-world applications. Therefore, clear guidelines involving developers and end users in the DSS development process are imperative to build trust and foster the implementation of new solutions.

This section addresses the second research question (see Section 2.5). It presents guidelines for selecting appropriate methodologies for DSS tailored to the specific requirements and constraints of the given use case and end user.

3.5.3.2 Innovation beyond the State-of-the-art

In our exploration, we have extensively reviewed the existing literature on Decision Support System (DSS) classifications (Musbah, Omar and Ayodeji, 2019[98]; Alter, 1980[99]; Holsapple and Whinston, 1996[100]; Power, 2004[101]). Despite identifying various classifications, none provide comprehensive guidelines for developers to navigate the selection of appropriate methodologies for generating recommendations.

Moving forward, we aim to bridge this gap by proposing a novel DSS classification that supports developers with the necessary tools to align AI methodologies with end users' specific needs, thereby fostering trust and facilitating the integration of advanced technologies into practical applications.

3.5.3.3 Description of Functionality

The proposed DSS classification addresses the gap in DSS classification by providing a decision matrix that brings clarity and structure while facilitating understanding and comparing methods. It is a valuable tool for developers and end-users, enabling them to make an informed selection and implementation of methodologies for generating recommendations.

3.5.3.4 Results

This study addresses crucial research questions by providing a structured categorisation of DSSs into four distinct classes: rule-based, optimisation-based, simulation-based, and learning-based (see Table 3). It identifies essential

[98] Musbah, J.A., Omar, A.N. and Ayodeji, T.A. (2019) 'Decision support systems classification in industry', Periodicals of Engineering and Natural Sciences (PEN). Available at: https://api.semanticscholar.org/CorpusID:201143299.
[99] Alter, S. (1980) Decision support systems: Current practice and continuing challenges. (Addison-Wesley series on decision support). Reading Mass.: Addison-Wesley Pub.
[100] Holsapple, C.W. and Whinston, AB (1996) Decision support systems: A knowledge-based approach. Minneapolis/St. Paul: West Pub. Co.
[101] Power, DJ (2004) 'Specifying an Expanded Framework for Classifying and Describing Decision Support Systems', Communications of the Association for Information Systems, 13 (12pp). doi: 10.17705/1CAIS.01313

criteria for selecting methodological approaches within DSSs, addressing challenges in explaining methodologies to end-users. Furthermore, the proposed classification system seamlessly aligns with user needs and the evolving manufacturing landscape, ensuring the adaptability and effectiveness of the proposed classification.

Table 3: Decision matrix for DSS selection (Olbrych et al., 2024).

		problem type			
		structured procedures	optimisation problems	simulation problems	pattern recognition
interpretability	high	rule-based DSS	optimisation-based DSS		
	medium		optimisation-based DSS	simulation-based DSS	learning-based DSS
	low				learning-based DSS

3.5.4 Industrial Scheduling Optimisation

3.5.4.1 Motivation and Reference to FAIRWork Use Case

This study addresses the third research question (see Section 2.5). We aim to explore how AI, particularly RL, can optimise scheduling processes, focusing on the 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.

3.5.4.2 Innovation beyond the State-of-the-art

The literature proposes various sparse and dense reward functions for the JSP. Each approach customizes its reward function within a distinct context. The observation space, action space, and modelling approach exhibit significant variations across the methods.

Samsonov et al. (2021) [102] created a discrete time simulation for the JSP, where tasks are processed on simulated machines. The agent decides which tasks go to which machines, using a set of fixed processing times for available jobs. The action space remains consistent regardless of problem size. Their reward system is sparse, giving rewards only at the end of each episode, with zero rewards for intermediate steps. Moreover, rewards are higher for improvements closer to the optimal solution, based on the make-span.

Tassel et al. (2021) [103], like Samsonov et al. (2021), present a discrete-time approach for the JSP, but with different state, action, and reward setups. Each step involves assigning jobs to machines, with rewards based on machine usage and scheduling efficiency. They aim to maximize scheduled area rather than directly targeting make-span, as it heavily depends on task waiting times. Their reward system allows for dense feedback, calculated at each step.

[102] Samsonov, V., et al.: Manufacturing control in job shop environments with reinforcement learning. In: ICAART (2), pp. 589–597 (2021)
[103] Tassel, P.P.A., Gebser, M., Schekotihin, K.: A reinforcement learning environment for job-shop scheduling. In: 2021 PRL Workshop-Bridging the Gap Between AI Planning and Reinforcement Learning (2021)

Zhang et al. (2019) [104] approach JSP using disjunctive graphs and a graph neural network (GNN) for state transformation. Their policy integrates GNN-transformed node information and the full graph at each scheduling step. The reward function focuses on critical path growth during scheduling. Through GNN, the agent handles diverse JSP instance sizes. Zhang et al. trained the agent with various JSP instances to assess its ability to learn a generalized solution procedure.

The study identifies a significant research gap concerning the effectiveness of diverse reward functions within RL methodologies applied to the JSP, particularly in minimising make-span. Despite multiple reward functions (Tassel et al., 2021; Samsonov et al., 2021; Zhang et al., 2020), their relative efficacy remains ambiguous. This underscores the necessity for comparative investigations to discern the most suitable reward-shaping strategies for enhancing scheduling optimisations within manufacturing environments.

3.5.4.3 Description of functionality

By systematically analysing the performance of these functions, particularly in terms of make-span minimisation, the study aims to determine which reward-shaping strategies are most effective for optimising scheduling solutions in manufacturing contexts. This functionality serves to enhance understanding and inform decision-making in the development of RL-based scheduling systems.

A disjunctive graph (see Figure 17) is used in this study. In the JSP, each job consists of multiple tasks that must be completed on specific machines in a particular order, leading to intricate task dependencies and machine constraints. A disjunctive graph provides a *clear* representation of these dependencies, where nodes represent tasks and edges represent the possible orderings between tasks on different machines. This graphical representation facilitates the modelling and optimisation of scheduling decisions, allowing for the efficient exploration of various scheduling strategies. Additionally, the chosen modelling approach ensures the generation of feasible and optimised schedules for the job shop problem.

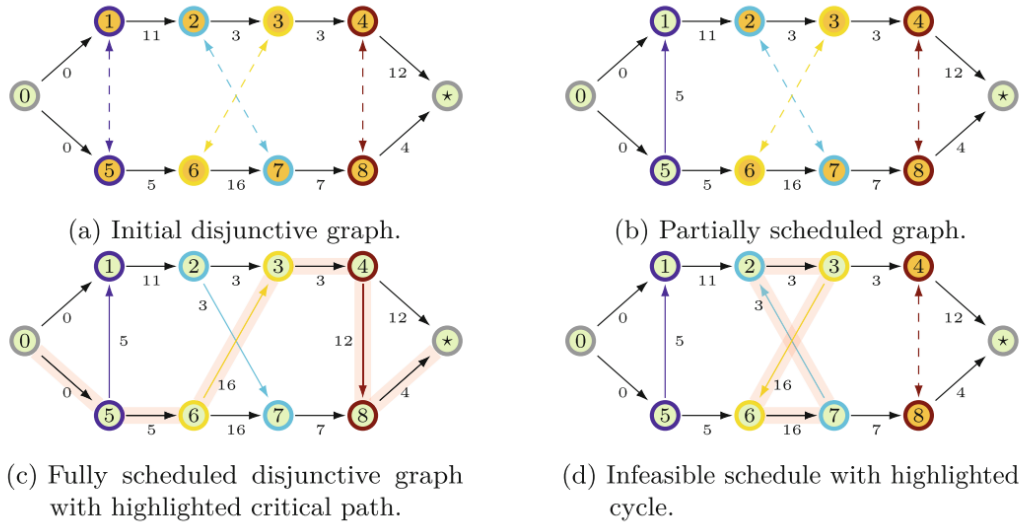


Figure 17: Disjunctive graph scheduling (Nasuta et al., 2024).

3.5.4.4 Experiments

Within this study, we compare reward functions using instances of different sizes from the literature. To learn policies, the study employs the Proximal Policy Optimization (PPO) algorithm with action masking, while Stable

[104] Zhang, C., Song, W., Cao, Z., Zhang, J., Tan, P.S., Chi, X.: Learning to dispatch for job shop scheduling via deep reinforcement learning. *Adv. Neural. Inf. Process. Syst.* 33, 1621–1632 (2020)

Baselines, and Weights and Biases[105] are used for experimentation and tracking, respectively. The focus is on individual instances because multiple runs are required per construction when comparing reward functions. The study runs each JSP instance 100 times with random hyper-parameters while tracking the optimality gap left shift percentage, and rewards. Finally, the study uses the PPO parameterisations of runs with the smallest optimality gap for long-term behaviour investigation and applies the tuned PPO parameterisations to other instances.

3.5.4.5 Results

Using the disjunctive graph approach, we compared different reward functions within a novel RL environment. We found that the choice of reward function, especially in the context of PPO parameterisation, significantly influences performance. Different reward functions perform differently across instances without a clear overall best performer. The reward functions based on machine utilisation show promising learning curves and near-optimal solutions. (Nasuta et al., 2024)

3.6 Methods and Services for Decision-Making Using Multi Agent Systems

3.6.1 Overview

The service developed has the objective of performing resource allocation for industrial environments through Multi-Agent Systems. The goal is to provide support to decision-making in decentralised way while ensuring a human-centric approach using human relevant data in use case scenarios. This service accesses the Knowledge Base for a retrieving the relevant data and it is connected to the Orchestrator to streamline workflow and necessary data exchange. It fits in the optimization field while also considering relevant data regarding human aspects in the decision process.

3.6.2 Multi-Agent Resource Allocation Service

3.6.2.1 Motivation and Reference to FAIRWork Use Case

Initially applied to the CRF Workload Balance Use Case Scenario, this service can be expanded to resource allocation scenarios where complexity and scalability are present as a challenge. This solution can support decision-making in situations where a manager needs to decide where to assign a worker considering many variables. These variables encompass production related data and worker's relevant conditions to the given production processes. The service provides a viable path to consider human related aspects into the decision-making process. The Multi-Agent Systems stands as an interesting approach to systems containing a great variety of actors that interact in many forms. Such actors are modelled in a way that the relevant data and the stakeholders interaction structure of the business process can be considered and quantified in a dynamic context. Many conflicts can arise from such interactions. MAS is a valuable technology for offering a means for conflict resolution.

3.6.2.2 Innovation beyond the state-of-the-art

The Multi-Agent Systems elevate Industry 5.0's decision-making by introducing autonomous, nuanced processes that integrate diverse data, reflecting a human-centric approach to problem-solving. These systems go beyond mere communication tools, acting as collaborative entities that enhance human network intelligence. By simulating complex social interactions, MAS offer insights into decision fairness and adapt over time, learning from outcomes to refine processes, aligning with Industry 5.0's state of technology as a human collaborator. They provide a decentralized decision-making model, reducing biases and incorporating a broad spectrum of perspectives,

[105] Biewald, L. (2020). Experiment Tracking with Weights and Biases. [Software]. Retrieved from https://www.wandb.com/

embodying Industry 5.0's vision of technology that supports adaptive, transparent, and fair human governance. It's a technology that fosters democratic decision-making through enhanced worker participation in the decision-making processes. The interaction mechanisms in MAS aligned to human data in a human-centric perspective, i.e. in a perspective that supports positive values for the humans, e.g. worker wellbeing in industry, can be designed[106] to promote fairer decisions based not only in production goals, but also in human condition, state and desire, going beyond of the current state-of-the-art.

Resource allocation in a workload balance scenario using Multi-Agent Systems is a breakthrough that requires more study. The agentification of stakeholders in a decision-making process allows for the consideration of a diversified range of inputs coming from multiple actors in a decision process, therefore promoting decision aware of human needs resulting in increased satisfaction. Further analysis of mechanisms of agent interaction and role[107] for fairer allocations is a requirement for realistic balance of workload in a production-oriented environment considering human and production parameters related to processes goals. MAS aligned to human factors enable a human-centred approach in decision-making for developing an increased level in worker satisfaction while balancing job demand with worker well-being[108] in complex environments with multiple actors and configurable goals. Some studies have analysed democracy aspects in open agent societies[109]. However, such an approach in closed societies with a high level in constraints as in industry settings has not yet been profoundly studied.

The exploration of MAS in smart factory environments, particularly in the context of Industry 4.0[110], marks a significant stride in optimizing production processes through advanced simulations and data-driven decision-making. However, the leap to Industry 5.0 with a human-centred perspective pushes the envelope further, integrating fairness into the decision-making processes. In Industry 5.0, MAS doesn't just streamline operations or enhance efficiency; it actively ensures that decisions reflect human participation, prioritize worker well-being, and foster inclusivity. This advanced application of MAS transcends mere operational optimization, embodying a paradigm where technology and human values converge to create socially responsible production environments. In this evolved context, MAS balance productivity with ergonomic considerations, adapt processes to accommodate diverse workforce needs, promoting equitable distribution of benefits and opportunities, illustrating a profound shift from technology-centric to human-centric industrial advancement.

3.6.2.3 Description of functionality

The service access data in the Knowledge Base periodically to offer an updated assessment on the resource allocation recommendation. It runs the algorithm in the Multi-Agent System to provide a possible solution to the given allocation challenge. It provides a recommendation before each shift in worker allocations using relevant data from the available workers related to the task required to perform and data related to the product and lines available to produce them. These data are process in the algorithm to suggest a feasible allocation configuration given the multiple possible configurations.

3.6.2.4 Interfaces

Interface with Knowledge Base and Orchestrator.

---

[106] Helbing, D., Mahajan, S., Fricker, R. H., Musso, A., Hausladen, C. I., Carissimo, C., … Pournaras, E. (2023). Democracy by Design: Perspectives for Digitally Assisted, Participatory Upgrades of Society. *Journal of Computational Science*, *71*, 102061. doi:10.1016/j.jocs.2023.102061

[107] van de Merwe, K., Mallam, S., & Nazir, S. (2024). Agent Transparency, Situation Awareness, Mental Workload, and Operator Performance: A Systematic Literature Review. *Human Factors*, *66*(1), 180-208. https://doi.org/10.1177/00187208221077804

[108] Isham, A., Mair, S., & Jackson, T. (2021). Worker wellbeing and productivity in advanced economies: Re-examining the link. *Ecological Economics*, *184*, 106989. doi:10.1016/j.ecolecon.2021.106989

[109] J. Pitt and J. Ober, "Democracy by Design: Basic Democracy and the Self-Organisation of Collective Governance," *2018 IEEE 12th International Conference on Self-Adaptive and Self-Organizing Systems (SASO)*, Trento, Italy, 2018, pp. 20-29, doi: 10.1109/SASO.2018.00013.

[110] Dornhöfer, M.; Sack, S.; Zenkert, J.; Fathi, M. Simulation of Smart Factory Processes Applying Multi-Agent-Systems—A Knowledge Management Perspective. *J. Manuf. Mater. Process.* 2020, *4*, 89. https://doi.org/10.3390/jmmp4030089

3.6.2.5 Experiments

The Multi-Agent System designed for workload balancing represents stakeholders as autonomous agents in a negotiation setting to allocate suitable workers to production lines. This system evaluates worker parameters indicating a worker's capability to perform tasks under specific conditions. Aspects such as previous experience, resilience, preference are considered for making a decision. The system calculates a score for each worker based on their resilience and preferences for working on particular production lines. Workers rate their line preferences, which, combined with their resilience, determine their allocation score, with predefined weights assigned to these factors. These weights can be adjusted to prioritize resilience or worker preference, reflecting the industry's needs or focus on worker well-being. This approach enables a dynamic and responsive worker allocation process, catering to both operational efficiency and worker satisfaction leveraging human participation in the decision-making process and bringing possibilities for a fairer decision when considering human aspects often neglected. In conflict situations, MAS enables conflict resolution in this experiment through weight balancing based on agent previous allocations aiming at a fairer outcome.

3.7 Model-based Knowledge Engineering for Decision Support

3.7.1 Overview

In this section we will introduce methods and services that were developed to fit in the research track introduced in Section 2.7. Therefore, we first will introduce a three-layered approach, supporting FAIRWork in identify and specify knowledge about the decision scenarios and prepare it so that it can be used for the DAI-DSS. Further, experiments and prototypes are discussed for configuring decision services through conceptual modelling and supporting the knowledge exchange during the method will be introduced.

The methodology is based on models with the assumption that modelling create transparency, understandability and hence contribute to trustworthiness. The idea is to use models for describing the problem settings as well as the abstract logic of applicable AI services for these. For this purpose, different types of models can be combined. The layers described through models in combination with verification by domain-experts or technicians aim to leverage trustworthiness of AI. Based on the previously presented method a model-based configuration of a rule-based decision service and a fuzzy logic approach is applied. Additionally, we discuss how the information exchange between the layers can be supported in a bidirectional way, allowing to reuse information form the high-level layers in lower-level layer but also enables to show information from the applied decision support to visualize it in the models to support explainability.

For the below described methodology and services a bottom-up approach was used. Therefore, the “Workload Balance” decision scenario, as defined in Deliverables D2.1 (Zeiner, 2023) and D5.1 (Chevuri, 2023[111]), was taken as a starting point, providing first insights.

3.7.1 Method: Conceptual Modelling for Knowledge Engineering – a three-layered approach

3.7.1.1 Motivation and Reference to FAIRWork Use Case

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. The proposed method applies three layers (Identification, Specification and Configuration), where each layer contributes to

[111] Chevuri, R. (2023). DAI-DSS FAIRWork Knowledge Base at Use Case Site. https://fairwork-project.eu/deliverables/d5-1/D5.1_DAI-DSS%20FAIRWork%20knowledge%20base_v1.0-preliminary.pdf (accessed: 09-04-2024)

bringing human knowledge into an IT-based system using conceptual models. In the FAIRWork project domain experts coming from the use case side have been consulted about their daily decisions making strategies and the resulting decision logic was represented in form of conceptual models (e.g. BPMN). On the technical side, an AI catalog containing algorithms and services is being developed as part of the project. These AI solutions need to be configured to meet the specific use case needs. This present research explores how models might serve as intermediary layers that connect requirements and capabilities, fostering mutual understanding and increased transparency. This proposed method is not limited to one specific use case in FAIRWork but investigates the utilization of models to identify requirements coming from different use case scenarios, to specify anticipated cases for an AI solution to address, and to configure AI services.

3.7.1.2 Innovation beyond the State-of-the-art

The combination of Conceptual Modeling (CM) with Artificial Intelligence (AI) approaches defines the CMAI domain, which aims to improve the strengths and address the weaknesses of each separate domain[112]. Currently, Artificial Intelligence (AI) is highly visible and has enormous societal expectations for transformation, instead Conceptual Modelling lacks equivalent attention from the general public. This should not imply that explicitly creating a model by hand that reflects a domain is no longer required or helpful during system development, rather, it would be more beneficial to investigate the relationship between AI and modelling[113]. This research can be positioned into the recent CMAI domain, but emphasizes on the decision making aspect. Bork et al. (2023)[112] indicate that based on their review especially conceptual modelling for AI Ethic 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)[114] highlight the importance of CM and AI to leverage reliability, interpretability and robustness of deep learning approaches by applying hybrid approaches. Mattioli et al. (2022)[115] emphasize hybrid AI-Approaches to achieve trustworthy AI and similarly they are proposing multiple steps for engineering AI for problem settings. Hence, as starting point, the presented methodology focuses on a very specific use case for supporting proceedings in model-based code generation for different AI methods, but also investigate in the next steps the modelling for machine or deep learning, NLP or generative AI. Additionally, concepts for ethical assessments can be applied based on the models.

3.7.1.3 Interfaces

For analysing the rule -based service and the fuzzy logic the hybrid modelling tool Bee-Up[116] is used. Bee-Up is based on ADOXX[117], which is an open-source metamodeling platform allowing to develop various modelling tools. The Bee-Up tool combines multiple modelling languages into a single prototypical implementation tool and it can be downloaded and used for free. It allows to create models in commonly used modelling languages such as BPMN, DMN and UML. 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. Additionally, ADOxx modelling environment is used for modelling and their prototypical certification with ADOxx-based services.

3.7.1.4 Experiments

The current research effort uses a bottom-up approach to develop experimental prototypes, that are then further investigated from the three-layer perspective proposed by the method under study. The decision-making

[112] Bork, D., Ali, S. J., & Roelens, B. (2023). Conceptual Modeling and Artificial Intelligence: A Systematic Mapping Study.
[113] Fettke, P. (2020). Conceptual Modelling and Artificial Intelligence: Overview and research challenges from the perspective of predictive business process management. In Companion Proceedings of Modellierung 2020 Short, Workshop and Tools & Demo Papers co-located with Modellierung 2020 (Vol. 2542, pp. 157–164). CEUR-WS.org. https://ceur-ws.org/Vol-2542/MOD-KI4.pdf
[114] Shlezinger, N., Whang, J., Eldar, Y., & Dimakis, A. (2020). Model-Based Deep Learning.
[115] 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. In Workshop on Artificial Intelligence Safety (SafeAI).
[116] https://austria.omilab.org/psm/content/bee-up/info
[117] https://www.adoxx.org/ (visited: 24.04.2024)

processes, which are inspected and represented in the Identification layer are based on use case scenarios known from the FAIRWork project (e.g. worker allocation to different production lines). Additional to the scenario several AI approaches (e.g. rules, fuzzy rules, artificial neural networks, etc.) need to be selected. We utilize existing AI services created during the project whenever feasible. We anticipate that the AI method largely impacts the Specification and Configuration layer. An investigation will be conducted on how the nature of the AI approach (knowledge-driven, data-driven, distributed) is influencing the selection of appropriate modelling languages and configuration possibilities. Analysing the experimental prototypes aims to determine the fundamental rules, conditions, and requirements for the three layers. The figure below displays the selected settings for the experimental prototypes.

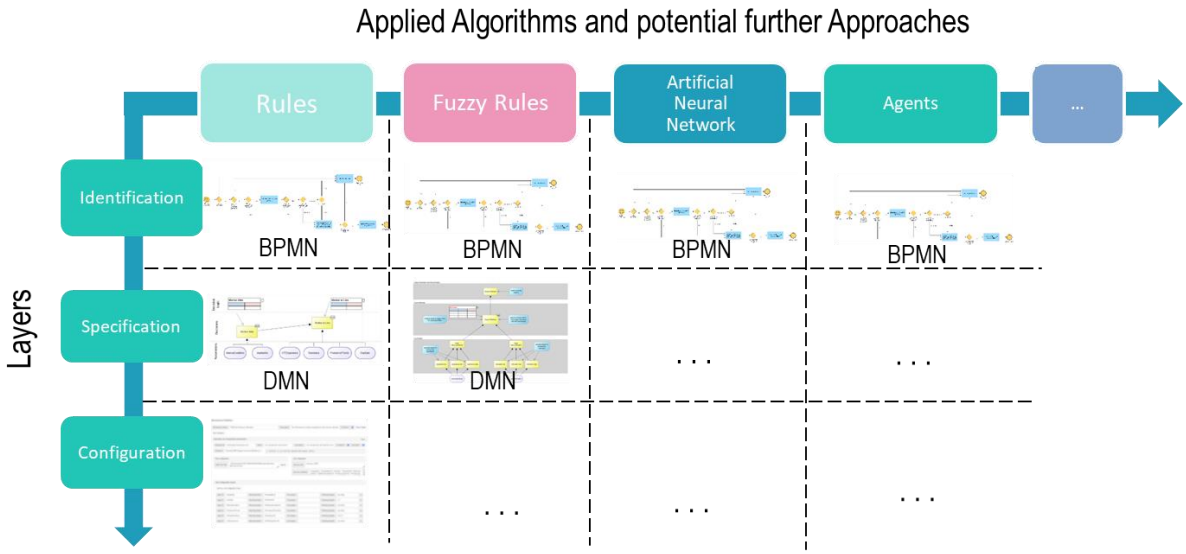


Figure 18: Current approaches and outlook.

3.7.1.5 Results - Fuzzy logic, Rules

The research is ongoing, but the first findings and observations can already be reported. For the **first AI approach**, the method is applied to a rule-based service configurable using DMN models. The service is presented in more detail in Section 3.3.3, but the following figure shows how the different parts of the service were assigned to the different layers.

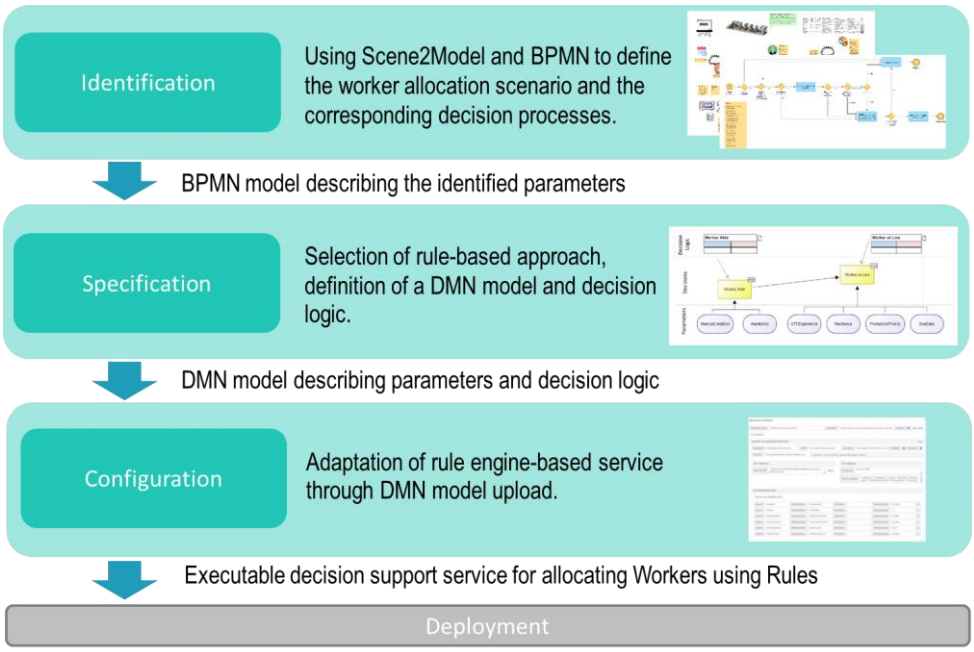


Figure 19: Rule-Based Allocation Service assigned to different layers.

For the **second AI approach**, the method is applied to fuzzy logic. The current application of the defined methodology including the three layers for the fuzzy logic approach is illustrated in Figure 20 below. The first and second layer were specified and modelled.

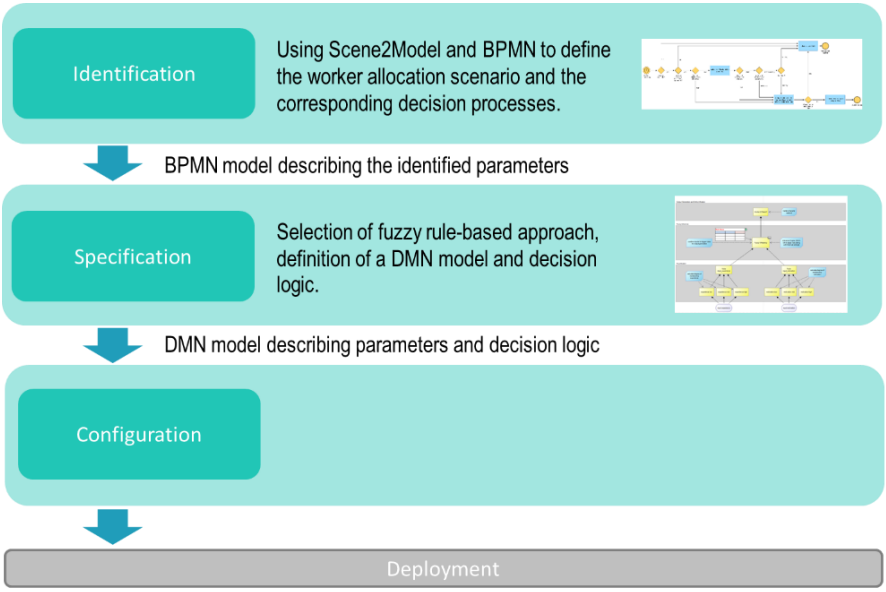


Figure 20: Fuzzy Logic approach assigned to different layers.

The ”Identification” layer consist of the concrete problem setting and in the Workload Balance use case covers the detailed decision making process and all gateways.

The “Specification” layer covers two steps. First, the decision model is analysed for suitable tasks or decision gateways where the human-like reasoning approach can be applied. Especially, gateways describing vague, uncertain decisions, or where no strict yes or no decision paths can be applied, are suitable application areas for

fuzzy reasoning. In the “Workload Balance” scenario the utilization of fuzzy logic fits to the efficiency of workers estimation or for defining physical resilience.

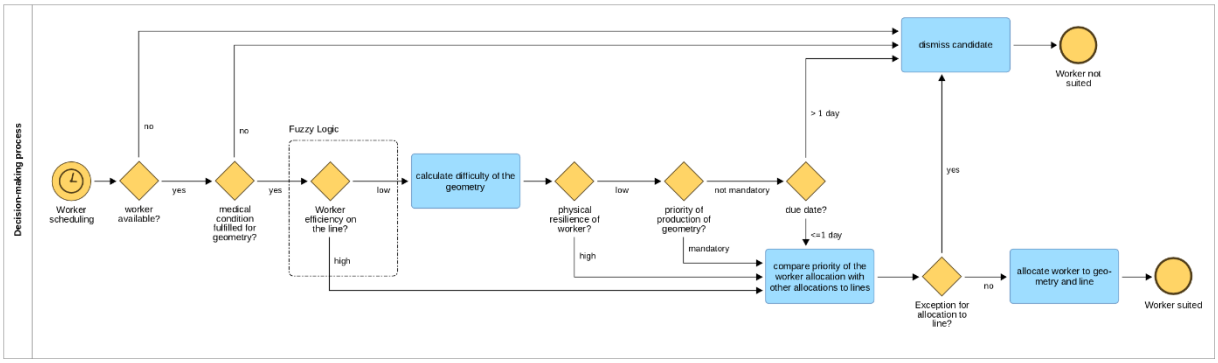


Figure 21: Identification for Fuzzy Logic application.

For the first parameter “efficiency”, a specification of the abstract logic for fuzzy reasoning was modelled with DMN by using input parameters “experience of the worker” and “motivation to work”. It details the logical connection of fuzzifying the concrete inputs for experience and motivation while determining each corresponding rule strength. Rule strength describes a value between 0 and 1 and describes the degree to which a particular rule is applicable or influential in making decisions within a fuzzy logic system. The relevant fuzzy input parameters “experience and motivation” where the rule strength is not zero along with their labels “high, mid or low” membership are then combined to determine the fuzzy output “efficiency”. A combination of all possible input pairs through “AND” or minimum operators determines the rules that are activated and their rule strength. This means if experience low (degree of membership of 0.2) and experience mid (degree of membership 0.7) and motivation low (degree of membership 0.5) are triggered all combinations of rules for defining the output parameter are relevant. In this example two rules are activated: a) IF “experience low (0.2)” AND “motivation low (0.5)”, THEN the “efficiency low” with the lower rule strength 0.2 is applied and b) IF “experience mid (0.7)” AND “motivation low (0.5)” THEN “efficiency mid” with the lower rule strength 0.5 must be considered. These IF-THEN statements are defined based on experts' knowledge and cover the connection of inputs with the corresponding output membership functions. The relevant output labels “efficiency low” and “efficiency mid” are defined through the membership functions for the output and the “height” of the output fuzzy areas are determined through the rule strength. These areas are overlapping and combined through fuzzy inference engines with an “OR” operator. This means that the overall area that is either part of efficiency low or efficiency mid membership functions are relevant as output. To get a crisp value from this area in the De-fuzzification step the centre of gravity calculation is applied resulting in an exact value for efficiency.

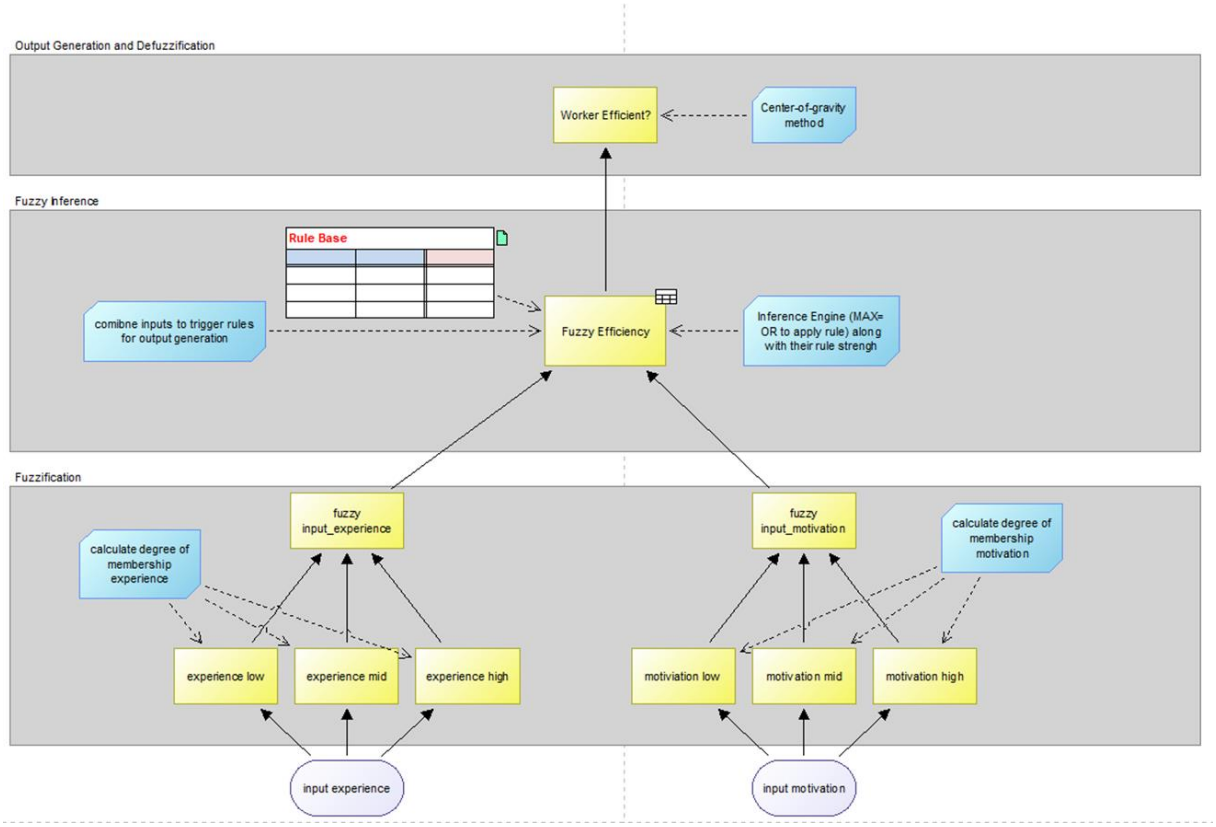


Figure 22: Fuzzy Logic abstract logic.

This Logic is described in the DMN model graphically for the Specification layer and could be applied to any fuzzy logic case where two inputs with membership low, mid, high for one output with a membership of "low, mid or high" are considered. The "configuration" layer is not yet defined.

The created models in the different layers can raise transparency and explainability and by performing certification and verification activities on them, this approach contributes to the Ethical Watchdog concept. As each of the layers of the methodology require different levels of trust, the idea is to support the three layers with certification and verification activities based on models. In order to enable trustworthiness, criteria such as correctness, completeness, robustness etc. must be fulfilled and approved by the appropriate experts. This certification process is elaborated in more detail in the Ethical Watchdog Section 3.9

3.7.1.6 Integration into the DAI-DSS Architecture

The results of the present research contribute to the DAI-DSS configurator component present in FAIRWork's architecture. The configurator consists of two subcomponents, the Configuration Framework, and the Configuration Integration Framework. The former enables the creation of decision models using several modelling environments. The resulting model data serves then as a basis for the second part of the component, which enables the generation of configuration files for the available services. The proposed method can be seen as a guide to realize a suitable service selection and configuration for the decision support system. Depending on the chosen AI service, it may be possible that not all layers are applicable. But in general, we expect to find insights that contribute to the Configuration Framework by investigating the first two layers (Identification, Specification), while the third layer (Configuration) highlights aspects of bringing the two architectural subcomponents together.

3.7.2 Service: Model-based configuration of Rule-Based Decision Services

3.7.2.1 Motivation and Reference to FAIRWork Use Case

This experiment was created with the CRF Workload allocation use case in mind, where the rules should support the decision, if a worker is allowed at a specific production line. But the experiment itself was focused on providing a way to easily adapt the used rule to different cases. Therefore, the overall prototype could also be applied to other cases, where the defined input parameters, containing strings, numbers or bool values should be mapped to an output value. The information to do so must be provided by an expert, who know how to decide it. Their knowledge is encoded into the rules.

3.7.2.2 Innovation beyond the State-of-the-art

Organizations and their information system are under pressure to quickly adapt to changing requirements. Low-code platforms, improving the productivity by offering tools to lower the complexity of creating information system are one way to tackle such dynamic adaptations (Bock & Frank, 2021)[118]. Low-code platforms cannot only be implemented as all containing applications, but following a microservice architecture, where the different capabilities which should be available are implemented as individual services, which can be configured to can be deployed and integrated to fulfil certain tasks (Falcioni & Woitsch, 2021)[119]. Here conceptual modelling can be used to support the configuration in way comprehensible even for domain experts not familiar with software engineering.

Conceptual modelling itself is not only used for documenting and designing system, but to use models as the basis for processing, increasing the model value (Bork et. al., 2018)[120]. Model value in this context mean that the added value for the users of the models and the corresponding modelling tools should be maximized, by processing the models. In this context of this service this means that the modelled knowledge is used to configure the decision services, by minimizing the needed manual tasks of the users.

In addition, the interest in connecting the fields of AI and conceptual modelling is rising, with a focus set of improving modelling methods through AI algorithms and a left potential for supporting AI through conceptual modelling (Bork et. al., 2023)[121].

This experiment service was used to analyse how conceptual modelling can be used as input to configure knowledge-based decision service and how the configuration of such decision service can be supported to minimize the users' manual tasks.

3.7.2.3 Description of Functionality

The experiment consists of two components: A modelling tool created for the configuration and a microservice implementing the decision service. Here the focus is set on the modelling and how the modelling tool can be used to ease the configuration of the existing service.

The modelling environment offers a graphical interface to create models and define the decision logic. In addition, it offers functionality helping with creating the decision knowledge, but the most important functionality is that the modelled knowledge can be exported to a microservice environment, which uses the information to instantiate a decision service, with a callable endpoint. The export functionality can be used to create a file for manually configuring the decision service. Or the information about the deployed microservice environment can be added to

[118] Bock, A. C., & Frank, U. (2021). Low-code platform. Business & Information Systems Engineering, 63, 733–740.
[119] Falcioni, D., & Woitsch, R. (2021). OLIVE, a Model-Aware Microservice Framework. In E. Serral, J. Stirna, J. Ralyté, & J. Grabis (Eds.), The Practice of Enterprise Modeling (pp. 90–99). Springer International Publishing.
[120] 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/
[121] Bork, D., Ali, S. J., & Roelens, B. (2023). Conceptual Modeling and Artificial Intelligence: A Systematic Mapping Study.

the model and then the knowledge can be directly deployed to the Olive (Falcioni & Woitsch, 2021)[119] microservice controller. Independent if the automated or manual approach is used, afterwards a microservice with a REST endpoint is available, which can be called from other systems, like the DAI-DSS orchestrator, to make a decision.

The rules for the different decisions are defined in decision tables, where each line contains one rule. The rules are saved in CSV format and written with a locally installed application, like Microsoft Excel or LibreOffice calc. The modelling tool offers functionality to create, open and import the tables into the modelling tool for further usage.

The decision service itself offers an endpoint, where a value is assigned to the provided parameters. Multiple sets of parameters can be provided at once and the service will return a value for each of this set, based on the configured rules. The rules can contain sub-decisions, which are evaluated and then are used as input for the further decisions. The result of the main decision is then returned. The input data for the decision must be provided by the calling system, as the service self does not gather data on its own.

Based on the model-based configuration different decision services with their own parameters and rulesets can be defined. The configuration is one important functionality of the service, as eases the reuse of the decision service. To achieve this the configuration functionality and managing the services lifecycle must be available. For this functionality we used the Olive microservice framework. Within this framework we implemented the configurable decision service. For evaluating the defined rules, we used the *Camunda DMN Engine* for our first prototype.

3.7.2.4 Interfaces

For creating the models, a modelling tool with a *Graphical User Interface (GUI)* is provided, allowing to create the diagrammatic models and trigger the available functionality.

For the decision service REST interfaces are used, where data is provided and consumed in JSON format. One interface consumes a JSON containing keys for the configured parameters and the corresponding values and then return answer in JSON format.

For the configuration out of another application the Olive framework offers endpoints to upload the file, containing the decision knowledge in XML format. And two interfaces one for creating a new microservice and one for updating an existing one. Here the configuration parameters defined for the decision service, like the ID of the uploaded file, the mapping of the parameters to the decision variables from the model and the ID for the decision, for which the answer should be returned, must be provided. The same information must be provided to updating an existing service.

For manually configuring a decision service, the Olive controller also offers a graphical web-based interface. Here the same information as above mentioned must be provided.

3.7.2.5 Experiments

For this service a prototype was created, which is also described in our deliverable 4.2 (Vieira, 2023)[122]. The prototype was created in an experiment, where our CRF Worker Allocation use case was taken and a decision service should be configured, deciding if a worker is allowed on a specific production line producing a specific product.

The experiment was used to analyse two parts, which are closely related to each other. One is how a model-based configuration of decision service can be made in a low-code manor and how the layers introduced in section 3.7.1 can be used together to create a usable service. The two parts are related, as the model-based configuration is

[122] Vieira, G. (2023). Initial DAI-DSS Prototype D4.2. https://fairwork-project.eu/deliverables/D4.2_Initial%20DAI-DSS%20Prototype%20v1.0a-preliminary.pdf

part of the layers. The separation here is made as it is once analysed from a technical view and one from a methodological one.

From the technical view we analysed how the needed knowledge can be encoded, regarding the decision and technical information, to allow the deployment of a service, so that the defined case can be used. The goal was to have a testing environment where a decision service is running and can be used by the DAI-DSS orchestrator to trigger the decision support. The decision service was then tested with the input from the before defined use cases, to see if the provided rules return the anticipated results.

For the methodological point of view, the experiment was used as an example to apply the design methodology which was now enhanced with the three-layered approach. The context of this service was used to manually translate the information captured in the models from the identification layer, established through workshops and discussions, into decision models, containing concrete decision knowledge. These models are used in the specification layer and define how and what a deployed decision service should decide. The last step is that the model is transformed in a format, so that the decision service can understand it and offer an endpoint, used in the DAI-DSS.

The experiment was used to identify the semantic gaps between the different layers, and what information is needed to come from an understanding the decision scenario to a usable decision service. Therefore, the CRF worker allocation use case was used to apply each layer and gather insights.

The created experiment is visualised in Figure 23. The identification layer was taken from the identified use cases and then the models in the specification and the functionality to export it to the Olive controller was designed and implemented.

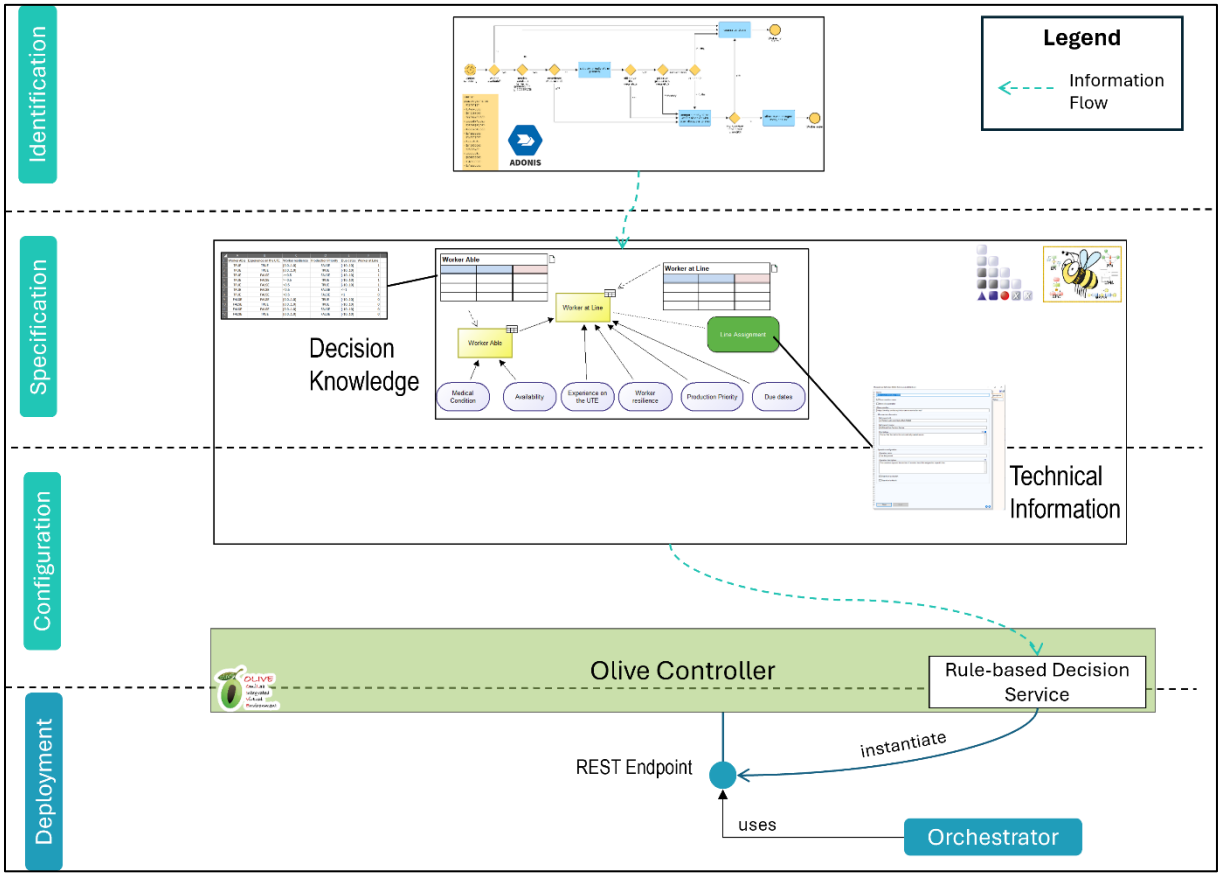


Figure 23: Visualization of the Experiment for the Model-based Designing and Configuring of Decision Services.

3.7.2.6 Results

For this service a first experiment prototype was created, which was integrated with the workflow orchestrator of the DAI-DSS prototype. This together combined with the knowledge base and the user interface provided a one prototype to see how the DAI-DSS can be further designed and implemented.

Information about the project and how it can be used can be found at: https://code.omilab.org/research-projects/fairwork/decision-services/bee-up-dmn-extension

This project contains currently the explanation on how to set-up the tool and service environment and how one can use the export of the decision knowledge and manually uploaded to create a decision service. This is also explained in the Deliverable 4.2 (Vieira, 2023). But the focus for implementing this service in WP4 was the decision service and how it can be integrated with the orchestrator. The focus for WP3 was set to the modelling and configuration.

3.7.2.7 Integration into the DAI-DSS architecture

This experiment is aligned with two components of the DAI-DSS architecture: the Configuration Framework and the AI Enrichment Services. The modelling environment is used for the configuration together with the abstract implementation of the decision service, managed by the Olive controller. The rule-based decision service itself is located in the AI Enrichment services.

The instantiated service can then be used by the orchestrator to integrate it in the overall decision process. The service itself is independent from the knowledge base and the orchestrator must retrieve the needed data from the knowledge base and provide it to the service.

3.7.2.8 Ethical issues

The service does not directly produce ethical issues, but its usage can. Meaning that the decision rules that are defined are based on the expert knowledge and therefore can contain unwanted biases or ethical. Therefore, during the configuration of this decision service, the users must consider such problems, or the DAI-DSS and FAIRWork offer approaches to support the reduction of ethical issues.

3.7.3 Service: Supporting FAIRWork's Design Methodology Through Supported Knowledge Transfer

This service is currently in the development phase of its first prototype. The goal of this service is to support the configuration and design methodology for the DAI-DSS.

3.7.3.1 Motivation within FAIRWork

One goal of the DAI-DSS is to enable flexible adaptation of its decision supports to new or changing decision scenarios in complex production processes. To achieve this a model-based configuration approach which is based on the design methodology (Zeiner, 2023)[31] and the layers introduced in section 3.7.1. This approach uses various conceptual modelling methods on different abstraction levels and levels of formality to support the understanding of the involved stakeholders and ease the configuration of the DAI-DSS.

The design methodology supports the *Identification Layer*, where high-level conceptual models are used with a focus on supporting communication and understanding between humans (Woitsch et. al., 2023)[123]. Here conceptual

[123] Woitsch, R., Muck, C., Utz, W., & Zeiner, H. (2023, September). Towards a democratic AI-based decision support system to improve decision making in complex ecosystems. Joint Proceedings of the BIR 2023 Workshops and Doctoral Consortium Co-Located with 22nd International Conference on Perspectives in Business Informatics Research (BIR 2023). https://ceur-ws.org/Vol-3514/paper94.pdf

modelling approaches used in design thinking workshops, e.g., Scene2Model (Miron et. al., 2019)[124], or *Business Process Model and Notation (BPMN)*[125] models are used to describe the decision processes. In the *Specification Layer* models containing concrete decision knowledge are used to formalize how decisions should be made. Finaly, in the *Configuration Layer* the information gathered in the models of the first two layers is translated into a configuration of the DAI-DSS, which can be used to suggest solution for decision problems.

Exchanging the information between within the design procedure and the between the layers is currently based on human interpretation and manual tasks. In this service, we research how this information exchange can be supported. Here a basic format and way of describing information will be analysed and established, which will utilize semantic rich knowledge representations, allowing to support the information exchange. Here knowledge graphs (Ehrlinger & Wöß, 2016)[126] should be utilised as a way that information can be described semantically rich and making it understandable to machines.

The information exchange should not only be supported in one direction, from the high-level models to the configuration, but bidirectional so that information can be used in the models to adapt them or to visualize concrete information supporting the understanding.

3.7.3.2 Innovation beyond the state-of-the-art

Conceptual modelling is not only used to visualize and document complex systems, but automated processing of models, can increase the model value and therefore further support users (Bork et. al., 2018)[30]. Processing is used to analyse the modelled information to support users in better understanding the system under study (Medvedev et. al., 2021)[127] or to integrate the modelled information into information system to reuse the already described information (Bock & Frank 2021)[118]. Often one modelling language is not enough to cover all needed viewpoints or abstraction levels, leading to an integration of multiple modelling languages (Karagiannis et. al., 2016)[29] (Bock & Frank 2016)[128] and even linking the modelled information to sources external to the current modelling tool (Fill, 2017)[130]. In this way a more holistic view on complex systems can be established and the processing can be improved as more information is available. To integrate models with external knowledge and make modelled information available to other applications, the models itself can be semantically enriched by integrating it with external data sources using RDF (Buchmann & Karagiannis, 2016)[129] (Fill, 2017)[130].

But integrating the models with applications external to the modelling tool is not only done in the direction from models to information systems, but also in the other direction, which is often referred to as *models at runtime* (Szvetits & Zdun, 2016)[131]. Data created and collected in information systems is automatically integrated into the models, which provide context, create visualization, and enable processing to further support the users and the system itself.

[124] 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.
[125] Business Process Model and Notation (BPMN), Version 2.0.2, (2013). http://www.omg.org/spec/BPMN/2.0.2
[126] Ehrlinger, L., & Wöß, W. (2016, September). Towards a Definition of Knowledge Graphs. Joint Proceedings of the Posters and Demos Track of 12th International Conference on Semantic Systems - SEMANTiCS2016 and 1st International Workshop on Semantic Change & Evolving Semantics (SuCCESS16).
[127] Medvedev, D., Shani, U., & Dori, D. (2021). Gaining Insights into Conceptual Models: A Graph-Theoretic Querying Approach. Applied Sciences, 11(2). https://doi.org/10.3390/app11020765
[128] Bock, A., & Frank, U. (2016). Multi-perspective Enterprise Modeling—Conceptual Foundation and Implementation with ADOxx. In D. Karagiannis, H. C. Mayr, & J. Mylopoulos (Eds.), Domain-Specific Conceptual Modeling: Concepts, Methods and Tools (pp. 241–267). Springer International Publishing. https://doi.org/10.1007/978-3-319-39417-6_11
[129] Buchmann, R. A., & Karagiannis, D. (2016). Enriching Linked Data with Semantics from Domain-Specific Diagrammatic Models. Business & Information Systems Engineering, 58(5), 341–353.
[130] Fill, H.-G. (2017). SeMFIS: a flexible engineering platform for semantic annotations of conceptual models. Semantic Web, 8(5), 747–763. https://doi.org/10.3233/SW-160235
[131] 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.

Knowledge graphs, e.g., implemented in RDF, in general can be used to integrate information and make it better understandable to applications (Ehrlinger & Wöß, 2016)[126]. Within knowledge graphs, not only instance data can be used, but also meta information, improving the integration of the data from different sources, like models and applications.

With this service we want to support the DAI-DSS by providing a way to transform the modelled information into RDF-based graph structures, which enables and flexible integration of information from different sources and supports a unified way of processing the information. But the information should not only come from the models, but other sources, like the DAI-DSS knowledge base. Then different services or applications within or adjacent to the DAI-DSS should be able to consume the information. For example, decision services or models to visualize them.

This research goes beyond the state of the art as analysis how information from different sources, with a focus of including conceptual models, can be integrated and then used in the decision making and support the explanation of decision making through visualizing made decisions as conceptual models.

3.7.3.3 Description of functionality

The need for this service was identified during the working on other prototypes for FAIRWork, especially in the context of the model-based configuration (see section 3.7.2). Therefore, no sharable prototype is yet available and this section will contain the description of the planned functionality.

The main functionality will be to export created diagrammatic, conceptual models in a knowledge graph structure, to make it easier accessible to other applications. To encode the knowledge graph, we will use the *Resource Description Framework (RDF)*, where the models and meta information will be transformed into RDF statements, which and then be used in other application to apply reasoning and gather the needed information abased on pattern matching then a pre-defined structure as used in XML or JSON. This export functionality will first be provided as a plug-in to ADOxx[132]-based modelling tools, such as Scene2Model and Bee-Up (Karagiannis et. al.,2016)[133], ADOXX is an open-source metamodeling platform allowing to develop various modelling tools. To achieve this, a way to effectively provide plugins within the ADOxx platform will be analysed and created. By making this available on the platform level, it can easier be reused for concrete modelling tools implemented within, allowing also future modelling methods used in FAIRWork, to use the RDF export.

The plugin will create a trigger in the modelling tools, with which the models can be exported in RDF structure, containing the information form the models, the meta model and additional information to ease the usage with other applications. Therefore, a basic structure will be defined, which will be included in the RDF export.

The common RDF structure will also be used to provide information to be shown in models. For example, that information from other modelling methods or the DAI-DSS can be used as input to be shown in models. In this way the available information can be visualized in other means, supporting a explanation of the decision scenarios within FAIRWork.

3.7.3.4 First Experiment Prototype

For the first experimental prototype we created a transformation from the conceptual model to a correct RDF structure. This is done by using XSLT-based transformation, based on a generic data structure provided by ADOxx.

[132] https://www.adoxx.org/ (visited: 02.04.2024)
[133] Karagiannis, D., Buchmann, R. A., Burzynski, P., Reimer, U., & Walch, M. (2016). Fundamental Conceptual Modeling Languages in OMiLAB. In D. Karagiannis, H. C. Mayr, & J. Mylopoulos (Eds.), Domain-Specific Conceptual Modeling: Concepts, Methods and Tools (pp. 3–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-39417-6_1

The transformation is defined as a set of rules, which can be interpreted by the system and then applied to the model information.

The resulting structure was imported into the graph database GraphDB[134], to evaluate if the result is in the right format and the information can be queried using SPARQL, as a standard query language for RDF based graph structures.

3.7.3.5 Interfaces

For using the functionality in the modelling tool, a trigger must be created, over which the export can be started, like a menu entry or a button. The export should not have a complex interface, but the information should be available in the models and the trigger should not require a lot of additional information.

How this information is then integrated into the DAI-DSS and the design methodology and the layer approach is still an open question. But to exchange the information, additional interfaces between the different technical components may be necessary. Here a way to save the data and make them accessible to other DAI-DSS components will be established. Therefore, interfaces are needed which allow to provide and extract the information.

3.7.3.6 Integration into the DAI-DSS architecture

This service will be used within the configuration component of the DAI-DSS architecture, as here the design methodology and the layer approach are used. The plug-in will be directly linked to the modelling tool and the corresponding service will be established also in the configuration component. How the information is shared and related to the knowledge base is still an open point, as the software for the knowledge base does not directly support knowledge graphs.

3.7.4 Outlook

After having a first formulation of a methodology and knowledge-driven experiments applied in this research track, the focus will shift in the next phase towards selecting and applying data-driven techniques to the concept. Therefore, we are currently investigating initial prototypes for machine learning-based worker to line allocation and optimal assignment. Adding further AI techniques aims at promoting the discussion on a common meta-model and its relevant parameters for the describing the three-layered approach. But not only the application of the introduced modelled-based design methodology will be researched, but also how it can be supported by utilizing a semantic rich and bidirectional data exchange between its steps and the corresponding tools.

3.8 Methods and Services for Reliable and Trustworthy AI

3.8.1 Overview

The following section focuses on the methods used for the research conducted on the topic of reliable and trustworthy AI, as described in Section 2.8. First, we focus on the studies that we did on transparency in general, under Sections 3.8.2 and 3.8.3 and afterwards we talk about the application of our findings (3.8.3) and the evaluation (3.8.4) of the DAI-DSS.

As reliability and transparency are important prerequisites for trust[135] and more additional research is necessary, especially with lay users, we conducted a qualitative focus group study, as detailed in Section 3.8.2, to find out more about the requirements, the users have towards AI systems regarding their transparency. This approach was

[134] https://www.ontotext.com/products/graphdb/?ref=menu (visited: 02.04.2024)
[135] Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. Academy of Management Annals, 14(2), 627-660.

combined with a qualitative study, described in Section 3.8.3, focusing on the comparisons of the different types of transparency that can be introduced to an AI system and can have different impact on users trust into the system.

The goal is of course to implement the findings of our studies into the DAI-DSS, to ensure a good level of transparency and trust into the system. To ensure that this takes place, as described under 3.8.4, we conducted a workshop with the different FAIRWork Partners, developing services for the DAI-DSS and will conduct more in-depth discussions in the future. Lastly, we also provide insight into our workshop, talks and questionnaires that we did with employees at the use case partner FLEX in Althofen and Timisoara in Section 3.8.5, to gather their input and obtain insights into the status quo before the introduction of a DAI-DSS.

3.8.2 Method: Qualitative Focus Groups about AI Transparency

3.8.2.1 Motivation and Reference to FAIRWork Use Case

In order to obtain trustworthiness in the DAI-DSS, the user has to be taken into account. Broadly spoken, good performance and the communication that technology provides such results reliably, establish trust[136]. What is more, Explainability has long been claimed as a way to increase trust and acceptance (Arrieta et al., 2020[137]; Miller, 2018[138]). For a long time, mainly developers and computer scientists have been researching this rather technological aspect of transparency. However, different stakeholders need to be addressed differently. IT experts will expect other information from a system that lay users (Mohseni et al., 2021[139]; van Nuenen et al., 2020[140]). However, this perspective neglects the high importance of understandability and the users' perspective on enhancing usage. As in FAIRWork, the users will mainly be people who are not computer experts but domain experts or workers their understanding of transparency has to be investigated. Lay users and their expectations, requirements, and usage of AI systems have to be taken into account to identify ways of how to build trust. That is transparency that is implemented successfully should increase the trust, acceptance, and usage of a DAI-DSS and provide autonomy to the involved people – which could be regarded as central for a legitimated representation.

Previous research shows that transparency can increase trust in systems (Mohseni et al., 2021). In case of mistakes, transparency has been regarded as a way to maintain trust and usage (Werz et al., 2021)[141]. On the other hand, previous research has found inconclusive results on the effects of transparency on trust and usage. In part, this goes back to the fact that there is no common understanding of what transparency is or what people require from it.

3.8.2.2 Innovation beyond the State-of-the-art

Therefore, to set up transparency in a DAI-DSS, it is necessary to gain a comprehensive understanding of transparency with a focus on what lay people require from it. This might depend on different system factors such as error relevance of a result or privacy concerns, they were under investigation, too.

[136] 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.

[137] 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

[138] Miller, T. (2018). *Explanation in Artificial Intelligence: Insights from the Social Sciences* (arXiv:1706.07269 [cs]). arXiv. http://arxiv.org/abs/1706.07269

[139] 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

[140] 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

[141] 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* (pp. 504–511). Springer International Publishing. https://doi.org/10.1007/978-3-030-50726-8_66

Therefore, the following trajectory looked into lay people's understanding of transparency and their requirements toward systems and their transparency. An initial overview of the results is currently being published in Werz et al (in press).

3.8.2.3 Description of Functionality

In order to increase trust, it is important to understand the different requirements, the users have towards AI systems with regard to their transparency. We need to understand these requirements independently of what is technically possible. Based on these insights, the AI services have to develop ways to make the DAI-DSS trustworthy and reliable.

3.8.2.4 Experimental method

A qualitative analysis of three focus group discussions took part to identify user requirements for AI transparency. The analysis focused on lay users' transparency needs beyond technical aspects and examined how different AI system factors influence these requirements. Three focus groups had taken place in which participants discussed three fictitious AI applications for lay usage: a financial investment app, a mushroom identification app, and a music selection app, each representing distinct system factors. A pre-test ensured users perceived the apps according to their assumed system factors. A total of 26 individuals, including 15 women, participated in the sessions.

In three rounds, the participants discussed three exemplary AI apps along with the following questions:

1. What does the AI app need to explain?
2. Under what conditions would you use the app?
3. How do you react when you realize that the app is wrong?

The analysis involved transcribing and anonymizing video recordings, followed by qualitative content analysis to identify categories such as "local" and "global" transparency across the three apps, revealing insights into the impact of system factors on transparency requirements.

3.8.2.5 Results

After transcribing, clustering, and analysing the results, the results indicate that the lay understanding of transparency goes beyond technical explanations. Three pillars of transparency requirements emerged. First, the importance of the domain of the application and prior experiences with domain and system(s), second the importance of background information beyond local and global Explainability, and third the effect of the system factor error-significance.

Participants' prior experiences significantly shaped their attitudes and transparency needs. For instance, scepticism towards the Finance App stemmed from negative financial experiences with apps and banks, while the Music App perceived the highest willingness for testing, reflecting users' familiarity with music services like Spotify. Prior experiences encompassed technical systems, interactions with institutions, and negative sentiments towards entire domains. Participants' experiences with similar systems influenced their expectations and demands for transparency in new ones. The analysis revealed that transparency concerns often focused on specific aspects rather than entire systems. Discussions on the Finance App centred on background processes, leading to demands for security measures and information about the app's business model. In contrast, discussions on the Music App focused on data privacy due to the perceived sensitivity of voice data.

The study also highlighted that for lay users the boundaries between global and local transparency are blurred. They did not make a clear distinction between the two, with transparency concerns extending beyond Explainability to include other aspects. Furthermore, the system factor of error significance played a crucial role in shaping

transparency requirements. Higher error significance, as perceived in the Mushroom and Finance Apps, prompted greater demand for background information and assurances.

The influence of novelty was evident, particularly in discussions about the Music App, where transparency needs extended beyond music selection to include language analysis and mood identification. The results showed the importance of user involvement in the design process to identify sensitive aspects and tailor transparency measures accordingly.

Overall, transparency requirements varied based on the AI type, application domain, and users' previous experiences with both the domain and specific systems. The results highlight the dynamic nature of transparency demands, which evolve with changes in system features and user experiences. Lastly, the factor of error significance intensified transparency concerns and requirements across all applications, underscoring the need for transparent and accountable AI systems.

3.8.2.6 Integration into the DAI-DSS Architecture

Together with the next study, the results flow into guidelines on how to provide transparency in AI services. What is more, the transfer of the results into the developers' perspective and the FAIRWork DAI-DSS architecture will be described in more detail in Section 3.8.4 covering the topic of the application of transparency.

3.8.3 Method: Quantitative Experiment comparing AI Transparency Methods (completed)

3.8.3.1 Motivation and Reference to FAIRWork Use Case

While computer researchers have been claiming that AI transparency will increase trust, the results of social science are inconclusive. Results show that depending on the context, implementation, and target group, transparency can have paradoxical effects and even diminish trust (Daschner & Obermaier, 2022[142]; Springer & Whittaker, 2018[143]; Yu et al., 2017[144]). Beyond the fact that there is no common understanding of transparency – which has been investigated with the previous approach – there are many ways to technically implement transparency. There are various ways to distinguish types of transparency, but it is important to note that they are not always clearly separate, and there are many overlaps in terminology (Ali et al., 2023[145]; Mohseni et al., 2021). Besides other classifications, one prominent and content-related differentiation is local and global explanations. Local explanations refer to single results and explain WHY an algorithm came to a certain result. Global explanations explain the entire system, specifying HOW it works (Molnar, 2019[146]). Global explanations are often prospective, meaning the transparency explanation is determined before the system's response is given. On the other hand, local explanations are often retrospective, meaning the explanation for the system's response is presented after or with the system's response (Carvalho et al., 2019[147]; Molnar, 2019). One potential advantage of local explanations is their continuous presentation on a "case-by-case" basis. Studies have shown that trust and performance outcomes were more positive in the local condition compared to the global condition (Wanner et al., 2022[148]; Herm

[142] 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

[143] Springer, A., & Whittaker, S. (2018). What are you hiding? Algorithmic transparency and user perceptions. *AAAI Spring Symposium Series*, 1–4.

[144] Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., & Chen, F. (2017). User Trust Dynamics: An Investigation Driven by Differences in System Performance. *Proceedings of the 22nd International Conference on Intelligent User Interfaces*, 307–317. https://doi.org/10.1145/3025171.3025219

[145] 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

[146] Molnar, C. (2019). *Interpretable Machine Learning (1st edition)*. Christoph Molnar (CC Attribution 2.0). https://christophm.github.io/interpretable-ml-book/index.html

[147] 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

[148] Wanner, J., Herm, L.-V., Heinrich, K., & Janiesch, C. (2022). A social evaluation of the perceived goodness of explainability in machine learning. *Journal of Business Analytics, 5(1),* 29–50. https://doi.org/10.1080/2573234X.2021.1952913

et al., 2023[149]). However, in some cases, such as in a field experiment by Kizilcec (2016)[150], trust decreased when both global and local explanations were provided about the functionality of a grading algorithm. The global explanation alone was perceived as more trustworthy and understandable. Similarly, research by Lim and Dey (2009)[151] found that individuals preferred global explanations about the general function of an algorithm.

The results of local or global Explainability are therefore inconclusive in their effects on end users' attitude towards AI. What is more, several technological approaches such as heatmaps or – simpler – security scores of AI have not yet been compared regarding their effects on trust and usage of AI, while the claim for transparency measure prevails. Together with the claim for more general and informational transparency we identified in the focus groups (see Chapter 3.8.3), these findings led us to the question of how different types of transparency effect usage and trust of systems.).

3.8.3.2 Innovation beyond the State-of-the-art

In addition to the previous approach that qualitatively investigated the requirements of lay users towards (transparent) AI systems, an additional approach was taken with a quantitative experiment. In this experiment, we intentionally investigated trust and usage, despite other researchers that mixed these two concepts (e.g., Schmidt et al., 2020; [152]Zhang et al., 2020[153]). However, other results suggest that while trust can be an important prerequisite of usage, they are different concepts (Daschner & Obermaier, 2020[154]).

Based on the described findings, several hypotheses have been formulated:

- Transparency in algorithms compared to no transparency leads to changed usage of the algorithm.
- Transparency in algorithms compared to no transparency leads to changed trust in the algorithm.
- Different types of transparency differ in terms of usage.
- Different types of transparency differ in terms of trust.

A study was set up to test these hypotheses through a Wizard-of-Oz experiment, where participants believe they are interacting with an AI system, while it is actually a pre-programmed sequence. Usage is operationalized as Weight of Advice (WoA). Five conditions are compared in a within-design to achieve a high effect size.

3.8.3.3 Experimental Method

The conducted experiment aimed to understand how different types of transparency influence trust and usage.

After excluding incomplete as well as outliers and other deviating data, the data set consisted of n=151 participants. In the study, we set up a 2x2 design of transparency, i.e. we different between local and global as well as between information about functionality (Explainability) and about accuracy. The resulting four transparency conditions can be seen in Table 4.

[149] Herm, L.-V., Heinrich, K., Wanner, J., & Janiesch, C. (2023). Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability. *International Journal of Information Management*, 102538. https://doi.org/10.1016/j.ijinfomgt.2022.102538

[150] Kizilcec, R. F. (2016). How much information? Effects of transparency on trust in an algorithmic interface. *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, 2390–2395. https://doi.org/10.1145/2858036.2858402

[151] Lim, B. Y., & Dey, A. K. (2011). Investigating intelligibility for uncertain context-aware applications. *Proceedings of the 13th International Conference on Ubiquitous Computing*, 415–424. https://doi.org/10.1145/2030112.2030168

[152] Schmidt, P., Biessmann, F., & Teubner, T. (2020). Transparency and trust in artificial intelligence systems. *Journal of Decision Systems, 29(4)*, 260–278. https://doi.org/10.1080/12460125.2020.1819094

[153] Zhang, Y., Liao, Q. V., & Bellamy, R. K. E. (2020). Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency,* 295–305. https://doi.org/10.1145/3351095.3372852

[154] 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

Table 4: Four transparency conditions of the study.

		global	local
Functionality		Global functionality (Condition B)	Local functionality (Condition E)
Accuracy		Global accuracy (Condition C)	Local accuracy (Condition D)

These four conditions were compared to one another as well as with an initial, non-transparent condition A.

All participants ran through all conditions: While the first condition was always A, the other four were randomized in order. Within each condition, participants completed three trials of estimation task.

During the course of each task, participants initially provided their estimate of the weight of fruits and/or vegetables depicted in a photo. Using a slider, they could set their response on a visual analogyue scale ranging from 0 g to 500 g. Clicking "Next" took the participants to the next page, where the algorithm's estimate was presented, and they could provide their final answer. After each “algorithm”, i.e. each transparency condition, participants answered questions regarding their trust into the previous algorithm.

The four transparency conditions were as follows:

- B: Global functionality: General information about the algorithm, who developed it, how it is working and how it was tested
- C: Global accuracy: One percentage value for the algorithm indicating its accuracy
- D. Local accuracy: One range per result indicating its uncertainty regarding this specific estimation (between 4 and 7 grams)
- E: Local functionality: Heat map picture for each estimation task that was indicated to “depict how the algorithm worked”

This set-up enabled us to calculate so-called Weight of Advice, which indicates the influence of algorithm advice on the final result of each estimation representing the usage of the algorithm. It indicates the usage between 0 (no usage) and 1 (complete adoption of algorithmic advice). In addition, the trust scale provided a mean trust measurement for each transparency condition. Different ANOVAs were conducted to compare the conditions and test the hypotheses.

3.8.3.4 Results

The results of this study supported the assumption that the presence of transparency significantly influences trust (H3) and usage (H4) of algorithmic recommendations. Furthermore, the results indicated that trust significantly varied among the four examined types of transparency (H1). However, the usage of the algorithms did not vary depending on the type of transparency, thus failing to confirm H2.

The following Figure 24depicts the mean usage in the neutral (A) and four (B-E) transparency conditions. The whiskers indicate standard deviations (SD) for each mean.

Figure 24: Effect of different transparency conditions on usage of the algorithm.

The results indicate that transparency measure does make a difference and increase trust – but they are differently effective in doing so. The most successful transparency measure were the two global transparency types. Regarding the one about global functionality, this is in accordance with the focus group results where participants demanded background information about the developers, their work and the ways the algorithm was tested. This seems to indicate that such general information is important for building trust in the first place.

The not as successful results of local explanations do not indicate that this type of transparency is superficial – to the contrary. As other research shows, they seem to be very important during the usage of algorithms especially when they fail or reach inconclusive results (Alam & Mueller, 2021[155]; Kim et al, 2023[156]).

As the study and results have just been finished at the beginning of 2024, we are currently finalizing the analysis and are going to publish the respective results in the following months.

3.8.3.5 Integration into the DAI-DSS Architecture

As described above, the results of the two studies flow into guidelines on how to provide transparency in AI services. The developer's perspective and the FAIRWork DAI-DSS architecture will be described in more detail in the following chapter.

[155] Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making, 21(178)*. https://doi.org/10.1186/s12911-021-01542-6

[156] Kim, S. S. Y., Watkins, E. A., Russakovsky, O., Fong, R., & Monroy-Hernández, A. (2023). "Help me help the AI": Understanding how explainability can support human-AI interaction. *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, 1–17. https://doi.org/10.1145/3544548.3581001

3.8.4 Method: Practical Application of Transparency in Different DAI-DSS Services (ongoing)

3.8.4.1 Motivation and Reference to FAIRWork Use Case

It is very important, to apply the knowledge gained in the other methods described under 3.10 to the FAIRWork Use case and the different Services of the DAI-DSS. Therefore, this method aims to provide a practical application to gain from previous results.

3.8.4.2 Innovation beyond the State-of-the-art

As has been described in 3.8.2 and 3.8.3, 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 for the individually best ways to enable transparency and apply the findings from 3.8.2 and 3.8.3.

3.8.4.3 Description of Functionality

Through Workshops and one-on-one meetings, the different DAI-DSS services are to be better understood and analysed, to find the potential of providing transparency to the different services and therefore enhance their trustworthiness. The solution must be tailored to each service, which is why close cooperation and consultation with partners is important.

3.8.4.4 Experiments

A workshop was conducted with different FAIRWork partners to identify different ways of providing transparency to the FAIRWork Services that will find application in the FAIRWork DAI-DSS. This workshop served to raise awareness with the partners for the importance of transparency and different possibilities to apply it to their specific service. It also served to gain first insights into and knowledge about the different services and how to enable them to be more transparent.

With the insights from this workshop, one-on-one meetings with the different partners will be conducted to work more detailed on applying transparency for their specific services, to enable higher trustworthiness. The goal of these meetings is to answer the following questions: How to set up transparency for the respective service? How reliable are the services and how can the extent of that reliability be made transparent? How to measure transparency and reliability? How can the results of these measurements be presented to lay users? How to measure trustworthiness regarding the DAI DSS Services?

Lastly, a comparison of (aspects of) transparency applications will be carried out with lay users in the form of an experiment or a vignette study. The goal for this is to verify the previous results and to evaluate our success in enabling the trustworthiness of the DAI-DSS Services through increased transparency.

3.8.4.5 Results

From the first workshop, a few preliminary results have been gathered: 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 services processing of the data. Therefore, these processes must be regarded for each service on its own, in greater detail.

The different DAI-DSS services vary in their complexity to explain, as some are much more easily understood and therefore explained than others. But the complexity also varies, depending on the use case and how much data is being processed.

Two points were identified, to be of special interest for explanations: The first one is whenever user expectations are violated. If the output of the system is something different than 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.

3.8.4.6 Integration into the DAI-DSS Architecture

The integration into the DAI-DSS architecture will happen together with the partners, that provide services. Here in workshops and one-on-one meetings, solutions tailored to each service are being developed and will then be implemented by the partner responsible for the corresponding service.

3.8.5 Method: Evaluation of different DAI DSS services concerning Trustworthiness (ongoing)

3.8.5.1 Motivation and Reference to FAIRWork Use Case

To validate previous results, contact with and feedback from the end users and FAIRWork use cases is essential. Therefore, this method aims to gather input from the users and check the acceptance of the developed solutions in the use cases.

3.8.5.2 Innovation beyond the State-of-the-art

The DAI-DSS system developed in FAIRWork is new and not yet tested. In order to transfer the system into practice, the integration and trust of its potential future users is essential. This is why the trust and acceptance of the employees need to be evaluated, to prove that the system can realize its desired potential.

As a pre-test and to gain a baseline, a questionnaire study was conducted. This questionnaire consisted of three parts, in addition to the demographics section. The employees' workload, their attitude regarding decisions made by their supervisors as well as their attitude regarding automated systems.

The employee workload could give important insights regarding specific tasks or employee groups, that benefit the most from a DAI-DSS as well as regarding any changes in the workload through the introduction of the DAI-DSS. To assess the employees' workload, the NASA Task Load Index (NASA TLX)[157] is a common tool, that is a well-validated and proven tool. However, in its original shape, the NASA TLX refers to the performance of one specific task. To gain an understanding of the general workload, during a typical workday, this questionnaire was slightly altered, to refer to not a specific task, but rather the last week.

The second part of the questionnaire was concerning the employees' attitude towards management decisions. This was done to gauge the current level of trust regarding supervisors and the attitude regarding their decisions before a DAI-DSS is introduced. For this, the Trust Scale by Merritt (2011)[158] was chosen, as it is a very short and easily understood scale that is still widely used and captures both the general construct of trust, but also the trustworthiness components of ability and benevolence (Mayer et al. 1995)[159]. For the present use, the scale was adapted so that it refers to decisions made by the direct superior.

The questionnaire closes with a few questions regarding attitude regarding automated systems, to have a baseline for the employees' attitude in that regard. The questions have been taken from the Propensity for Trust Subscale

[157] Hart, S. (NASA Ames Research Center Moffett Field, CA United States) (1983) *Task Load Index – NASA TLX*
[158] Merritt, S. M. (2011). Affective processes in human–automation interactions. Human Factors, 53(4), 356-370.
[159] Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. Acad. Manag. Rev. 20, 709–734. doi: 10.5465/amr.1995.9508080335

of the Trust in Automation Scale by Körber (2019)[160]. This scale had the advantage of being a validated scale in the German language and therefore didn't need to be changed or translated to be used with the Austrian employees.

3.8.5.3 Description of Functionality

With the use case partner FLEX, especially employees from their factory in Austria, an evaluation concerning trustworthiness takes place. For this, a status quo survey took place and the introduction of DAI-DSS services will be accompanied by further studies to identify potential changes in attitude compared to the status quo and evaluate the implementation of the services.

3.8.5.4 Experiments

To explore the trustworthiness of the DAI-DSS services in the FAIRWork project from a user-centred perspective, the lead use case is the FLEX company site in Austria. This is due to its easier reachability and missing language barrier, while CRF in Italy will provide additional input, where needed.

As a status quo investigation, a questionnaire study was conducted with 14 participants from Flex in Austria. The questionnaire contained a demographics section, as well as three additional topics: the employees' workload experienced over the last week, their attitude regarding decisions made by their supervisors as well as their attitude regarding automated systems. This was done, to gather insights into the status quo before the implementation of a DAI-DSS.

Additionally, we gathered further input from employees at the FLEX company site in Romania, during a workshop at the science fair, which was conducted during the FAIRWork Partner-Meeting in November 2023. Here we gathered employees fears and hopes for a decision support system.

The introduction of DAI-DSS services will be accompanied by further surveys to identify potential changes and evaluate the implementation.

3.8.5.5 Results

So far, the pre-test has happened via a questionnaire study at FLEX in Austria. From this, we will provide some results in the following:

Apart from the demographics, the questionnaire contained three parts (see 3.10.5.5). In the first part, regarding the workload experienced over the last week, the workers described their workload as moderate in total. On average, a manageable workload with a balanced distribution across various dimensions of the NASA TLX scale. One notable result was that supervisors reported a relatively high mental demand during their week, which might be an indication, that a DSS might be especially useful for supervisors.

In the second part of the questionnaire, the employees reported highly positive views toward decisions made by their supervisors. This suggests a consensus among participants in perceiving their supervisors' decisions as favourable, though influences of social desirability are possible. Since employees seem to be content with the decision-making of their supervisors, this contentedness must not be allowed to deteriorate through the introduction of the DAI-DSS.

The third part of the questionnaire contained questions relating to the attitude regarding automated systems. Here a moderate level of trust and attitude toward automated systems was displayed by the employees. However, one

[160] Körber, M. (2019). Theoretical considerations and development of a questionnaire to measure trust in automation. In Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20 (pp. 13-30). Springer International Publishing.

item scored lower than the rest. This displayed a need, to be careful with unknown automated systems. So, the participants were generally open to automated systems. However, employees see the need to be cautious with unknown systems at first. Which makes a careful introduction of the DAI-DSS even more important.

Regarding the qualitative results of the workshop during the science fair, the employees were able to name many hopes as well as fears connected to the introduction of a DAI-DSS. Here a division took place regarding a system supporting the employees' own decisions as well as a system being used in the decisions of their supervisors. Regarding a decision support system for their own decisions, the main fears are that the system might propose wrong or suboptimal solutions and that there might not be a way to be sure that the proposed decision is the best, due to missing transparency. A related fear was that the database and input data couldn't be checked by the employees, whether it is correct and up to date. For all these reasons, the system might lead the worker to a wrong action, which he wouldn't have chosen without the system's suggestion. Other fears mentioned by the employees were that the system might only give one solution for a given task so that the employees don't have a choice, that the system might require a lot of input from them or that one would become too comfortable with and reliant on the system, not questioning its proposed solutions and therefore making mistakes one wouldn't make now. However, they also see the potential for the use of a DAI-DSS. The employees hope for faster and easier answers and decisions for their problems, through the use of a decision support system. They wish for help in their work by supporting them in difficult situations. Lastly, they see a potential to get better and easier insights into company priorities, for example regarding budget or energy consumption) which can be taken into account by using the decision support system.

When asked about a situation, where their supervisors would use a decision support system, the fears and hopes were different. Here the employees feared that people might not understand why a decision was made about them in a specific way and might therefore feel like they were treated unfairly. Additionally, inaccurate or out-of-date information in the database might also lead to an unfair treatment of the employee. On the other hand, the hope is that people can understand the decisions and are being helped by the system. The system might be more neutral than a human decision-maker and therefore treat the employees more fairly and unbiased. The system could also include not only the performance parameters of the workers but also their preferences for different tasks or workplaces, which might lead to higher work satisfaction.

3.8.5.6 Integration into the DAI-DSS Architecture

The Integration of the gained knowledge and results into the FAIRWork DAI-DSS architecture takes place within the practical application that is described in detail in the previous chapter.

3.8.6 Outlook

The future outlook has been detailed in the above sections already, however, to summarise:

- Guidelines for transparent AI development and design are to be developed and the aim is to build a transparency matrix that relates system factors with required transparency measures that have to be taken into account to foster trust.
- Regarding the application of our findings to the DAI-DSS, one-on-one meetings with the different partners will be conducted to work more detailed on applying transparency for their specific services, to enable higher transparency and trustworthiness of the DAI-DSS system and its services.
- We aim to continue the questionnaire study that was conducted at FLEX in Althofen, with employees of CRF in Turin
- The introduction of DAI-DSS services will be accompanied by further surveys with employees at the production sites of the use case partners, to identify potential changes through introducing the DAI-DSS and evaluate the implementation.

- Lastly, a comparison of (aspects of) transparency applications will be carried out with lay users in the form of an experiment or a vignette study. The goal for this is to verify the previous results and to evaluate our success in enabling the trustworthiness of the DAI-DSS Services through increased transparency.

3.9 Ethical Watchdog

3.9.1 Overview

The ethical watchdog concept can be interpreted as a collection of tools that create trust, transparency, and fairness. This could involve certification, a certain type of modelling or models, questionnaires, evaluation tools or other features and services to estimate compliance with ethical topics.

3.9.2 Model-based assessment of ethical criteria to ensure compliance

3.9.2.1 Motivation and Reference to FAIRWork Use Case

Society's great expectations for AI's transformative potential and significant achievements in utilizing data for machine learning can lead to the conclusion that the explicit and manual creation of models is no longer required or beneficial in system development. One should be cautious about drawing such conclusions, as AI in general and machine learning in particular come with specific prerequisites and limitations that must be considered when applying and implementing their methods in various scenarios[161]. Challenges and limitations can include the availability and quality of data, the possible manifestation or even amplification of bias introduced in the trainings data or algorithms[162] and the "black box" nature of some AI methods, which makes is difficult to trust and interpret the produced results[163]. Some of these limitations can be tackled when combining conceptual modelling with AI. In FAIRWork we use conceptual modelling to depict decision processes and these processes are used as basis for configuring the decision support system. As these models influence the outcome of the decision support system and therefore the decision makers and other stakeholders involved in the resulting decisions, it is important to assess them before they are approved for configuration. The assessment should consider ethical criteria and ensure compliance with them. Once a model has passed evaluation and received final approval, it requires assurance that no modifications will be made to the model afterwards. BOC proposes the involvement of digitally signing the model to achieve two objectives: firstly, to authenticate the signatory's identity, and secondly, to enable tracking of any changes made to the model after it has been signed.

3.9.2.2 Initial Experiments

Building upon research results of previous project complAI[164], certification and compliance services are applied to FAIRWork specific- configuration. For this purpose, the methodology presented in previous Sections serves as starting point for considerations (e.g., different layers might require different levels of trust and proper experts that are qualified for the certification).

[161] Fettke, P. (2020). Conceptual Modelling and Artificial Intelligence: Overview and research challenges from the perspective of predictive business process management. In Companion Proceedings of Modellierung 2020 Short, Workshop and Tools & Demo Papers co-located with Modellierung 2020 (Vol. 2542, pp. 157–164). CEUR-WS.org. https://ceur-ws.org/Vol-2542/MOD-KI4.pdf

[162] Belenguer, L. (2022). AI bias: Exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry. Ai and Ethics, 2(4), 771–787. https://doi.org/10.1007/s43681-022-00138-8

[163] von Eschenbach, W. J. (2021). Transparency and the Black Box Problem: Why We Do Not Trust AI. Philosophy & Technology, 34(4), 1607–1622. https://doi.org/10.1007/s13347-021-00477-0

[164] https://complai.innovation-laboratory.org/

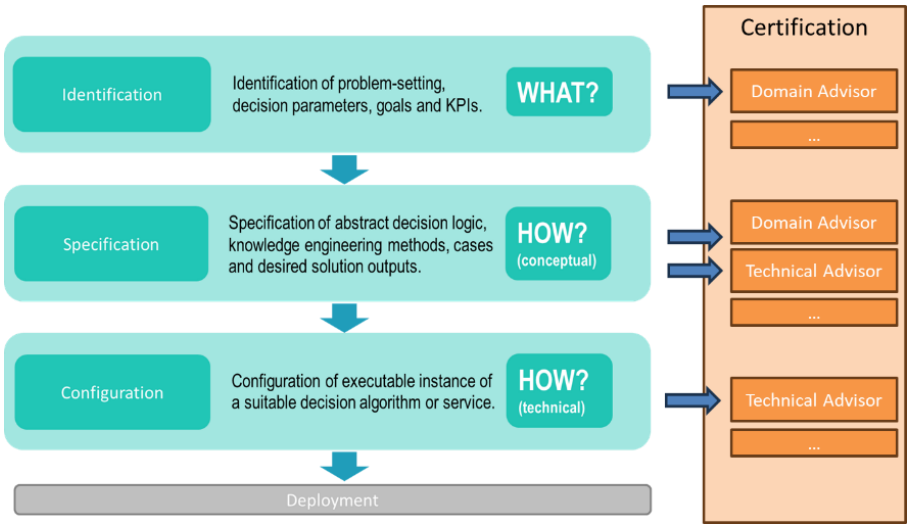


Figure 25: Methodology extended with certification aspects.

The FAIRWork "Workload Balance" scenarios serves as a specific example and as starting point for certification considerations for AI services. Due to the advantages of models to create transparency and common understanding they can be used as basis for the expert signatures or for applying questionnaires. One example for such a validation process can be seen in Figure 1Figure 26, for the Fuzzy Logic approach. Domain experts are important to verify the used input parameters to determine "efficiency of workers", the ranges of the membership functions for all input and output parameters, but also for the formulated rules in the rule base. Technical advisors on the other hand are important for signing the correctness of the applied calculations, Fuzzy Inference and Defuzzification methods.

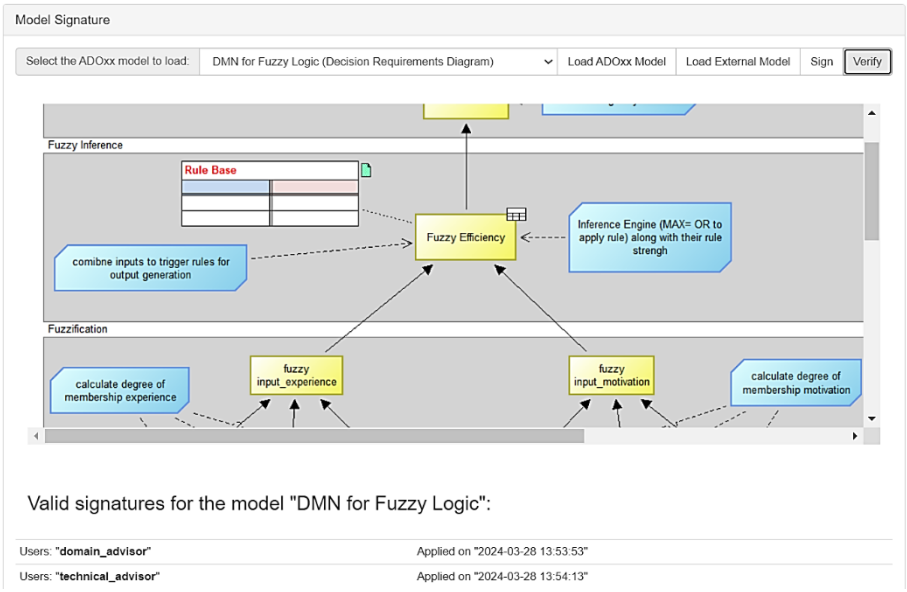


Figure 26: Example of model-based signing service of "Specification Layer".

Another example based on modelling is by adding questionnaires for different parts. For example, the input parameter "motivation" of a worker is very sensitive information, and supervisors should not have access to this knowledge due to privacy reasons. For this purpose and in addition to the certification services, questionnaires to examine compliance of sensitive topics can be utilized. In Figure 27 you can see an example for such a situation

evaluated with questionnaires. One exemplary questionnaire on data privacy and awareness of workers regarding the AI services is illustrated in Figure 27.

If the questions are answered in favour of concepts like transparency, trustworthiness, privacy etc., “green flags” indicate that there is no issue expected for the evaluated topic (e.g. data privacy of workers for the input “motivation”) otherwise “red and yellow flags” mark the corresponding item (e.g. input “experience”) and highlight problematic elements and values in the model. An example is illustrated in Figure 28

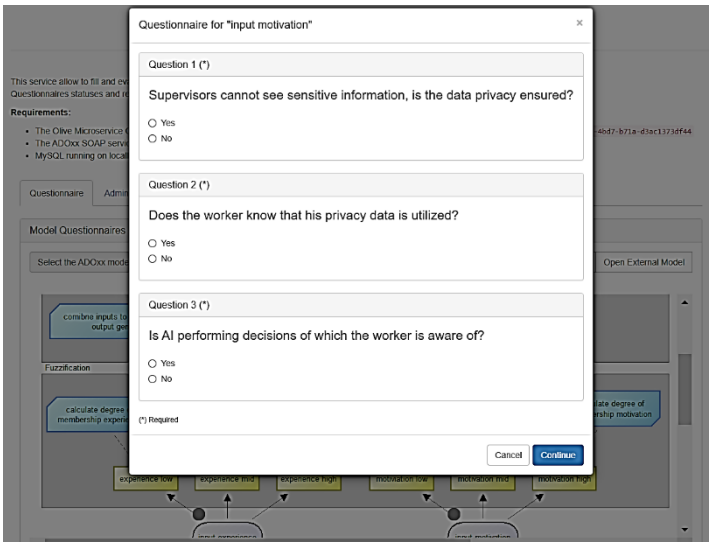


Figure 27: Exemplary questionnaire.

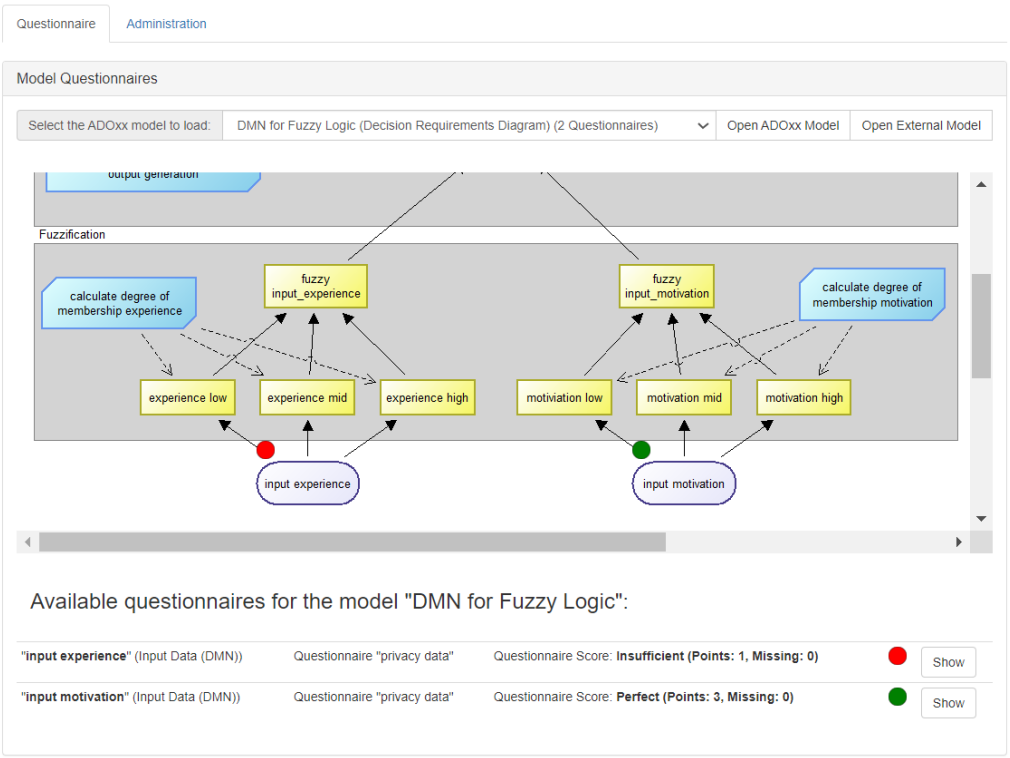


Figure 28: Evaluation of the Fuzzy Logic with the Questionnaire.

Based on these two examples, many considerations can be made. Questions can include but are not limited to:

- **Certification**: Who is allowed to certify which layers? How many certifications are required so that the overall certification can be achieved? Is certification democratic, in the sense of what is the weight of the individual certificates? If one expert is not available, who is the replacement?
- **Questionnaire:** Are the questions appropriate? Where must they be applied? When are they indicated as fulfilled or not? Who answers these questions?

3.9.3 Concept: Features of Human-centred Machine Learning Model

3.9.3.1 Motivation and Reference to FAIRWork Use Case

With the upcoming of Industry 5.0 methodologies there will be an increasing number of machine learning models in use that process human-centred data either implicitly or explicitly. There are many features that are necessary to describe the level of the consideration of ethical aspects of these models that either eventually adapted to a final state or that would represent an adaptable model that can be parametrised by future incoming human-centred data.

The transparency about these ethical aspects would enable the human decision maker but also any automatised service that would depend on that input of the machine learning model to weigh the relevance of the model for any further impact in the socio-technical system in an appropriate manner.

In FAIRWork, we are developing machine learning models on the basis of wearable biosignal sensor-based measurements. These models are then input to computations about resilience scores and these are further considered, e.g., by the worker allocation component. In the future, these components are considered to be adaptive by taking periodic, potentially, even daily input from the workers. It is of high importance for these services to understand the relevance of these models, such as, the validity for the purpose that they are implemented for, in particular, the risk for discriminating any persons of the worker community.

3.9.4 Concept: Supporting Decision Explanations through Conceptual Modelling

This section introduces a concept which aims to use diagrammatic, conceptual modelling as a way to support the explanation of made decisions within the DAI-DSS. The concept is based on (Muck et. al, 2024)[165].

It is not always easy for people to trust decision support systems, especially if AI algorithms are used, as they often do not completely understand how the decision-making process looks like and why specific decisions were suggested. In such cases explainability is important, as it explains the decision in a way that can be interpreted by the involved humans (Ashoori & Weisz, 2019)[166]. Such explanations must be tailored to the stakeholders, who should understand it, which means that often it is not sufficient to just explain the AI algorithm but also to provide context from the domain of the user (Bayer et. al., 2022)[167]. Therefore, we want to use diagrammatic, conceptual models to support the explanation of made decisions within the DAI-DSS. Conceptual models and their diagrammatic representation support the understanding of complex systems by using spatial information (Larking & Simon, 1987[168]; Harel & Rumpe, 2000[169]). Such an understanding is further improved if concepts from the domain

[165] Muck C, Tschuden, J., Zeiner, H., Utz W. (2024). Explainability of Industrial Decision Support System using Digital Design Thinking with Scene2Model, Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences, Nice, France, July 2024. (accepted – but not published)

[166] Ashoori, M., & Weisz, J. D. (2019). In AI We Trust? Factors That Influence Trustworthiness of AI-infused Decision-Making Processes. ArXiv, abs/1912.02675. https://api.semanticscholar.org/CorpusID:208637106

[167] Bayer, S., Gimpel, H., & Markgraf, M. (2022). The role of domain expertise in trusting and following explainable AI decision support systems. Journal of Decision Systems, 32(1), 110–138. https://doi.org/10.1080/12460125.2021.1958505

[168] Larkin, J. H., & Simon, H. A. (1987). Why a diagram is (sometimes) worth ten thousand words. Cognitive Science, 11(1), 65–100.

[169] Harel, D., & Rumpe, B. (2000). Modeling languages: Syntax, semantics and all that stuff, part i: The basic stuff.

are taken, which further improves the understanding for domain experts (Karagiannis et. al., 2016[170]; Frank, 2013[171]).

As FAIRWork already uses conceptual models to design the concrete decision support scenarios and support their implementation as introduced in Section 2.7. As we already create the models for designing and configuration, we also want to use them to support the explanation. Therefore, the aim is not to replace other explanation efforts, but so support them through the diagrammatic representation. Through the model-based design methodology we already have models and a mapping of models used to the decision scenarios. These can be reused to support the explanations of made decisions.

In addition, we want to analyse within this concept, how we can use data collected from the decisions made within the DAI-DSS to be shown in models to tailor it to concrete decision scenarios. Therefore, the models can be enriched with information form a concrete decision made within the DAI-DSS and adapt to visualize it to support the understanding. This should not only be done to have visual representation, but having a model and not just a picture, allows us to process the model information to further support the explanation. One possibility would be to use generative AI to create textual explanations based on the modelled information. But this investigation is still ongoing, and no initial prototype or experiment is yet available.

As we use multiple modelling methods within FAIRWork, this approach should be supported through a semantically rich information exchange framework, which allows to provide information from the models to the DAI-DSS, but also to enable to provide data back to the models in the future. This is a part of which will be researched in the concept described in Section 2.8.

[170] Karagiannis, D., Buchmann, R. A., Burzynski, P., Reimer, U., & Walch, M. (2016). Fundamental Conceptual Modeling Languages in OMiLAB. In D. Karagiannis, H. C. Mayr, & J. Mylopoulos (Eds.), Domain-Specific Conceptual Modeling: Concepts, Methods and Tools (pp. 3–30). Springer International Publishing. https://doi.org/10.1007/978-3-319-39417-6_1

[171] Frank, U. (2013). Domain-Specific Modeling Languages: Requirements Analysis and Design Guidelines. In I. Reinhartz-Berger, A. Sturm, T. Clark, S. Cohen, & J. Bettin (Eds.), Domain Engineering: Product Lines, Languages, and Conceptual Models (pp. 133–157). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-36654-3_6

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 D3.2 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 D3.3 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 **Explainable AI (XAI)**, often overlapping with Interpretable AI, or Explainable Machine Learning (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, 2023[172]; Longo et al., 2024[173]). 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[174]). 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[175]).

With the upcoming widespread us of artificial intelligence (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[176]).

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

[172] Mihály, Héder (2023). *Explainable AI: A Brief History of the Concept*. ERCIM News (134): 9–10.

[173] Longo, Luca; et al. (2024). Explainable Artificial Intelligence (XAI) 2.0: A manifesto of open challenges and interdisciplinary research directions. *Information Fusion*. 106. doi:10.1016/j.inffus.2024.102301

[174] Vilone, Giulia; Longo, Luca (2021). Notions of explainability and evaluation approaches for explainable artificial intelligence. *Information Fusion*. December 2021 - Volume 76: 89–106. doi:10.1016/j.inffus.2021.05.009

[175] Castelvecchi, Davide (2016). Can we open the black box of AI? *Nature*. 538 (7623): 20–23. doi:10.1038/538020a.

[176] Mehrabi Ninareh, Morstatter Fred, Saxena Nripsuta, Lerman Kristina, and Galstyan Aram (2021). A survey on bias and fairness in machine learning. *ACM Comput. Surv. (CSUR)* 54, 6 (2021), 1–35.

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 of AI Services

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 LIME (Local Interpretable Model-agnostic Explanations[177]), SHAP (Shapley Additive Explanations; Lundberg & Lee, 2017[178]; Aal et al, 2021[179]) or the What-if tool (Wexler et al., 2020[180]).

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 Locally Interpretable Model-Agnostic Explanations (LIME), DeepLIFT, and Layer-Wise Relevance Propagation.

XAI tools make it possible to explain and interpret the predictions of machine learning models. These 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.

4.3 Fairness in AI Services

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 FAIRWorks, “fairness” plays a role in several dimensions, as described in more detail, as follows, by several viewpoints.

4.3.1 Observations on Fairness in FAIRWork

Fairness in Decision Making

First, it is clear that numerous decisions are made by groups. Social choice theory[181] 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

[177] Ribeiro, M.T., Singh, S., and 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.
178 Lundberg, Scott M.; Lee, Su-In (2017). A Unified Approach to Interpreting Model Predictions. *Advances in Neural Information Processing Systems*. 30: 4765–4774. arXiv:1705.07874. Retrieved 2021-01-30.
[179] Aas, Kjersti, Martin Jullum, and Anders Løland. Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values. *Artificial Intelligence,* 298 (September 2021).
[180] J. Wexler, et al. (2020). The What-If Tool: Interactive Probing of Machine Learning Models, in *IEEE Transactions on Visualization & Computer Graphics*, vol. 26, no. 01, pp. 56-65, 2020.doi: 10.1109/TVCG.2019.2934619
[181] Sen, A. (1986). *Social choice theory*. Handbook of mathematical economics, 3, 1073-1181.

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[182]. 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 and Lu, 2019[183]).

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.[184]

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[185]).

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[186]. 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.

[182] De Jong, S., Tuyls, K., & Verbeeck, K. (2008). Fairness in multi-agent systems. The Knowledge Engineering Review, 23(2), 153-180.
[183] Jiang, J., & Lu, Z. (2019). Learning fairness in multi-agent systems. Advances in Neural Information Processing Systems, 32.
[184] 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
[185] Tolan, S. (2019). Fair and unbiased algorithmic decision making: Current state and future challenges. arXiv preprint arXiv:1901.04730.
[186] 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. Presented at the Estoril, Portugal. Richland, SC: International Foundation for Autonomous Agents and Multiagent Systems.

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[187]. 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[187]. 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 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[188,189]. 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 reproduce 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[190]).

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 Measures of Fairness

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[191]) 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

[187] 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

[188] 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. doi:10.3390/make4020026

[189] Jiang, J., & Lu, Z. (2019). Learning Fairness in Multi-Agent Systems. CoRR, abs/1910.14472. Retrieved from http://arxiv.org/abs/1910.14472

[190] 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.

[191] Pedro Saleiro, Benedict Kuester, Abby Stevens, Ari Anisfeld, Loren Hinkson, Jesse London, and Rayid Ghani. (2018). *Aequitas: A Bias and Fairness Audit Toolkit.* arXiv preprint arXiv:1811.05577 (2018).

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[192]). 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 Guttag (2019[193]) 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[194]) 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 & Guttag as well as from Olteano et al., integrating the survey of Mehrabi et al. (2021), as follows:

- **Measurement Bias**. Measurement, or reporting, bias arises from how we choose, utilise, and measure particular features (Surash and Guttag, 2019). 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 [195]).

[192] Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. (2018). *AI fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias.* arXiv preprint arXiv:1810.01943 (2018).

[193] Harini Suresh and John V Guttag. 2019. A Framework for Understanding Unintended Consequences of Machine Learning. arXiv preprint arXiv:1901.10002 (2019).

[194] Olteanu A, Castillo C, Diaz F and Kıcıman E (2019) Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries. *Front. Big Data* 2:13. doi: 10.3389/fdata.2019.00013

[195] Colin R Blyth. 1972. On Simpson's paradox and the sure-thing principle. J. Amer. Statist. Assoc. 67, 338 (1972),364–366.

- **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[196]) studied fairness definitions in political philosophy and tried to tie them to machine learning. Authors in [70] studied the 50-year history of fairness definitions in the areas of education and machine-learning. Hutchinson and Mitchell (2019[197]) listed and explained some of the definitions used for fairness in algorithmic classification problems. Saxena et al (2019[198]) 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[199]).

- **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.

[196] RDP Binns. 2018. Fairness in machine learning: Lessons from political philosophy. Journal of Machine Learning Research (2018).
[197] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un) fairness: Lessons for Machine Learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency. ACM, 49–58.
[198] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. 2019. How Do Fairness Definitions Fare?: Examining Public Attitudes Towards Algorithmic Definitions of Fairness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. ACM, 99–106.
[199] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare). IEEE, 1–7.

- **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):

- 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 artificial intelligence 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):

- **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.

Examples of some existing work and their categorisation into these types are shown in detail Mehrabi et al. (2021).

5 SUMMARY AND CONCLUSIONS

This report consists of three main parts. The first part (Section 2) provides an overview of the research tracks, covering research directions such as democratization of decision-making, digital Human Factors analytics, reliable and trustworthy Artificial Intelligence, as well as Artificial Intelligence and Multi-Agent Systems for improving decision support systems in manufacturing. 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 5) provides an overview on the publication efforts, categorised into already published and submitted work under review.

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.

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 Artificial Intelligence and Multi-Agent System technologies in decision support systems.

6 Scientific Dissemination

6.1 Publications Developed in the Context of the Research Collection

Published:

Muck, C., & Utz, W. (2023). A Recognition Service for Haptic Modelling in Scene2Model. In Proceedings of the AAAI2023 Spring Symposium on Challenges Requiring the Combination of Machine Learning and Knowledge Engineering (AAAI-MAKE 2023), CEUR Workshop Proceedings.

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. LOD 2023. Lecture Notes in Computer Science, vol 14505. Springer, Cham. https://doi.org/10.1007/978-3-031-53969-5_16

Paletta, L., Ayaz, H., Asgher, U., Eds. (2023). Cognitive Computing and Internet of Things, ISBN: 978-1-958651-49-0. DOI: 10.54941/ahfe1003983.

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 (2023) International Conference. AHFE Open Access, vol 73. AHFE International, USA. http://doi.org/10.54941/ahfe1003971.

Woitsch, R., Muck, C., Utz, W., & Zeiner, H. (2023). Towards a Democratic AI-based Decision Support System to Improve Decision Making in Complex Ecosystems. In: Joint Proceedings of the BIR 2023 Workshops and Doctoral Consortium co-located with 22nd International Conference on Perspectives in Business Informatics Research (BIR 2023). CEUR Workshop Proceedings. https://ceur-ws.org/Vol-3514/paper94.pdf

Zeiner, H., Unterberger, R., Tschuden, J., Quadri, M.Y. (2023). Time-Aware Optimisation Models for Hospital Logistics. In: Liu, S., Zaraté, P., Kamissoko, D., Linden, I., Papathanasiou, J. (eds) Decision Support Systems XIII. Decision Support Systems in An Uncertain World: The Contribution of Digital Twins. ICDSST 2023. Lecture Notes in Business Information Processing, vol 474. Springer, Cham. https://doi.org/10.1007/978-3-031-32534-2_4

Accepted:

Paletta, L., Schneeberger, M., Tschuden, J., Zeiner, H. (2024). Resilience Scores from Wearable Biosensors for Decision Support in Manufacturing. *Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences*, under review, Nice, France, July 2024.

Pszeida, M., Mosbacher. J.A., Schneeberger, M., Draxler, S., Weiss, W., Russegger, S., Albert, D., Paletta, L. (2024). Towards the Assessment of Resilience Parameters from Wearable Biosignal Sensors and Oculographic Features. *Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences*, Nice, France, July 2024.

Schneeberger, M., Carballo-Leyenda, B., Rodríguez-Marroyo, J., Paletta, L. (2024). Validation of Machine Learning-based Estimation of the Physiological Strain Index Using Gaussian Process Regression. *Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences*, Nice, France, July 2024.

Werz, J. M., Borowski, E., Isenhardt, I. (in press). Explainability as a Means for Transparency? Lay Users' Requirements towards Transparent AI. *Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences*, Nice, France, July 2024.

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. In: Complex Systems Informatics and Modeling Quarterly (CSIMQ) (under review)

6.2 Organisation of Scientific Events

International Conference on Applied Human Factors and Ergonomics 2024 (Nice, France):
Organisation of the **Session "Democratic Decision Support in Industrial Scenarios"**
(July 27, 2024, 13:30-14:30)

- Paletta, L., Schneeberger, M., Tschuden, J., Zeiner, H. Resilience Scores from Wearable Biosensors for Decision Support in Manufacturing. Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences, Nice, France, July 2024.
- Werz, J. M., Borowski, E., Isenhardt, I. Explainability as a Means for Transparency? Lay Users' Requirements towards Transparent AI. Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences, Nice, France, July 2024.
- Olbrych S., Nasuta A., Kemmerling M., Abdelrazeq A., Schmitt, R. H. From Simple to Sophisticated: Investigating the Spectrum of Decision Support Complexity with AI Integration in Manufacturing. Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences, Nice, France, July 2024.
- Gheibi, N., Böschen, S. Democratization in Industry via Multi-Agent Systems? The case of a production company. Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences, Nice, France, July 2024
- Muck C, Tschuden, J., Zeiner, H., Utz W. (2024). Explainability of Industrial Decision Support System using Digital Design Thinking with Scene2Model, Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences, Nice, France, July 2024.
- Schneeberger, M., Carballo-Leyenda, B., Rodríguez-Marroyo, J., Paletta, L. (2024). Validation of Machine Learning-based Estimation of the Physiological Strain Index Using Gaussian Process Regression. Proc. 15th International Conference on Applied Human Factors and Ergonomics (AHFE 2024) and the Affiliated Conferences, Nice, France, July 2024.

7 ANNEX A: LIST OF ABBREVIATIONS

Table 5: Abbreviations and long version explanation.

Abbreviation	Long version
AAA	Authentication, Authorization and Accounting
AI	Artificial Intelligence
CES	Cognitive-emotional Stress
CM	Conceptual Modelling
DAI-DSS	Democratic Artificial Intelligence-based Decision Support System
DL	Deep Learning
DSS	Decision Support Systems
GUI	Graphical User Interface
JSP	Job Scheduling Problem
MAS	Multi-Agent Systems
ML	Machine Learning
PSI	Physiological Strain Index
RL	Reinforcement Learning
XAI	Explainable AI