Uniwersytet Mikołaja Kopernika w Toruniu

Wydział Nauk Ekonomicznych i Zarządzania

Interdyscyplinarna Szkoła Doktorska Academia Copernicana

Stephan Wronkowski-Elster

Nr. albumu: 503049

Dyscyplina: ekonomia i finanse

Rozprawa doktorska

THE DESIGN AND DEVELOPMENT OF A REFERENCE ARCHITECTURE FOR TRUSTWORTHY AI

with a
Focus on Corporate Planning and
Decision-Making in the Process Industry

Promotorzy:
dr hab. Sylwester Bejger prof. UMK
dr Krzysztof Rykaczewski asst. prof. UMK

Toruń 2023 r.

May/June 2023

ABSTRACT

Various decisions relating to critical areas, such as medicine, finance, security, and defence, are being passed on to Artificial Intelligence (AI) algorithms with increasing regularity. Therefore, it is of immense importance to understand why an artificial intelligence model makes a particular decision or performs a specific action. This so-called *explainability* means that developers, users, and society remain able to comprehend -- as well as trust – certain judgements or results of activities; this, in turn, will allow for better decisions in cooperation between man and machine, or among autonomous machines. The research investigating these aspects is known as Explainable Artificial Intelligence – (short XAI[1]).

Strategy, corporate planning, and decision-making are core managerial functions. IT systems and approaches like Business Analytics, Predictive and Prescriptive Analytics have increasingly supported these tasks in the past twenty years. Artificial Intelligence is the next step in developing such systems to support managerial capacities.

However, if AI activities cannot be explained with respect to human communication, understandability, and readability, their users (especially the managerial and supervisory bodies) will not trust them (Chiusi et al., 2020; Been Kim et al., 2016; Bejger & Elster, 2019; Bejger & Elster, 2021; Christoph Molnar et al., 2019; Gilpin et al., 2019; de Graaf & Malle, 2017; Gunning, 2016; High-Level Expert Group on Artificial Intelligence, 2019; Lipton, 2017;Marco Tulio Ribeiro et al. 2016; Miller, 2018; Sameer Singh et al., 2019; Wachter et al., 2017; etc.) . AI's decisions and actions must be kept transparent, interpretable, and explainable to be considered trustworthy and reliable. The AI models (AI models, AI agents, and AI systems are terms used synonymously in this work) must also be held accountable for the decisions and actions they undertake (Doshi-Velez et al., 2017; Doshi-Velez & Kim, 2017, Wachter et al., 2017), as a lack of trust may lock substantial untapped potential for the increased growth of this new and promising technology, particularly in the field of integrated Corporate Planning and its objectives.

[1] Following Chari et al. (2020), the short term XAI was first coined by DARPA and therefore focus on a specific project. - Explainable AI is much broader, but this thesis is using both the term explainable AI and the abbreviations XAI or xAI.

The process industry (in short PI) plays a significant role in worldwide business, mainly in the pharmaceutical industry; it plays a leading role in protecting against (or even prevention of) deadly pandemics, for instance, as well as in the fight against cancer. Therefore, the PIs are central to transforming raw materials by way of applying suitable systems or methods into finished products. PIs can be classified with regards to their feedstock type or products used, for example, petroleum refining, mineral processing, chemical processing, fertilizers, food, and pharmaceuticals (Brennan, 2020). This work focused primarily on the chemical and pharmaceutical (in short, pharmacy) industry.

Corporate planning is the core process in the management cycle and deals with the prediction and achievement of future desired states that deviate from the current state. The importance of this process is particularly relevant for companies that produce with high fixed costs and thus must guarantee capacity utilisation in the future, as their production capacity cannot be flexibly adjusted. This is even more true for companies in the process industry due to their networking in highly complex supply chains up- and downstream (Elster, 2009). Therefore, optimised planning and decision-making, which uses, for example, modern tools such as artificial intelligence for strategic and tactical planning, is of immense importance for these companies. This pertains to chemicals and pharmacy, specifically scenario planning, integrated business planning, and decision-making.

The use of AI, especially subsymbolic black-box models, presents the above challenges.

Hypothesis:

By developing a reference architecture for an explainable AI system that could combine both subsymbolic and symbolic approaches, confidence in AI models and, thus, decision-making in corporate planning can be improved.

The primary goal of this dissertation is to establish and create a reference system architecture that promotes explainable artificial intelligence, with the aim of improving decision-making capabilities to facilitate better business planning within the process industry. The research has resulted in a reference system architecture for trustworthy AI in corporate planning, which is the main contribution of this work. To the author's knowledge, there are no previous or other comparable works in this domain.

This work examines a crucial research question: How can an explainable artificial intelligence system, or agent, be created and integrated into the planning framework of the process industry to increase trust in decision-making AI systems by improving their transparency and decision quality?

The research method is based on Design Science (Hevner et al., 2004; Wieringa, 2014). The reference architecture is supported by relevance and scientific rigour.
Hevner et al. (2004) define seven guidelines for understanding, executing and evaluating design-science research. The current research produced the reference system architecture for trustworthy AI in corporate planning in the process industry (Galster, 2011; Nakagawa, 2014). The issue addressed in this research holds significant importance for both theoretical and practical applications, with recent studies supporting its relevance. (Chiusi et al., 2020; Bejger & Elster, 2019; Bejger & Elster, 2020; Molnar, 2019; Singh et al., 2019; Tulio Ribeiro et al., 2016; Been Kim, et al, 2016; Gunning, 2016; High-Level Expert Group on Artificial Intelligence, 2019; Wachter et al., 2017; Chakraborti et al. 2020; Willms & Brandenburg, 2019; etc.) The evaluation of the design was conducted in two stages. Firstly, the design process was assessed based on the design principles outlined by Hevner et al. (2004). Secondly, a survey was conducted with experts in the fields of architecture and corporate planning to evaluate the design further. The survey questions were based on criteria derived from best practices in research and studies. (Bass et al., 2021; Vasconcelos et al., 2005). They were introduced using the method described by Saunders et al. (2023) and Sekaran and Bougie (2019). After conducting a survey, statistical methods were used to analyse the results. The group of experts provided additional concerns, requirements, and constraints identified as gaps in the thesis. These gaps will be considered during the next iteration of the design cycle. The survey results confirm the hypothesis that the developed reference architecture can serve as a viable solution to the stated problem.

The design science approach mentioned above was used to design and build a reference architecture called “Re_fish” (s combination of “**Re**jewski” and “Babel**fish**”, as a tribute to Marian Rejewski -- the leading Polish scientist who broke the Enigma code, and the

Babelfish – a fictional entity and universal decoder for any form of language in the universe[2]) that can be used in a (corporate) planning context within the process industry, by using a knowledge-based hybrid approach (Hitzler et al. 2021; Hitzler & Sarker, 2022; Hochreiter, 2022; Niu et al. 2022; Futia & Vetrò, 2020; Marcus, 2020; Marcus & Davis, 2021; Sohrabi et al., 2018; Tiddi, et al., 2020).

[2] "The Babelfish," said the Hitch Hiker's Guide to the Galaxy quietly, "is small, yellow and leech-like, and probably the oddest thing in the Universe. It feeds on brainwave energy […]" (Adams, 2010, p. 60)

ABSTRACT

Różne decyzje w krytycznych obszarach, takich jak medycyna, finanse, bezpieczeństwo i obrona narodowa, są coraz częściej przekazywane algorytmom sztucznej inteligencji (AI). Dlatego niezwykle ważne jest zrozumienie, dlaczego model sztucznej inteligencji podejmuje określoną decyzję lub wykonuje określone działanie. Tak zwana wyjaśnialność oznacza, że programiści, użytkownicy i społeczeństwo są w stanie zrozumieć i zaufać pewnym osądom lub wynikom działań, co z kolei umożliwia podejmowanie lepszych decyzji we współpracy człowiek-maszyna lub między autonomicznymi maszynami. Badania zajmujące się tymi aspektami nazywane są Explainable Artificial Intelligence (w skrócie XAI).

Strategia, planowanie biznesowe i podejmowanie decyzji to podstawowe funkcje zarządzania. Systemy i podejścia IT, takie jak Business Analytics, Predictive i Prescriptive Analytics, w coraz większym stopniu wspierały te zadania w ciągu ostatnich dwudziestu lat. Sztuczna inteligencja jest kolejnym krokiem w rozwoju takich systemów wspierających możliwości zarządzania. Jeśli jednak działań AI nie można wyjaśnić w kategoriach ludzkiej komunikacji, zrozumiałości i czytelności, ich użytkownicy (zwłaszcza organy zarządzające) nie będą im ufać (Chiusi i in., 2020; Been Kim i in. 2016; Bejger & Elster, 2019; Bejger & Elster, 2021; Christoph Molnar i in, 2016; Bejger & Elster, 2019; Bejger & Elster, 2021; Christoph Molnar et al, 2019; Gilpin et al, 2019; de Graaf & Malle, 2017; Gunning, 2016; High-Level Expert Group on Artificial Intelligence, 2019; Lipton, 2017; Marco Tulio Ribeiro et al, 2016; Miller, 2018; Sameer Singh et al, 2019; Wachter et al, 2017; etc.). Decyzje i działania sztucznej inteligencji muszą być przejrzyste, możliwe do zinterpretowania i wyjaśnienia, tak aby można je było uznać za godne zaufania i wiarygodne. Modele AI (modele AI, agenci AI i systemy AI są używane w niniejszym dokumencie jako synonimy) muszą również ponosić odpowiedzialność za podejmowane przez siebie decyzje i działania (Doshi-Velez i in., 2017; Doshi-Velez & Kim, 2017, Wachter i in., 2017), ponieważ brak zaufania może zablokować ogromny i niewykorzystany potencjał wzrostu tej nowej i obiecującej technologii, zwłaszcza w dziedzinie biznesu, a zwłaszcza w planowaniu w przedsiębiorstwie.

Przemysł przetwórczy (w skrócie PI) odgrywa znaczącą rolę w globalnej gospodarce, zwłaszcza w przemyśle farmaceutycznym. Przemysł farmaceutyczny odgrywa wiodącą rolę np. w ochronie przed (lub nawet zapobieganiu) śmiertelnym pandemiom i np. w walce z rakiem. PI mają zatem kluczowe znaczenie dla przekształcania surowców w gotowe produkty poprzez zastosowanie odpowiednich systemów lub metod. PI można sklasyfikować zgodnie z charakterem wykorzystywanych surowców lub produktów, np. rafinacja ropy naftowej, przetwarzanie minerałów, przetwarzanie chemiczne, nawozy, żywność i farmaceutyki (Brennan, 2020). W niniejszej pracy skupiono się przede wszystkim na przemyśle chemicznym i farmaceutycznym (w skrócie farmaceutycznym).

Planowanie w przedsiębiorstwie jest podstawowym procesem w cyklu zarządzania, którego zadaniem jest stawianie prognoz i osiąganie przyszłych stanów docelowych, które odbiegają od stanu rzeczywistego. Znaczenie tego procesu jest szczególnie istotne dla firm, które produkują z wysokimi kosztami stałymi, a tym samym muszą zagwarantować wykorzystanie mocy produkcyjnych w przyszłości, ponieważ ich zdolności produkcyjne nie mogą być elastycznie dostosowywane. Dotyczy to szczególnie firm z branży przetwórczej ze względu na ich wzajemne powiązania w wysoce złożonych łańcuchach dostaw wyższego i niższego szczebla (Elster, 2009). Zoptymalizowane planowanie i podejmowanie decyzji, z wykorzystaniem np. nowoczesnych narzędzi, takich jak sztuczna inteligencja do planowania strategicznego i taktycznego, ma zatem ogromne znaczenie dla tych firm. Dotyczy to sektora chemicznego i farmaceutycznego, w szczególności planowania scenariuszy, zintegrowanego planowania i podejmowania decyzji.

Dzięki wykorzystaniu sztucznej inteligencji, a zwłaszcza podsymbolicznych modeli czarnej skrzynki, pojawiają się powyższe wyzwania.

Hipoteza:

Opracowując architekturę referencyjną dla wyjaśnialnego systemu sztucznej inteligencji, który może łączyć zarówno podejście subsymboliczne, jak i symboliczne, można zwiększyć zaufanie do modeli sztucznej inteligencji, a tym samym podejmowanie decyzji w planowaniu biznesowym.

Głównym celem niniejszej rozprawy jest opracowanie architektury systemu referencyjnego, która promuje wyjaśnialną sztuczną inteligencję w celu poprawy zdolności decyzyjnych i umożliwienia lepszego planowania przedsiębiorstw w przemyśle przetwórczym. Wynikiem

badań jest architektura systemu referencyjnego dla godnej zaufania sztucznej inteligencji w planowaniu biznesowym, która stanowi główną część niniejszej rozprawy. Zgodnie z wiedzą autora, nie ma wcześniejszych lub porównywalnych prac w tej dziedzinie.

W niniejszej rozprawie kluczowym pytaniem badawczym jest: W jaki sposób można stworzyć wytłumaczalny system sztucznej inteligencji lub agenta i zintegrować go z ramami planowania przemysłu procesowego, aby zwiększyć zaufanie do decyzyjnych systemów AI poprzez poprawę ich przejrzystości i jakości decyzji?

Metoda badawcza opiera się na Design Science (Hevner i in., 2004; Wieringa, 2014). Architektura referencyjna jest wspierana przez relewancję i rygor naukowy.
Hevner et al. (2004) definiują siedem wytycznych dotyczących rozumienia, prowadzenia i oceny badań z zakresu design science. W aktualnych badaniach opracowano architekturę systemu referencyjnego dla godnej zaufania sztucznej inteligencji w planowaniu biznesowym w przemyśle przetwórczym (Galster, 2011; Nakaga-wa, 2014). Temat poruszany w tych badaniach ma ogromne znaczenie zarówno dla zastosowań teoretycznych, jak i praktycznych, o czym świadczą ostatnie badania. (Chiusi i in., 2020; Bejger & Elster, 2019; Bejger & Elster, 2020; Molnar, 2019; Singh i in., 2019; Tulio Ribeiro i in., 2016; Been Kim i in., 2016; Gunning, 2016; High-Level Expert Group on Artificial Intelligence, 2019; Wachter i in., 2017; Chakraborti i in., 2020; Willms & Brandenburg, 2019; itp.) Ocena projektu została przeprowadzona w dwóch etapach. Po pierwsze, proces projektowania został oceniony w oparciu o zasady projektowania nakreślone przez Hevnera i in. (2004). Po drugie, przeprowadzono ankietę z ekspertami w dziedzinie architektury i planowania biznesowego w celu dalszej oceny projektu. Pytania ankietowe opierały się na kryteriach zaczerpniętych z najlepszych praktyk w badaniach i analizach. (Bass i in., 2021; Vasconcelos i in., 2005). Zostały one wprowadzone zgodnie z metodą opisaną przez Saunders i in. (2023) oraz Sekaran i Bougie (2019). Po przeprowadzeniu ankiety wyniki przeanalizowano przy użyciu metod statystycznych. Grupa ekspertów przedstawiła dodatkowe obawy, wymagania i ograniczenia, które zostały zidentyfikowane jako luki w tej pracy. Luki te zostaną uwzględnione w kolejnej iteracji cyklu projektowania. Wyniki ankiety potwierdzają hipotezę, że opracowana architektura referencyjna może zapewnić realne rozwiązanie określonego problemu.

Powyższe Design-Science zostało wykorzystane do zaprojektowania i stworzenia architektury referencyjnej o nazwie "Re_fish" (połączenie słów "**Re**jewski" i "Babel**fish**", jako hołd dla Mariana Rejewskiego - czołowego polskiego naukowca, który złamał kod Enigmy, i Babelfish - fikcyjnego podmiotu i uniwersalnego dekodera dla każdej formy języka we wszechświecie), który może być wykorzystywany w kontekście planowania (korporacyjnego) w przemyśle przetwórczym, przy użyciu hybrydowego podejścia opartego na wiedzy (Hitzler i in. 2021; Hitzler & Sarker, 2022; Hochreiter, 2022; Niu et al. 2022; Futia & Vetrò, 2020; Marcus, 2020; Marcus & Davis, 2021; Sohrabi et al., 2018; Tiddi, et al., 2020).

BIOGRAPHICAL SKETCH

Stephan Georg Elster was born on January 14th, 1970, in Gummersbach, Germany. After completing his secondary education at the Hollenberg Gymnasium Waldbröl and his military service as an officer cadet (his current rank is captain), he attended University of Siegen in 1991 and graduated in 1999 with a degree as a Diplom-Kaufmann (M.Sc. in Management). He then started his professional career in parallel. From 2002 until 2006 he studied business informatics with focus on software engineering, management information systems and decision support systems. He graduated with a Diplom- Wirtschaftsinformatiker (B.Sc.) degree in 2006. In 2007 he attended the Westfälische Wilhelms University of Münster and graduated in 2009 with a Master of Business Administration degree. Always keeping in touch with research, he joined Prof. Reinhard Selten (Noble Prize in economics) in 2014 as an academic assistant at the Nobel Laureate Meeting conference at Lake Constance. In 2018 he attended the interdisciplinary doctorate academy, the Academia Copernicana, as a PhD student in the faculty of Economics Sciences and Management – applied Informatics and Mathematics department.

Parallel to his education and research, Stephan has made an outstanding professional career. He has worked for companies such as COGNOS, IBM Germany and IBM Switzerland, in 2016 was rated "high potential" being the Head of Sales Consulting at T-Systems (Deutsche Telekom). In 2017, he joined SAP Germany in the industry division Process, Life Sciences and Consumer Goods as a business development lead and principal architect. He was appointed to chief architect at the beginning of 2023.

DEDICATION

One evening, June 2014, at Lake Constance, as we were working together on a presentation paper for a conference, I asked Reinhard Selten how he had managed to develop such an extraordinary mathematical understanding and memory. He promptly replied: "I had a long walk home after school and was able to think a lot."

In memory of Reinhard Selten

We are all on our way home somehow - and we should use the time on the way wisely.

ACKNOWLEDGMENTS

Writing a doctoral thesis is like a rollercoaster ride. The joy when it goes uphill and you gain insights, and the downhill when you realise that there is still a lot of work ahead of you. Writing a doctoral thesis alongside a full-time job is like riding a hyper rollercoaster. Despite everything, you have a smile on your face when you reach the finish line, no matter what the ride was like.

I would like to thank everyone who has supported me along the way.

First and foremost, I would like to thank my supervisor Prof. Dr. Sylwester Bejger, who always kept his cool throughout the entire journey. I owe a lot to him and his approach to the topics of artificial intelligence. I would also like to thank my second supervisor, Prof. Dr. Krzysztof Rykaczewski, who always responded super-fast when I had a question or needed help.

My special thanks go to everyone who helped me - Apple Music, the experts, Ariel, Yakoota, Aava & Pukki, who haven't seen me much in the last four years, my ~~manager~~ mother.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

ABBREVIATION	MEANING
A	
ADD	Attribute Driven Design
ADM	Architecture Development Method
AI	Artificial Intelligence
AIS	Autonomous Intelligent Systems
AISOP	AI-based Scenario planning to Predict crisis situations
API	Application Programming Interface
APO	Advanced Planning and Optimization
APS	Advanced Planning and Scheduling
APQC framework	American Productivity & Quality Center
ASR	Architecturally Significant Requirements
ATP(aATP)	Available to Promise and Advanced Available to Promise
B	
BBIB	Building Blocks Information Base
BI	Business Intelligence
C	
CALO	Cognitive Assistant that Learns and Organizes
CBIS	Computer-Based Information System
CEFIC	European Chemical Industry Council
COGS/COS	Cost Of Goods Sold/ Cost Of Sales
COMPAS	Correctional Offender Management Profiling for Alternative Sanctions
CRISP-DM method	CRoss Industry Standard Process for Data Mining
CSR	Corporate Social Responsibility
D	
DARPA	Defense Advanced Research Projects Agency
DDD	Domain Driven Design
DFKI	Deutsches Forschungszentrum für Künstliche Intelligenz
DSS	Decision Support Systems
DLG	Distributed General Ledger
DNNs	Deep Neural Networks
E	
EES	Explainable Expert System Framework
EIS/ESS	Executive Information Systems – Support Systems
ELI5	“Explain like I am 5”
ERP	Enterprise Resource Planning
ETL	Extract Transform and Load
EU27	European Union 27 members
F	
FDA	Food and Drug Administration
FRAM	Functional Resonance Accident Model
G	
GDPR	General Data Protection Regulation
GHG	Greenhouse Gas

GMP	Good Manufacturing Practices
I	
ICEE	Integrated Cognitive Explanation Environment
IBP	Integrated Business Planning
INFORMS	Institute for Operations Research and Management Science
IW	Inference Web
J	
JSON-LD	JavaScript Object Notation for Linked Data
K	
KB Systems	Knowledge Based Systems
KG	Knowledge Graphs
KPI	Key Performance Indicator
KRR	Knowledge Representation and Reasoning
L	
LIME	Local Interpretable Model-Agnostic Explanations
LRP	Layer Relevance Propagation
M	
MAPE	Mean Absolute Percentage Error
MC	Management Cycle
ML	Machine Learning
MRP	Manufacturing Resource Planning
N	
NAFTA	North American Free Trade Agreement
NLP	Natural Language Processing
O	
OWL	Web Ontology Language
P	
PML	Proof Markup Language
Q	
Q4	here Q = quarter
R	
RA	Reference Architecture
R&I	Research & Innovation
R&D costs	Research & Development costs
RAG	Resilience Analysis Grid
PAL	Personal Assistant that Learns
RDF	Resource Description Framework
RDFS	RDF Schema
REACH	Registration, Evaluation, Authorisation and Restriction of Chemicals
ROI	Return on Investment
RRI	Responsible Research and Innovation
S	
SARIMA	Seasonal Auto-Regressive Integrated Moving Average
S&OP	Sales and Operations
SCHOLAR	Expert system
SG& A expenses	Selling, General and Administrative Expenses
SHAP	SHapley Additive exPlanations

SIB	Standards Information Base
SKU level	Stock Keeping Unit
SPA	Scenario Planning Adviser
STRIPS	Stanford Research Institute Problem Solver
SWRL	Semantic Web Rule Language
SVA	Shareholder Value

T

TOGAF	The Open Group Framework

W

WWW	World Wide Web

X

XAI	Explainable Artificial Intelligence

"The following presentation is also intended to take into account the latest results and, by placing the emphasis not on the theories but on the facts, both to try to give a picture of what is currently known of Explainable AI, as well as to indicate the directions in which the current impetus of research seems to run" (following Julius Elster and Hans Friedrich Geitel (1897),Elster & Geitel (1897), Compilation of the results of new work-by the inventors of the photocell (1893)-underlined changes by the author)

"When we were new, Rosa and I were mid-store, on the magazines table side, and could see through more than half of the window. So, we were able to watch the outside – the office workers hurrying by, the taxis, the runners, the tourists, Beggar Man and his dog, the lower part of the RPO Building." (Ishiguro, Kazuo (2021). Klara and the Sun. Chapter 1)

1 Introduction

Facts and circumstances that are valid today may undergo rapid change and become invalid virtually overnight; human decision-makers, who by virtue of their jobs steer the operations of a system (e.g., a company) in a target-oriented manner, are confronted with this potential uncertainty on a daily basis. Often, their training only allows for decision-making by going on a so-called "good gut feeling". In the decision-making phase, there may be a need for information, or a need may arise to evaluate it in a target-oriented manner, or a training gap may appear while using available information in a target-oriented way. Planning is one of the primary tasks embedded in the management cycle. The starting point of planning is the existence of states in a system which are considered unsatisfactory in comparison to other states, or unacceptable (no longer acceptable) concerning external specifications or requirements, according to an affected party (the decision-maker) (Macharzina & Wolf, 2023; Hammer, 2015; Kaplan & Norton, 2014; Klein & Scholl, 2011). In this dissertation, the reference architecture of an AI system for explainable AI[3] within corporate planning context is examined and developed – such a system should help decision makers make broader and better use of AI within planning contexts.

Since AI plays a significant role in this thesis in the context of improving decision-making in corporate planning in the process industry, AI is considered from both a macroeconomic and microeconomic perspective in chapter 1.1 and chapter 1.2. The results of these two preliminary chapters are referenced again and again in the following chapters, before the

[3] Explainable AI, xAI or XAI are used synonymously

motivation and relevance (in the sense of design science research, see chapter 1.5) are considered in chapter 1.3. Chapter 1.4 describes the Research Goal and the Research Question. Chapter 1.5 describes the research theory of the design science approach and the research design of this thesis. Chapter 1.6 explains the structure of the thesis and the different ways of reading the thesis.

1.1 The Macroeconomic Perspective of AI

The importance of AI is growing in our personal and professional lives, having a significant impact on our world. Various countries, such as China, the USA, and Europe, compete fiercely to establish themselves as leaders in AI technology (here and in the following Szczepanski, 2019). Finally, the significant investment programs and "AI strategies" implemented by each country demonstrate their commitment to establishing a strong foundation for their economies in regard to AI.

According to OECD reports (OECD, 2017), the field of AI has seen significant growth in patent applications. The "AI patents worldwide, 2000- 2015" report highlights that Japan, Korea, and the USA account for two-thirds of AI patents, while Europe only contributes 12%. Additionally, China and Taiwan have seen a significant increase in patent applications, which indicates substantial investments in AI in these countries. It is worth noting that, according to the World Intellectual Property Report (Abbott, 2019), the growth of patents related to AI in Japan, the USA, and Europe has decreased compared to the period from 2000 to 2005. However, the number of scientific articles on AI has increased. Interestingly, machine learning applications had the highest number of patent applications during the period covered by the report. Despite Europe's reputation for having a strong foundation in AI, particularly in developing governance rules and standards, companies in Europe are behind the USA and China in terms of patent applications.

The potential impact of AI on modern societies is significant and far-reaching. Nowadays, an increasing number of people are utilizing AI assistants in their daily lives, such as when using their iPhones (Siri was built by Apple on using parts from DARPA PAL project, s. chapter 5.2.3), or when writing texts, shopping online or finding prices, etc. AI must be seen from several perspectives. Looking at AI through an economic lens, it is considered a significant contributor to growth. However, when viewed from a social and cultural standpoint, it is met with scepticism and seen as a potential threat, despite its benefits that

enhance society. To better understand AI's current and future impact on the economy, summaries of various studies conducted by four reputable consultancies have been reviewed.
According to a report titled "Why Artificial Intelligence is the Future of Growth" by Accenture (Purdy & Daugherty, 2016), the significance of capital investment and labour as a means of economic growth can no longer be relied upon. Instead, the focus should be on new transformative technologies. Accenture discusses the emergence of an AI era, emphasizing the ability of AI to surpass the limitations of capital and labour and tap into novel sources of value and expansion. Based on their analysis of twelve countries, the potential exists for these nations to increase their yearly economic growth rates twofold by 2035. Further, Accenture suggests that policymakers and business leaders should recognise AI as a transformative technology that offers diverse growth opportunities rather than solely improving productivity, as commonly assumed.
Economists, including Robert Gordon (Gordon, 2016), predict a continued decline in growth rates. Gordon (2016) refers to his thesis of “secular stagnation” and that the productivity-enhancing potential of new technologies is overestimated. (A similar discussion already took place with earlier waves of innovation, for example Solow (1987) did not recognise any positive effect from the increased use of computers (Menzel & Winkler, 2018).According to their analysis, this will be the new normal for the next 25 years since the factors of production, such as labour and capital, have plateaued, and there are no foreseeable innovations on the horizon that can replicate the likes of steam propulsion and telegraphy. This lack of investment paired with unfavourable demographic trends, such as growing educational inequality, will only add to stagnation or even decline. Also, the authors argue that economists have overlooked the transformative potential of AI. While it is commonly recognized that AI can boost production, it is often overlooked that AI is a hybrid factor of production that combines capital and labour. The speed at which AI can perform work activities is unparalleled, surpassing human capabilities by a thousandfold. Interestingly enough, there are some ways to turn AI into capital, such as using it in robotics or intelligent machines. Moreover, AI has enhanced its functionality over time by learning from work processes. As processes. As stated by Purdy and Daugherty (2016), there are three potential opportunities for growth via AI, mainly through intelligent automation. With the help of AI, intricate and strenuous physical tasks can now be replaced. Additionally, virtual work can also be carried out through software agents, which are capable of replacing non-physical tasks such as matching outgoing invoices with payments, within the framework of robotic process automation. There is also a potential for advancement by

building upon existing work, as outlined in this dissertation, which could ultimately exceed human capabilities. Another opportunity for growth arises when innovations spread from one area to another, resulting in increased efficiency through the use of AI and leveraging synergies. This will lead to a significant increase in productivity, projected to reach up to 40 percent by 2035.

According to a study conducted by PricewaterhouseCoopers (Gillham, 2018) in 2018, the global GDP is projected to increase by approximately 14% by 2030. This translates to a staggering amount of around 16 trillion US dollars worldwide (Gillham et al., 2018). These remarkable gains are attributed to the widespread adoption of AI technologies and services, with the USA and China leading the charge. By offering these services, a significant and precise amount of data will be generated, leading to a positive cycle where AI products and services can utilize this data to improve their offerings. This will ultimately result in even better products and services being developed. Europe will experience some benefit from this economic upswing, although not to the same extent as the US and China. It is believed that the adoption of new technologies and services in China will take a longer time compared to the US, mainly due to its size and infrastructure. Moreover, developing nations are expected to have a lower involvement in this growing trend.

In a study titled "Notes from the Frontier: Modelling the Impact of AI on the World Economy" (McKinsey, 2018), McKinsey (in the following in short McK) analysed the influence of AI on the global economy. The study reveals that by 2030, around 70% of companies are likely to adopt AI technology to some extent, but full adoption is not expected. In contrast, it is assumed that less than 50% of companies will fully embrace all available AI technologies by 2030. Surprisingly, around 30% of companies are not planning to adopt any AI technology at all. This raises the question of why and whether this lack of adoption is due to a lack of trust in the capabilities and decision-making abilities of the new technology. According to McK's projections, the global GDP is predicted to grow at a rate of 1.2% annually until the year 2030. This growth is primarily attributed to the rise of automation and the replacement of traditional labour with AI, as well as the innovation of AI in various products and services. The negative impacts of AI, which will be discussed later, can cause a disruption in the job market resulting in higher costs for transitioning workers and a decrease in local spending because of increased unemployment.

As reported by Petropoulos et al. (2019), the global economy is currently experiencing a significant shift towards digitalization powered by AI. Despite the potential advantages of

adopting these new technologies, such as improved production processes, productivity levels do not seem to have increased significantly. This is true for both the US and the EU28, including the UK, countries. For example, in the period from 2005 to 2016, the average growth in the production rate was only 1.3%, compared to an average increase in the production rate of 2.8% in the years from 1995 to 2004. While the financial crisis hit the EU harder than the US, both regions experienced similar stagnation in production growth after recovering. This can be seen by examining the total factor productivity (TFP) growth rate, which has steadily declined for the countries analysed (USA, UK, EU, and Japan) and is currently at a low level. (s. figure 1) (Bergeaud, et al., 2018)

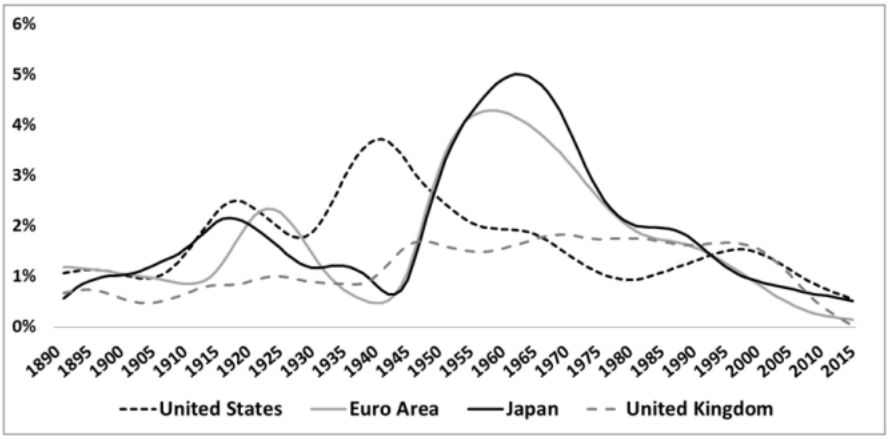

Figure 1: Total factor productivity (Bergeaud et al., 2018)

The phenomenon of decreased productivity despite the introduction of technology is known as the modern productivity paradox. This was previously observed during the early 1980s with the advent of computers and information technology. According to Brynjolfsson et al. (2017), there are four possible explanations for this paradox:

1. Explained by the failure to meet AI expectations

Here, Brynjolfsson et al. (2017) presents a rather pessimistic view of Gordon (2016). In his opinion, AI cannot contribute significantly to the growth of productivity increases, not least because of its limited added value, e.g., compared to electricity or the combustion engine. However, according to Petropoulos (2019), all inventions and technology improvements take time to act productivity. Technological advancements, such as the steam engine, electricity, the internal combustion engine, and the computer, have demonstrated this

phenomenon. And, undoubtedly, these technological advancements and inventions improved productivity directly and spurred innovations in other areas, as mentioned above. This led to a positive synergistic effect, complementing each other's impact Brynjolfsson et al. (2017).

2. Explained by errors in data collection

One of the advocates of this explanation is Hal Varian (Varian, 2018), a former economics professor who currently works at Google. Varian (2018) gives an example that in the past, photos were taken from films produced by only three companies worldwide, and due to digitalisation, this market has nearly vanished altogether. Photos are now exchanged among acquaintances and not sold, whereas the GDP is designed for market transactions. Brynjolfsson and Collis (2019) use a similar example with the messenger Whatsapp. On the contrary, Ahmad et al. (2017) argue that some products and services are freely available but cannot plausibly explain the change on this scale.

3: Explained by an unfair distribution of AI benefits

It seems that only the biggest companies in the industry can reap the rewards of increased productivity that come with incorporating AI technology. As a result, they are able to increase the gap between themselves and smaller companies, ultimately leading to a decrease in the motivation for smaller companies to implement AI technology. The McK study mentioned above also agrees that the pioneer companies are more likely to be involved in the adoption and use of AI and thus have advantages.
(McK, 2019; Andrews et al, 2016; Gutierrez & Philippon, 2017; De Loecker & Eeckhout, 2017; Autor et al. 2020; Furman & Orszag, 2015)). So, Petropulous (2019) suggests stricter regulation through a well-designed policy framework. This can promote a fairer distribution of productivity gains and minimise the potential risks of AI usage.

4: Explanation through barriers to implementation

As previously stated, Brynjolfsson et al. (2017) believes that adopting and widespread using AI and other technologies takes time, often requiring innovation in different fields, before they can significantly contribute to productivity growth. This has already been seen in the two periods from 1870 to 1940, for electricity, and from 1970, for information technology. These two periods thus show similar patterns to those of the present day. Produc-

tivity growth decreased during the twenty-five years that followed technology implementation in the previous two periods, but later experienced a significant acceleration. Taking the above into consideration, it is assumed that AI will not only achieve a simple increase in productivity, but this technology also has the potential to revolutionise the entire economy. The emergence of AI as a new factor of production has led to a hybrid combination of labour and capital. This has resulted in a decrease in the significance of both labour and capital in global economic growth. The incorporation of AI has been made possible through three channels. Firstly, automation has allowed for the automation of areas in the physical world. Secondly, virtual AI agents have been integrated across previous industrial boundaries. Lastly, AI has enabled the creation of new industries altogether.

Below, the role of the so-called singularity and the implementation of R&D by machines will be shown. There is potential for additional growth through the augmentation of labour and capital, which can be achieved by utilising the collaboration between humans and AI in various fields. Here, human capabilities are significantly enhanced by AI. In addition, many scientists see a further growth opportunity through the diffusion of AI into other areas, which thus influence each other positively through AI.

The Polish economist, Jakub Growiec, (Growiec, 2022), has created a hardware-software called framework to explore the topics mentioned above and determine potential combinations for growth. He sees two scenarios as the most likely and examines them in terms of their effect on increasing the productivity rate. These two scenarios pivotally revolve around the so-called Technological Singularity, i.e., the point in time when, as predicted by Kurzweil (2005) and other scientists and visionaries (Kurzweil, 2005; Davidson, 2021; Hanson & Yudkowsky, 2013; Roodman, 2020; Sandberg ,2013), AI surpasses human intelligence. It is important to note that AI has the ability to replace not only simple and low-skilled work but also highly demanding cognitive work, such as Research & Development (R&D). As stated by Growiec (2022), the singularity will bring about explosive growth in this field.

The three growth channels mentioned above are relevant to this work, presented here at the macroeconomic level. Specifically, the first two - automation, and labour and capital augmentation. Furthermore, the results of the McK study indicate that companies leading the way in AI implementation stand to gain the most, potentially leaving slower adopters behind.

The macroeconomic effects described above, which will result from the transformation of the economy through AI, will have a profound impact on labour and capital. Thus, the

impact of AI on individuals' work is inevitable, but the specifics are still unclear and there are only a few empirical studies on the topic to date. The criteria mentioned above also apply to the labour market, where AI impacts it through three channels. These channels replace human labour with AI, particularly in physical work or repetitive activities like text or image recognition. Secondly, AI can enhance productivity and boost human capabilities by augmenting their skills. The third channel involves the implementation of AI in new business areas, creating fresh tasks that impact both companies and employees. This channel leads to innovation and is closely linked to the diffusion of AI into various other fields. In the following, three renowned studies and their approaches, which may differ considerably, will be briefly presented. In one of the best-known studies on the impact of AI on the labour market, the study conducted by Frey and Osborne (Frey & Osborne, 2017), the researchers conclude that 47% of today's occupations could be replaced by AI in the medium term. Frey and Osborne use the Standard Occupational Classification (SOC for short) used in the USA to classify over seven hundred occupations as "replaceable" or "not replaceable" on the basis of 70 subjectively selected occupations based on interviews with experts (s. Graus et al., 2021). An occupation is considered replaceable if all the activities of the occupation are considered replaceable in the case of a non-substitutable activity, the entire occupation is also considered non-substitutable. In the further statistical treatment for calculating the replacement probability, the occupations and their activities, whether "replaceable" or not, are fully included in the statistics as non-substitutable occupations if only one activity is not replaceable. The replacement probabilities obtained in this way are then regressed on occupation-describing variables, which then calculate the replacement probabilities of the other remaining occupations. The resulting statement of 47% of the occupations is replaceable since they also include occupations whose activities include, for example, critical thinking, problem-solving, etc., which are not currently considered replaceable by AI. The study by Frey and Osborne (2017) can, therefore, only be seen as a starting point for a more in-depth investigation. Other approaches, such as that of Autor et al. (2003), seem more plausible in that they first assign replacement probabilities to the individual activities, add these up and weigh them according to the proportion in the specific occupation (s. Graus et al., 2021). Manyika et al. (2017) use this approach in that the authors define the automatability of core skills and then determine the substitutability of activity within the above-mentioned work database, also taking into account the proportion of the activity in the occupation. They come to the less spectacular conclusion that less than 5% of occupations are fully automatable, but 60% of occupations contain at least 30%

automatable activities. By using the approach of Autor et al. (2003), Bonin et al. (2015) conclude that only about nine to twelve percent of jobs are threatened by the replacement of AI (s. Graus et al., 2021). Studies that look at the gross effects, i.e. both the job losses and the job gains due to the new technology, show a more positive picture - McK (2017), for example, concludes that the use of AI will create around 10 million new jobs by 2030, and Acemoglu and Restrepo (2018) also come to the conclusion that the losses can be offset by compensatory mechanisms in which new jobs are created (Menzel & Winkler, 2018; Graus et al., 2021).

All these studies are, therefore, to be regarded as more or less well-grounded estimates. Autor et al. (2003) and Autor and Acemoglu (2011) assume that the adoption of automation and robotics will likely result in the replacement of tasks that involve repetitive and routine skills, as previously described. Whereas AI typically offers complementary support to tasks that require non-routine activities, serving as a helpful augmentation. For this reason, it is probable that AI will have a greater impact on workers with medium-level skills who perform routine tasks that can be automated, rather than those with high or low levels of skill. This, in turn, would lead to a polarisation of the labour market. (Goos & Manning, 2007; Autor & Acemoglu, 2011). Brynjoffsson et al. (2018) also conclude that AI, due to its intrinsic learning capacity, can create new opportunities for human-machine collaboration (e.g., in medical diagnosis or prognosis, etc.), while human labour will tend to be replaced in purely "codifiable" activities (s. Graus et al., 2021).

As for individuals and companies, some empirical studies already show results that suggest that the use of AI in companies will change the skill profile requirements of employees. For example, an OECD study (Samek, Squicciarini, & Cammeraat, 2021) found that it will be more important for "human" workers to acquire interpersonal and social skills that complement technical skills where appropriate. This is not only about understanding AI, but also being able to ensure that it is used according to e.g., compliance requirements (Samek, Squicciarini, & Cammeraat, 2021).

The impact of AI and three growth channels ultimately affect the labour market is unclear. However, there are opportunities and risks. Brynjolfsson et al. (2014) see the chance that AI can help humans improve their skills and thus help humanity with current problems and, for example, eradicate poverty, cure diseases better, provide better education for all or understand and control a possible negative change in the climate. In contrast, there are current statements against the background of the introduction of ChatGPT, such as those of Jeffrey Hinton (Hanna, 2023), the Turing Award winner, who, like many others, warns

against the consequences of AI. Hinton speaks of an "existential threat" and uses a term coined by the philosopher Nick Bostrom (2003, 2005, 2017). This and the ideas which Hinton reveals in his post-exit interview seem to reflect the idea of "effective altruism", which tries to combine neoliberal economics and ethics. The underlying concept here is that the world is full of suffering, and therefore, the limited resource of aid should be utilised in the most possible and effective way. "Earn to Give" and so-called "Longtermism"[4], in which future human lives are equated with current human lives, are further ideas. The movement also includes what is known as transhumanism.

Finding 1: Impact of AI on economy

The impact of AI on economy can be described as the importance of former production factors, labor and capital, become less important, or grow together into a single factor.

Finding 2: Potential growth opportunities through AI:

1. Intelligent automation. With the help of AI, intricate and strenuous physical tasks can now be replaced. – Replacement case.
2. Additionally, virtual work can also be carried out through software agents, which can replace non-physical tasks such as matching outgoing invoices with payments, within the framework of robotic process automation (RPA[5]). – Replacement case.
3. There is also a potential for advancement by building upon existing work, as outlined in this dissertation, which could ultimately exceed human capabilities. - Augmentation case.
4. Another opportunity for growth arises when innovations spread from one area to another, resulting in increased efficiency through the use of AI and leveraging synergies – Raising synergies through diffusion.

Finding 3: Impact on Labour:

The impact of AI on the labour market is not viewed uniformly. There are different opinions about the strength and direction of the impact. However, the impact can be differentiated according to the growth drivers outlined in Finding 2. For example, some work will

[4] https://de.wikipedia.org/wiki/Longtermism , accessed 18.06.2023

[5] Robotic Process Automation, s. e.g., https://www.sap.com/germany/products/technology-platform/process-automation/what-is-rpa.html , accessed on 18.06.2023

fall under the so-called replacement case, others under the augmentation case, and new work will be created, for example, through the diffusion of innovation into other areas.

1.2 The Microeconomic Perspective of AI

In the years to come, one of the crucial tasks for companies is to embrace digitalisation in their industries. This involves the so-called Industry 4.0, i.e., the fusion of various technologies such as Big Data Analytics, IoT etc. and certainly AI into a cyber-physical system. The data obtained then is used by the machine learning systems to enhance their learning process and optimise themselves.

In a 2018 conducted study McK (2018) divided companies into three groups: the "leaders" or "frontrunners", the "followers" and the "laggards". One of the findings was that the leaders will reap the greatest benefits from the adoption of AI - the "front runners", 10% of the companies, turned out to be the companies that will use AI technology extensively and across the enterprise in the next five to seven years (Rogers, 1983). This can lead to a so-called "the winner takes it all" effect so that so-called “superstar” companies establish themselves in the sectors. The laggards can only catch up with the “superstar” companies if there is a delay in the diffusion of technology (Autor et al., 2017). The so-called followers, 20-30% of the companies, on the other hand, are adopting AI technology, but with caution so that the changes in cash flow are slower and less noticeable. In the McK (2018) study, 60-70% of the companies are the so-called Laggards. These can lose up to 23% of today's cash flow (based on the simulation used in the study) and also lose out in the AI race. The response of these companies will be to minimise costs and reduce investments, which in turn may further widen the gap with the laggards. These companies ultimately risk being forced out of the market. Still, AI is expected to affect global competition and individual companies in various ways, such as through regulatory alignment within the framework of trade agreements, as well as at the global level, for example with regard to pricing algorithms, which may pose a potential risk of promoting collusion (Monopolkommission, 2018).

Furthermore, Cockburn et al. (2018) provide an interesting and important result. They predict that AI will have a significant impact on markets by revolutionising the way innovation is approached. According to Cockburn et al. (2018), AI is expected to bring about advancements in research and development that rely more on data-driven algorithms and less on traditional human research methods. As a result, data is becoming increasingly important

for companies. Merck KGaA, a process industry company, serves as a good example. They have entered into a strategic partnership with the American company Palantir Inc. and the German SAP AG to effectively execute such scenarios.[6]

In light of the significant value of AI, it is crucial to explore the barriers that hinder companies from adopting this technology. It is also important to understand why AI needs to be explainable, even when taking into account its macro- and microeconomic implications. (s. Kraus et al., 2022). Explainable AI is considered of utmost importance in various sectors of the economy such as healthcare, manufacturing, construction, finance, and the process industry. This is because understanding and explaining why certain decisions have been made is a prerequisite for users to accept those decisions. As technology advances, it is becoming increasingly important to understand the decisions made by algorithms. The growing importance of explainable AI (XAI) reflects this need for transparency and accountability. With black box models becoming more prevalent, it is crucial to have tools and techniques in place to interpret their outputs. This will not only help build trust in AI but also ensure that the decisions made are fair and safe. What particularly is to be noticed is the importance of deep neural networks. These are neither interpretable, i.e., the "inner" mechanisms are not apparent to the user, nor are the decisions made by the models. Individual decisions can be explained, for example, using tools, which means that so-called local explainability is possible. However, the tools used for this can only be used by experts. These are, see Chapter 3, e.g., LIME, SHAP, Integrated Gradients, LRP, DeepLift, GradCAM, ELI5 etc. Tools that can also be used by non-experts include saliency maps, counterfactual explanations, prototype, or surrogate models, etc. These are explained in chapter 3. Basically, these models are machine learning ones. They are then divided into so-called white or glass box and black box models. The decisions of these black box models and the mechanisms for making these decisions cannot be explained to experts.

In terms of AI as a whole, a distinction can be made between Symbolic or GOFAI (Good Old-Fashioned AI) and subymbolic AI or connectionist AI (Newell & Simon, 1988; Winograd, 1971; Fikes & Nilsson, 1971; Ilkou & Koutraki, 2020). These methods can also be classified historically. Thus, the models of symbolic AI "originate" from the first AI wave and were authoritative until around the mid -1980s, while the subymbolic AI methods that

[6] https://www.merckgroup.com/de/news/palantir-healthcare-acceleration-partnership.html and https://news.sap.com/germany/2023/02/digitalisierung-cloud-merck/ , both accessed 18.06.2023

emerged then are still very popular and resonant today due to their enormous performance (Kautz, 2020). Experts now speak of the so-called third wave of AI methods, in which the two categories of methods are combined. In this context, experts sometimes refer to neuro-symbolic AI, whereby there are several gradations and different categorisations made depending on the author.

Symbolic AI	Subsymbolic AI
Symbols Logical Serial Reasoning v. Neumann machines Localised Rigd and static Concept composition and expansion Model asbtraction Human intervention Small data Literal/precise input	Numbers Associative Parallel Learning Dynamic Systems Distributed Flexible and adaptive Concept creation and generalisation Fitting to data Learning from data Big data Noisy/ incomplete input

Table 1 Differences between symbolic and subsymbolic AI

As shown in table 1, the two categories differ quite clearly in their characteristics. According to Ilkou and Koutraki (2020), the two categories differ in three aspects: firstly, in the results, in the way people or, more precisely, users interact with the models, and in the data provided as input to the models. The symbolic methods provide logical conclusions, whereas the subymbolic methods provide associative results. Intervention and, therefore, the initiation of "learning processes" by the user is common in the symbolic methods, whereas this is not provided for in the subsymbolic methods and the models learn from the data given. The symbolic methods work and deliver the best results when they work with relatively few but precise data, whereas the subsymbolic models require a large amount of data, which also contain a large part of so-called *noisy data*. A detailed explanation of different categories of neuro- symbolic AI is given in Chapter 3.2.3 (McCulloch & Pitts, 1943; Minsky & Papert, 1969; Rosenblatt, 1958; Hopfield, 1982).

As already mentioned above, acceptance by the user only results from the explicability of AI. But also, other regulations, e.g., the compliance requirements resulting from the transparency requirements of the European GDPR (s. Blackman, 2022). In certain industries, e.g., in the process industry, they are indispensable for the certification of processes within

the framework of the so-called GxP (Good Manufacturing Practices - GMP) requirements or for the fulfilment of requirements by the FDA or REACH. Thus, AI systems, including a concrete instantiation of the reference architecture to be developed here, may have to be subjected to a certification process in order to meet the extensive regulatory requirements, especially in the process industry.

According to a recent study by McKinsey (Grennan et al., 2020), companies that already receive 20% of their EBIT from AI are more likely to implement practices that help explain how their AI models work (McK, 2021). A noteworthy discovery is that businesses relying on AI models' explainability to engage with their customers witness a yearly revenue and EBIT growth of at least 10%. Grennan et al. (2022) see different requirements for XAI depending on the stakeholder (see chapters 2.3.4 and 3.5). In Grennan's (2022) findings, XAI offers five key benefits to companies. First, it increases productivity. This allows the system to be monitored more effectively by detecting errors or biases in the data thanks to explainability, which enables timely corrections. Second - Building trust and thereby greater adoption- Explainability is important for building trust. Society, users, etc. (see Stakeholder Chapter 2) have an interest in AI making a fair and confident decision. By ensuring that the decisions made by the models are explainable, trust and acceptance can be built between the company and its customers. This also applies to internal users. For example, if sales teams are able to understand the decisions of the AI model, they will trust the model and use it even if it makes decisions that are "incomprehensible" at first glance, such as a navigation system that suggests an alternative route based on information that is not (yet) available to the driver. Thirdly XAI can be used to uncover new value-generating interventions. For instance, one may use process analysis to optimize operations in process-driven companies or determine the reasons for high customer churn rates. Similarly, an insurance company may analyse feature combinations to better handle accidents and optimize their tariffs accordingly. Furthermore, fourthly, XAI can create added value for companies by allowing business to verify the anticipated business benefits of their AI models and provide the desired value. At least, fifthly, as previously mentioned, XAI offers a significant benefit in ensuring that an AI system (CSR) adheres to all relevant regulations, ethical standards, and laws. This confirmation is crucial for maintaining compliance.

Explainable AI is a sine qua non, especially for the process industry, for example to meet the certification requirements of certain regulators. In addition, significant economic factors can also be demonstrated.

The findings discussed above from both macro and microeconomic perspectives will be a reference point in the following, not to mention the subdivision of AI into symbolic and subymbolic AI methods.

Finding 4: Front runners participate most. This could lead to “supercompanies”.

Finding 5: XAI can help to overcome barriers against AI- XAI can be also a needed requirement for specific industries to use AI (s. regulations in process industry- s. chapter 2)

Finding 6: The impact of AI on the situation of work at the company level, as a competition for the greatest talents and the best skills, is closely linked to the "front runner" benefit most. Because the "front runner" companies will also gain the best talents and skills. As a result, according to studies, companies have the task of training their employees extensively in order to ensure the best possible use of AI in the company.

1.3 Motivation and Relevance

Artificial Intelligence (for short, AI) may be implemented to support decision-making and aid decision-makers in decision situations. In recent years, AI models have been used with increasing frequency to support decision-making processes; however, the newer and more successful models are primarily subsymbolic black box models which lack explainability.

Explainable AI aims to achieve the explainability of the AI models being used, as AI systems are able to make an increasing number of autonomous decisions without any human intervention. As computing power grows, AI technology and models will be implemented progressively in everyday life systems (e.g., edge AI). If the decisions being made affect and influence living things, the demand for trust in AI will become all the higher. Initially, the discussion on Artificial Intelligence (John McCarthy first coined “Artificial Intelligence” or “AI” in 1955, in a proposal for a summer research project[7] at Dartmouth College, McCarthy (1955)) gave research priority to symbolic AI – AI, which Simon defined as

[7] Took place in Summer 1956

"[…] aimed at programming computers to do things which, if done by a person, would be regarded as intelligent" (Simon, 1977, p.1187); more recently, this has been defined by Russell & Norvig (2022) as being focused on "the study and construction of agents (agents and models used synonymously) that do the right thing – 'the right' thing is being defined on the objective we give the agent."

AI can be distinguished in terms of symbolic AI methods and subsymbolic methods. Symbolic AI and models were used from the beginning of AI in 1956 (and even earlier), whereas statistical, connectionist, and subsymbolic methods grew in Machine Learning during the 1990s to become the more specific field of Deep Learning (using Deep Neural Networks, or DNN), especially DNNs with enormous parametric space and hundreds of layers comprising the so-called "black box" models. While connectionist models are more powerful, with regard to accuracy, they lack explainability and are more complex. It may be claimed that there exists a trade-off between accuracy and explainability or understandability (Breiman, 2001).[8] By contrast, models which are easy for users to understand are referred to as "glass-box" or "transparent" models.

As mentioned above, transparency is necessary when it comes to decision making, e.g., in medicine (and precision medicine in particular), to understand why a specific diagnosis was made, or in the military, when an AI system has identified an object as an enemy craft (e.g., a tank or a plane). A lack of transparency may mean a lack of understandability, which in turn leads to mistrust; subsequently, humans may act hesitantly, or reject the use of AI technologies. This may both slow down the adoption of new, promising technology and cause harm – for this reason, in medicine, there is a demand for having "a human in the loop", to make the final decision when it comes to a recommended medical treatment.

[8] In his 2001 article Breiman (2001) points out that Occams's Razor "the simpler is the better" is not working when it comes to evaluate between accuracy and interpretability- A linear regression is a good interpretable model how y,x relates- but it has a lower accuracy when it is compared to the less interpretable neural nets- The same is for random forests- instead of using one tree as a predictor there are fifty or more trees grown on the same data. The single tree rates an A+ for the interpretability - but they are fewer good predictors - while random forests are very good A+ predictors - regards interpretability they rate F. Therefore, Breiman states:"Accuracy generally requires more complex prediction methods. Simple and interpretable functions do not make the most accurate predictors." and "Using complex predictors may be unpleasant, but the soundest path is to go for predictive accuracy first, then try to understand why."

In terms of understandability, it is highly relevant which stakeholders the model is explained to, and in which situational context XAI is being used. In situations where a decision must be made in real time, an explainable model whose findings or possible alternatives require several minutes to explicate will be useless. This pertains to the efficiency requirement and the instance of an XAI system implemented in an autonomous-driving car. The usage of methods currently discussed and developed in XAI by statisticians and within mathematical functions of LIME[9] (Local Interpretable Model-agnostic Explanations) or LRP (Layer Relevance Propagation), for example, must be questioned in such a context, as humans tend to assign human-like traits to it; they are also more likely to utilise anthropomorphic terms when using AI models. Therefore, human-like traits need to be addressed by an XAI system – the one driven by the idea of creating something "human-like" though artificial, such as the Golem, a human-like being built from clay. More supportive research must be performed to understand how humans explain both a decision and actions to other humans (Miller et al., 2017; Görz et al., 2021; de Graaf & Malle, 2017) There is also growing doubt as to whether XAI is necessary per se, and whether it can become counterproductive or at least harmful to specific stakeholders when they are confronted with a given explanation. This could occur because humans do not necessarily provide insights into the workings of their brains while explaining their decision(s) to another person. The present work argues with this standpoint to an extent, as it comprises a certain shortcut regarding accountability, for instance, which is viewed differently when comparing AI and humans. The work will indicate that a precise analysis must be provided, and specific requirements ought to be identified in order to formulate a definition, namely what kind of explanation for a specific model this could be, and how that should be presented to a selected group of stakeholders. Further, since situations (contexts) might present a challenge to XAI, as already mentioned, the XAI methods must be analysed with regards to the cost of runtime and implementation complexity, while taking into account the fact that in an XAI system, two components are necessary to explain the decision and action to a stakeholder. The first of these is the explainer component, and the other a context-appropriate explanation user interface, whether that is by way of a natural language, graphic, or haptic explanation (Gunning, 2019).

As mentioned above, when van Lent, and Fisher used the wording "explainable artificial intelligence" or XAI in 2004, explainable artificial intelligence already looked back upon

[9] S. chapter 3.3.1

a much longer history (Lent, van, et al., 2004; Hansen & Rieger, 2019). The research in explainable artificial intelligence reaches back more than 50 years, when it emerged along with advances in connectionist and opaque Machine Learning/Deep Learning models used in AI. The research on explainable AI had already started with research on experts' systems, e.g., in the late 1960s, with the SCHOLAR system or Stanford's MYCIN. While SCHOLAR was designed to explain why a student's answer was wrong, MYCIN was designed with three components: a rule-based decision support component, an explanation module (or component), and a learning module (Shortliffe & Buchanan, 1984; Clancey, 1983; Hansen & Rieger, 2019).

Ryszard Michalski (1983) put forward his postulate of "comprehensibility" in 1983, claiming that the result of machine learning or the result of learning processes within artificial intelligence should be of symbolic representations: "[…] should be symbolic descriptions of given entities, semantically and structurally like those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single chunks of information, directly interpretable in natural language, and relate quantitative and qualitative concepts in an integrated fashion." (Michalski, 1983, p. 519) Indeed, most modern approaches to exploring the explainability of ML only make it comprehensible in terms of how the model generates output from a given input, thereby (nearly) recognising Michalski's postulate (Görz et al., 2021). In the light of other authors' beliefs, which are mostly influenced by cognitive sciences, this dissertation asserts the importance of doing more basic research to reach convincing results in XAI. A good example of such research is to determine which explanations are more convincing, and by which potential stakeholders they might be used in a specific context; this could include the findings of the longer research on the history which relates to intelligent systems and with rudiments in the social sciences, psychology, and the cognitive sciences. Therefore, those more complex ML explanations, which are supported by statistics and mathematics, do not address the requirements of all stakeholders. The whole explanation procedure ends up being an inappropriate exercise, like "inmates running the asylum" (Miller et al., 2017). This dissertation agrees that a plausible justification of a decision or action made on the basis of current scientific rationality is only possible by combining symbolic and subsymbolic approaches using so-called "hybrid" or knowledge-based approach (d'Avila Garcez & Lamb, 2023 resp. 2020; Ilkou & Koutraki, 2020; Bibal & Frénay, 2016; Caruana et al., 2015; Fürnkranz et al., 2020; Görz et al., 2021; Marcus, 2020; Marcus & Davis, 2021; Tiddi, 2020, Chari et al., 2020a, 2020b).

But, before proceeding further, some definitions of key terms used in this work must be presented.

Artificial Intelligence - AI

There are several definitions of *Artificial Intelligence* (*AI*). Some refer to how good the AI in terms of fidelity to human performances. Some authors define the *intelligence* in AI as rationality, in the sense of doing "the right thing". Others consider intelligence a property of an internal thought and reasoning process. In contrast, still others focus on intelligent behaviours, namely their external characterisation (a behaviouristic explanation). Human performance or rationality, along with thoughts or behaviours, may constitute four possible combinations when it comes to the definition of AI.

In this work, AI is defined, following Russell & Norvig's (2021) definition, as the "standard model" of AI. Thinking of an agent (which in this work is placed equal to model or algorithm) operating autonomously while perceiving the environment as persisting over a prolonged period, adapting to change, and creating and pursuing (the right) goals. AI focuses on the study and creation of agents that do the right thing; here, the right thing is defined by an objective provided to the agent (Russell & Norvig, 2021).[10]

Machine Learning

Machine Learning is defined as a part of *Artificial Intelligence* and is about improving an agent's performance through learning after making observations about the world. The agent will be a computer -- that is why it is called machine learning; it observes data, builds a model based on that data, and subsequently uses the model as a hypothesis about the world and an algorithm that can solve problems (Russell & Norvig, 2022) *Deep Learning* is a specific part of machine learning, which uses large neuronal networks.

Definition of an Expert System

An expert system, as a part of an *Artificial Intelligence system*, is a computer program designed to model a human expert's problem-solving abilities (Durkin 1994).

[10] Since perfect rationality is not possible in complex environments or only serves as a theoretical starting point, bounded rationality or bounded rationality is a reality. (Gigerenzer & Selten, 2002; Simon, 1957; Elster, 2009)

Definition of Accountability

Accountability refers to responsibility and justification of the decisions and predictions made (Rosenfeld & Richardson, 2020).

Transparency

A model is considered transparent provided that it is understandable. There are different levels of *transparency*: simulatability (entire model), decomposability (at the level of individual components -- input, parameter, and calculation), and algorithmic *transparency* (at the algorithm level). Therefore, a "black-box model" is one which is not transparent at the levels mentioned above. Transparency is the opposite of "opacity" or "black-box-ness" (Lipton, 2016).

Understandability

Barredo Arrieta et al. (2020) define *understandability* within the context of an ML system as the characteristic of a model which allows a human to understand its function and learn how the model works, which he places equal to intelligibility.

Interpretability

Interpretability can be defined as the science of comprehending what a model has done, e.g., by using LRP (Layer Relevance Propagation), a method specifically for making DNNs (Deep Neural Networks) interpretable by revealing which of the (input) features were more relevant for the classification that a picture is a dog and not a cat (Bejger & Elster, 2020). Nevertheless, being interpretable in terms of meaning provides visual cues to find the "focus" of a DNN. The identification of the most dominant classifiers by simplifying the problem space locally, using a more interpretable model (with a kind of intrinsic "explainability"), does not necessarily solve all the problems and questions that various XAI stakeholders might have. Therefore, explainability is needed, which is broader and covers interpretability (Bibal & Frénay, 2016; Gilpin, 2018). By implication, interpretability is insufficient for humans to be able to trust in the decisions of black box models. There is a need for explainable models that can summarise the reason for a specific neural network behaviour; by doing so, in producing insights about the causes of the model's decision, user trust may be gained. Explainability mechanisms also need to be able to defend

(or justify) their actions (demonstrate accountability), provide relevant responses to questions, and at least be auditable. Explainable models are interpretable, while the reverse is not always true (Gilpin et al., 2018).

Explainability

As already mentioned, explainability includes interpretability, and is therefore a more general concept than interpretability. It is more related to an "intrinsic" explanation of a machine learning model and how a function can be communicated to a user regarding completeness, which is a close enough approximation (Hansen & Rieger, 2019). Interpretability can be seen as the first step, though users (humans, stakeholders) require explainability to gain trust in the decisions made by the model. That implies there is a need for models which summarise the reasons for a specific behaviour and produce insights about the cause (and effect chains) of a given decision. The models should be "able" to defend their actions, provide relevant responses to questions and be audited (all in a stakeholder-specific, understandable form) (Gilpin et al., 2018; Hansen & Rieger, 2019). Explainability is also very much associated with an interface appropriate to a specific stakeholder group and within a specific context. Therefore, an explainable interface is one part -- besides an explanation module, component, or system -- used "to explain" in a form proper to the entity requiring it (be it a human, other living entity, or a machine) (Arnold et al., 2021; Gilpin et al. 2018).

The difference between *interpretability*/*explainabilit*y (the first is more the "*understandability*" of a model during "runtime" directly when the decision is made) versus an Explanation is as follows: the first is somewhat intrinsic, while the second is more or less explicit (done afterwards). Hence, there is a high correlation as to within what situation a given kind of XAI is being used.

Comprehensibility

Comprehensibility for ML models goes back to Michalski's postulate, stating that *comprehensibility* refers to the ability of a learning algorithm to represent its learned knowledge in a human-understandable fashion. According to Barredo Arrieta et al. (2020), "the results of computer induction should be symbolic descriptions of given entities, semantically and structurally like those a human expert might produce observing the same entities. Components of these descriptions should be comprehensible as single 'chunks' of information,

directly interpretable in natural language, and should relate quantitative and qualitative concepts in an integrated fashion" (Barredo Arrieta et al., 2020; Görz et al., 2021; Michalski, 1983).

Fair and Ethical Decision-Making

There is an increasing demand by the public for *fair and ethical decision-making* alongside *explainability*, e.g., concerns raised by politicians and other stakeholders that AI or algorithmic decision-making is influencing social life more and more, such as the COMPAS system. Pursuant to the GDPR of the European Union, individuals affected by any algorithmic decision have the right to file a claim (Bejger & Elster, 2020; Goodman & Flaxman, 2017; High-Level Expert Group on Artificial Intelligence, 2019; Lipton, G., 2016).

Reliability and Robustness

Reliability refers to the model of being objective and unbiased, and *robustness* refers to the strength of the model against gaming (e.g., "gaming the system") or conceptual drift. A conceptual drift arises when a decision made by a model changes its environment -- such that the model no longer fits the environment that it had learnt; this is also a challenge concerning model lifecycle management. This topic is particularly important with regard to the entire life cycle of an ML model (Suresh & Guttag, 2021).

Process Industries

The *process industries* have a crucial role in the commercial transformation of raw materials into finished products. The processes involved in this transformation typically require both physical and chemical changes, at times requiring biochemical changes; these engineered processes take place within processing plants. Most of the products of the *process industries* have well-defined specifications, and the industries themselves can be usefully classified according to the type of feedstock or product involved. Examples may include petroleum refining, mineral processing, chemical processing, and the production of fertilizers, food, and pharmaceuticals (Brennan, 2020). For the purposes of this dissertation, the focus is placed on chemicals and pharmaceuticals (in short pharmacy).

Corporate Planning (Scenario Planning and Integrated Corporate Planning)

Corporate planning refers to the rational anticipation of future operational events. Planning deals while thinking about the future, and with a goal-oriented approach, is central. Such

goals must be articulated clearly among the different areas and subareas of a given company. Aside from decision-making, planning is one of the core capabilities of management or leadership. It must be forward-thinking, overarching, and not limited to one company; it is related to many disciplines, e.g., finance, production, sales, distribution, investment, and so forth. The reason for planning originates when an affected person (or machine – e.g., in machine-to-machine communication) that is a planner, manager, or decision maker, encounters or discovers a specific state (or system state) which when compared to a desired state, needs improvement or change. That perceived gap between the two (!) states is not satisfactory or not acceptable (any longer), in relation to external requirements. Generally, one speaks of a problem - the deviation in a current or expected state from a desired state as described by established goals. One may also speak of a decision-related problem, as certain decisions have to be made and enforced in order to solve the problem, i.e., to eliminate the aforementioned deviation from the desired state (Klein & Scholl, 2011; Wild, 1980). Wild formulates it similarly, when describing planning as a "systematic, future-oriented thinking through and setting of goals, measures, means and ways not only company-related goal achievement" (Hammer, 2015).

Scenario Planning

Scenario Planning can be seen as a controlled method for possible imaging futures that companies have applied to various issues. The Shell oil company has been using scenarios since the 1970s to generate and evaluate strategic options. By identifying fundamental trends and uncertainties, a manager can then construct a series of scenarios that might help “to compensate for unusual errors in decision making – overconfidence and tunnel vision” (Schoemaker, 1995).

Integrated Business Planning

Traditionally, supply chain planning focuses on volume-based planning. More modern approaches emphasize a value-based approach with a greater focus on financial flows. This is the concept behind *Integrated Business Planning* (in short IBP) Still, many authors see IBP as a mere restatement of mature S&OP process characteristics. By contrast, others see IBP as a suitable interface between Sales and Operations Planning and Financial Planning (Coldrick et al., 2003; Willms & Brandenburg, 2019).

Reference Architecture

There is no commonly agreed definition of *reference model* or *reference architecture*. However, most authors see *reference architectur*e as a continuation of a *reference model* focusing on software technology; certain authors use the two terms synonymously. Bass defines them thus: "A *reference architecture* is a reference model mapped onto software elements (that cooperatively implement the functionality defined in the *reference model*) and the data flows between them. Whereas a *reference model* divides the functionality, a *reference architecture* is the mapping of that functionality onto a system decomposition" (Bass et al., 2022).

As the main goal of this work is to develop a comprehensive reference architecture for XAI systems, *reference architecture* is seen herein as an abstraction on a meta-level, whose idea and goal are to help the design, development, and implementation of systems. It provides knowledge in the sense of best practices, and satisfies requirements dictated by the environment and scientific rigour. It is a framework: a reference architecture combines and synthesizes a technical perspective with a domain knowledge (Reidt et al., 2018; Brocke, vom, 2003) *A reference architecture* has specific layers, such as business, functional, process, information, and system layers. Across those, there might be different perspectives, like "Governance", "Explainability", etc.

1.4 Research Goal and Research Questions

When decisions and actions made by an AI model in corporate planning scenarios and decision-making are not explainable to stakeholders, they are not trusted. As these models need to be more transparent, interpretable, or explainable, they are not used to their full potential (the difference between interpretability/ explainability and explanation depends on the situation in which the model is used). This work proposes that most managers and decision-makers in business need more mathematical and statistical knowledge to understand decisions or actions made by subsymbolic black-box machine learning and profound learning models. A sustained lack of stakeholder trust may slow down or even prevent the adoption of AI approaches and models within a corporate planning - business context.

Hypothesis:

By developing a reference architecture for an explainable AI system that could combine both subsymbolic and symbolic approaches, confidence in AI models and, thus, decision-making in corporate planning can be improved.

The primary goal of this dissertation is to establish and create a reference system architecture that promotes explainable artificial intelligence, with the aim of improving decision-making capabilities to facilitate better business planning within the process industry. The research has resulted in a reference system architecture for trustworthy AI in corporate planning, which is the main contribution of this work. To the author's knowledge, there are no previous or other comparable works in this domain.

This work examines a crucial research question: How can an explainable artificial intelligence system, or agent, be created and integrated into the planning framework of the process industry to increase trust in decision-making AI systems by improving their transparency and decision quality?

Corporate planning, or planning, entails the mental anticipation of future operational events, thus planning deals by thinking about the future, and doing so while having a goal-oriented approach. These goals must be stated clearly among the different areas and sub-areas of the company. Aside from decision-making, planning is one of the core capabilities of management or leadership, and goal-oriented, forward-looking thinking is not limited to one company. Planning is a core element of business and is central in all business disciplines. Therefore, planning is a decision problem, which may be examined from different perspectives, e.g., business administration follows a rationality paradigm, with a model of the rational thinking “homo oeconomicus”; the cognitive psychologists prioritise the processes in the mind of the decision maker; game theorists are interested in mathematical decision behaviour; the behavioural economists are interested in the changes in decision-making behaviour in particular contexts, etc. Of note here is that the quality of decision-making is significantly improved through the usage of AI models, as humans tend to bias decision-making with emotions and irrational behaviours. Humans also lack information about the situation the decision must be made within (bounded rationality). Humans tend to base their decision-making on subjective, past experiences - even when the context of the situation does not fit. (Gigerenzer & Selten, 2002; Dörner, 2001; Elster, 2009; Simon, 1957) Recent studies have found a machine-hybrid or augmented approach, which could beat the best chess computers within a game, for instance, and reach better results than AI or a human, alone (e.g., s. De Cremer & Kasparov, 2021).

In addition to the hypothesis already mentioned, this work also answers other research questions. These are:

RQ[11]1: What are the specifics of the process industry?

RQ 1.1: What are the main and differentiating characteristics of the process industry?

RQ 1.2: What are the specific market conditions of the process industry?

RQ 1.3: What does the planning process look like within corporate planning?

RQ 1.4: What special planning sub-processes in corporate planning are of particular importance for the process industry?

RQ 1.5: What decisions are made in these sub-processes that AI systems can/ will take over?

RQ 1.6: What are the requirements for explaining decisions made in the sub-processes?

RQ 2: What is Explainable AI and how can it support decision making in the corporate planning process?

RQ 2.1: What is AI

RQ 2.2: What is Machine Learning?

RQ 2.3: What are knowledge-based systems?

RQ: 2.4 What is explainable Artificial Intelligence?

RQ: 2.5 What are the Stakeholders of XAI and how do they relate to the stakeholders in corporate planning?

RQ 3: How is a Reference Architecture for an explainable AI system being designed and developed?

RQ 3.1: What are the various theoretical approaches for constructing a reference architecture?

RQ 3.2: What methodology for designing and developing a reference architecture can be provided?

[11] RQ = research question, G = goal

RQ 4: How to provide guidance on creating a reference architecture for explainable artificial intelligence in the operational planning context?

RQ 4.1: To create a reference architecture, what preparations and basic assumptions need to be taken into account? Moreover, what factors should be considered throughout the lifecycle to guarantee explainability?

RQ 4.2: What are some existing architectures that could be used as a foundation?

RQ 4.3: How can the requirements be summarised?

RQ 4.4: What is the Business Layer of Re_fish?

RQ 4.5: What is the Application Layer of Re_fish?

RQ 4.6: What is the Technology Layer of Re_fish?

RQ 4.7: What is the process for managing the lifecycle of an explainable AI system?

RQ 4.8: How can a reference architecture be evaluated?

RQ 4.9 What is the gap between the generic framework and expert opinion?

Main Goal

G[12]1: The main goal of this work, taken from the hypothesis, is to develop a reference architecture as a reference model which can be used for design development, as well as implementation and runtime of a trustful and reliable XAI system. The designed reference architecture is called "Re_fish" (in tribute to Marian Rejewski, the leading Polish scientist solving the Enigma code and the Babelfish – "a fictional universal decoder for any form of language in the universe" (Adams, 2010). The empirical relevance of the reference architecture will be developed with scientific rigour, within a process industry corporate planning context (Futia & Vetrò, 2020; Marcus, 2020; Marcus & Davis, 2021; Sohrabi et al., 2018).

In addition to the main objective G1, mentioned above, the thesis has the following secondary objectives or specific goals.

G1.1: The thesis will provide an overview of the actual status and research of the impact of Artificial Intelligence on the economy. This goal is mapped to Chapter 1.1 and 1.2

[12] G = goal

G1.2.: The thesis will provide an overview on the specifics of the process industry, challenges the process industry is facing and how AI can support business in the process industry. This goal is mapped to Chapter 2

G1.3: The thesis will provide an overview of the actual status and research in AI and XAI. This goal is mapped to Chapter 3

G1.4: The thesis will provide an approach on how to develop a reference architecture for a trustworthy AI (XAI) system. This goal is mapped to Chapter 4

G1.5: The thesis at least will provide a system reference architecture – Re_fish, which can be used by instantiating to build a trustworthy AI- XAI system. – This is a direct subgoal of the main goal (repeating it) and mapped to Chapter 5

1.5 Research Theory and Design

The research methodology for this dissertation is grounded on design science research, whose roots are based within engineering and artificial sciences (Simon, 1996). Its primary purpose is to be a problem-solving paradigm for creating new and innovative artifacts, and not to analyse an existing and observed phenomenon within the behavioural paradigm, with its roots in natural sciences. These artifacts must be built and evaluated in a rigorous design process and could be of different types, like models and methods (Hevner et al., 2004). Design science is comprised an object of study, as well as two main components: the design of the artifact and the investigation of the artifact in context, in accordance with a definition by Wieringa (2014).

These artifacts, in turn, define the ideas, practices, technical capabilities, and products through which the analysis, design, implementation, use, and management of information systems can be accomplished efficiently and effectively (Hevner et al., 2004). By contrast, the behavioural science paradigm (the principal research methodology in Anglo-Saxon research) seeks to develop and verify theories that explain or predict human or organisational behaviour. The goal of the design-science paradigm seeks to create new and innovative

artifacts and, in doing so, tries to extend the boundaries of human and organisational capabilities by providing intellectual and computational tools (Hevner et al., 2004). However, technology in IS research and behavioural science are not dichotomous. Indeed, they are inseparable; design science can serve as a bridge between those paradigms, to resolve the conflict among pragmatists arguing that truth and utility are two sides of the same coin, and that scientific research should be evaluated in terms of its practical implications (Hevner et al., 2004).

The design-science paradigm is based on the knowledge and understanding of a problem domain and its solution, which are reached and achieved by building and applying an artefact, designed in a systematic process for relevance (for solving a problem) and under the rigour of scientific research. IT artefacts can be differentiated into constructs (vocabulary and symbols), and models (abstractions and representation, e.g., reference models/architectures) (Hevner et al., 2004).

To validate the artefact, different tools can be used, such as field studies, which are for behavioural researchers to understand organisational phenomena in context. At the same time, designing, building, and using innovative artefacts, like by building a prototype, will help design science researchers understand the problem addressed by the artefact and validate and understand its feasibility.

Information systems support organisations, and they are "[…] complex, artificial and purposefully designed" (Hevner et al., 2004, p. 78). The connection between organisations and information systems can be seen in figure 2 (Hevner et al., 2004; Henderson & Venkatraman, 1993)

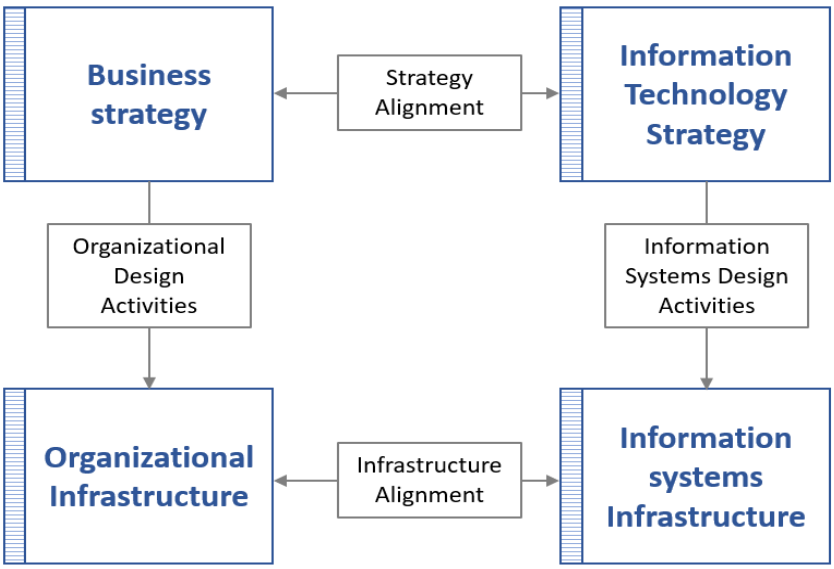

Figure 2: Business and Information Strategy and Organisation (Henderson & Venkatraman, 1993)

Business Strategy and Information System (IT) Strategy are aligned (Business IT Alignment) the Organisational Infrastructure is being built by organisational design activities, which are derived from the business strategy, while the information systems infrastructure is designed by activities derived from the information system strategy (Henderson & Venkatraman, 1993; Hevner et al., 2004). Design activities and the design itself are a process and a product. That means that it is a process of expert activities that produce an innovative artefact or product. The evaluation of the artefact then leads to a better understanding of the problem, and the manner in which the problem is solved gives feedback to improve the quality of the product as well as the quality of the design process itself (Hevner et al., 2004).

March and Smith (1995) identify in their research that there are two design processes and four artefacts within design science research.

The two processes of Design Science are:

1. To build (the build process)
2. To evaluate (the evaluation of what was build)

The four artifacts of design science are:

Constructs

Provide a language in which problems and solutions are defined and can be communicated (Schön, 1983).

Models

Use the constructs mentioned above to represent the real-world situation (Simon, 1996). “Models aid problem and solution understanding and frequently represent the connection between the problem and solution components enabling exploration of the effects of design decisions and changes in the real world” (Hevner et al., 2004, p. 78/79).

Methods

Define processes to guide how to solve problems.
Provide instantiations.
Show that a working system can implement constructs, models or methods.

Instantiations

Show that constructs, models and methods can be implemented in a working system, as they demonstrate feasibility, and provide a direct evaluation of a system in its intended, purposeful usage (Hevner et al., 2004).

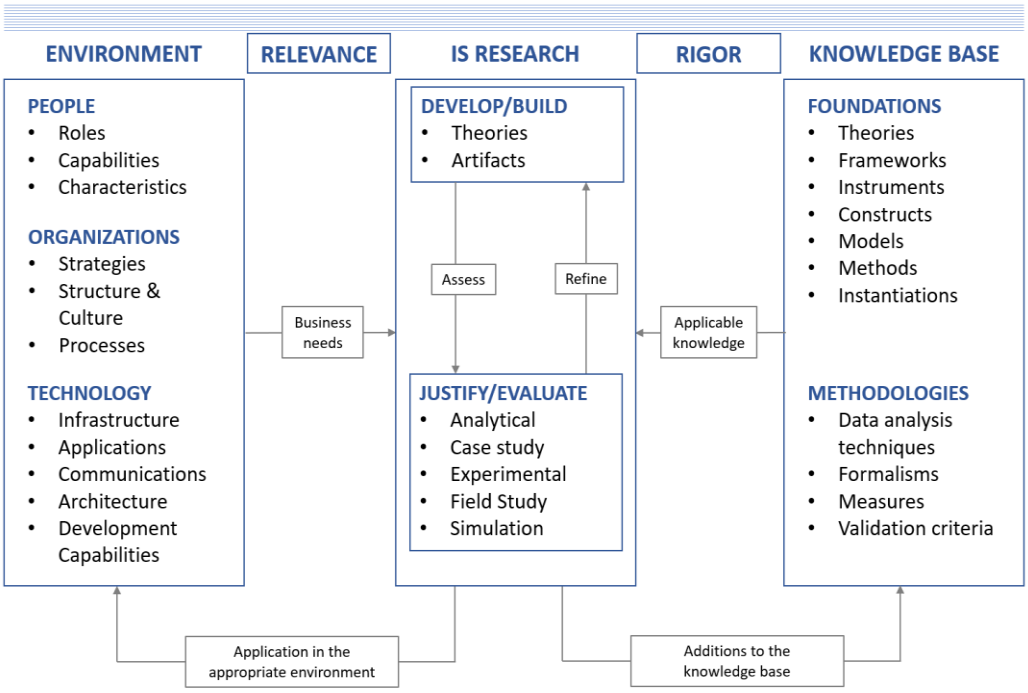

Figure 3: Concept of Design Science research, aligned with Hevner et al. (2004)

The concept of design science research built by Hevner et al. (2004) is shown in figure 3, where IS research can be found in the middle; it respects the relevance of business needs

"to solve a real-world problem", which is raised by the environment, including people, organisations, and technology. The environment defines this problem space (Simon, 1996). On the other side of the coin, IS research is based on the "knowledge base" (for providing scientific rigour) by building research on foundations (theories, frameworks, etc.) and methodologies (data analysis techniques, formalisms, etc.). By developing and building theories and artefacts and justifying/evaluating them, they can be seen as additions to the knowledge base and be applied in an appropriate environment to solve a particular business problem. The scientific rigour is reached by appropriately applying existing foundations and methodologies (Hevner et al., 2004).

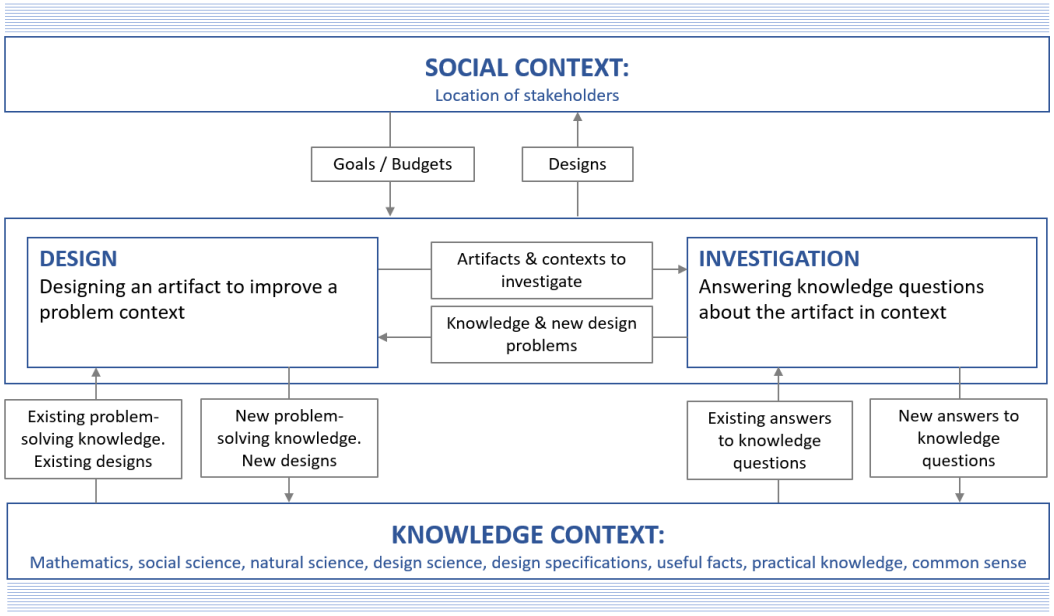

Figure 4: Design science framework by Wieringa (2014)

Wieringa has adjusted the design science framework by Hevner et al. (2004). As shown in figure 4, the social context (top to bottom) is expressed by identifying the stakeholders -- these are providing goals, which might differ from those of the researcher and budget for the research. They receive the artifact, the idea of which is to improve the context of the problem and satisfy specific requirements (within a certain margin of accuracy), so that the stakeholder goals are fulfilled. This is shown on the left side of the framework. The investigation addresses the knowledge of the artifact in the context. The knowledge context in Hevner's framework (Hevner et al., 2004) is the knowledge base, which consists of mathematics, social sciences, natural sciences, design science, etc. A given knowledge context provides existing answers for knowledge questions and receives new answers by

way of the investigation of the contextualised artifact. In the design part, there is the existing problem-solving knowledge derived from the knowledge context, new problem-solving knowledge, and a new design added to the knowledge context.

Hevner et al. (2004) define seven "Design Science Research Guidelines.":

Guideline 1: Design as an Artifact: In this dissertation, a reference architecture for building a trustworthy AI system within corporate planning will be designed, the scope of which will include the full lifecycle of an Artificial Intelligence system. The reference architecture will be a purposeful IT artefact, addressing a fundamental organisational problem: the design, construction, and running of a trustworthy AI system.

Guideline 2: Problem Relevance: The relevance of the business problem is derived from empirical analysis, e.g., that of existing literature and empirical studies. This can be seen as an unsolved business problem. While the goal of behavioural science goal is to research why a phenomenon occurs, design science aims to change the occurrence of a phenomenon. In this work, insufficient explainability of AI models (or the lack thereof) in corporate planning comprises such a problem.

Guideline 3: Design Evaluation: The design artefact utility and its efficacy must be evaluated using rigorous evaluation (Hevner et al., 2004, p. 85) [13]; the methods put forward by Hevner et al. (2004) emphasise that evaluation will be a crucial component to the design science research process. The evaluation process ensures valuable feedback, both to the design process of the artifact as well as to the artifact itself. As the design process is iterative, the quality of the process and the artifact itself will be improved. Hevner et al. (2004) differentiate among various rigour evaluation methods. In this work, two evaluation methods will be used: the first of these is an evaluation; the artifacts of the reference architecture will be shown to experts for assessment; and the second method will involve an informed argument evaluation. It will use information from the knowledge base to build a convincing case for the utility of the artifact. Gaps will be identified and documented.

[13] The evaluation can be done in terms of functionality, completeness, consistency etc. (s. chapter 5.3)

Guideline 4: Research Contributions: The research will use existing foundations and proven methodologies to provide a verifiable contribution to the design of artefacts, design foundation (e.g., reference architecture) and design methodologies (the evaluation), and the artefact itself. The artefact will be used as a starting point for further iterations (Hevner, et al, 2004, p. 87).

Guideline 5: Research Rigour: The work of the thesis is built upon applying rigorous methods in the construction, evaluation, and design of the artefact. In this work, the well-researched area of reference modelling as a foundation to knowledge used for artefact construction will be implemented. The evaluation will be done by testing the artefact – gaining expert opinions and thoroughly gathering valid arguments concerning the utility of the reference architecture.

Guideline 6: Design as a Research Process: The artefact utilises available means to reach desired ends and satisfies laws in the problem space (environment). However, design science is an inherently iterative process; therefore, this work can be seen as a starting point to search for the best and optimal solution for a reference architecture to build reliable, sustainable explainable AI systems. Therefore, it can be seen as a satisfactory solution – satisficing – without specifying all the possible solutions (as a "starting point", which can help to further investigate and help research – Simon (2019).

Guideline 7: Communication of Research: The artefact with respect to the research result of this thesis is effectively presented to both audiences – those which are technology-oriented (with sufficient detail to enable construction and implementation of the artefact) and business-oriented audiences (to enable them to use the artefact in a specific organisational context) (Hevner, 2004).

By using Hevner's et al. (2004) approach and its adjustment by Wieringa (2014), the following plan of research for the thesis is developed:

The research plan will follow the design science research methodology (see figure 5),

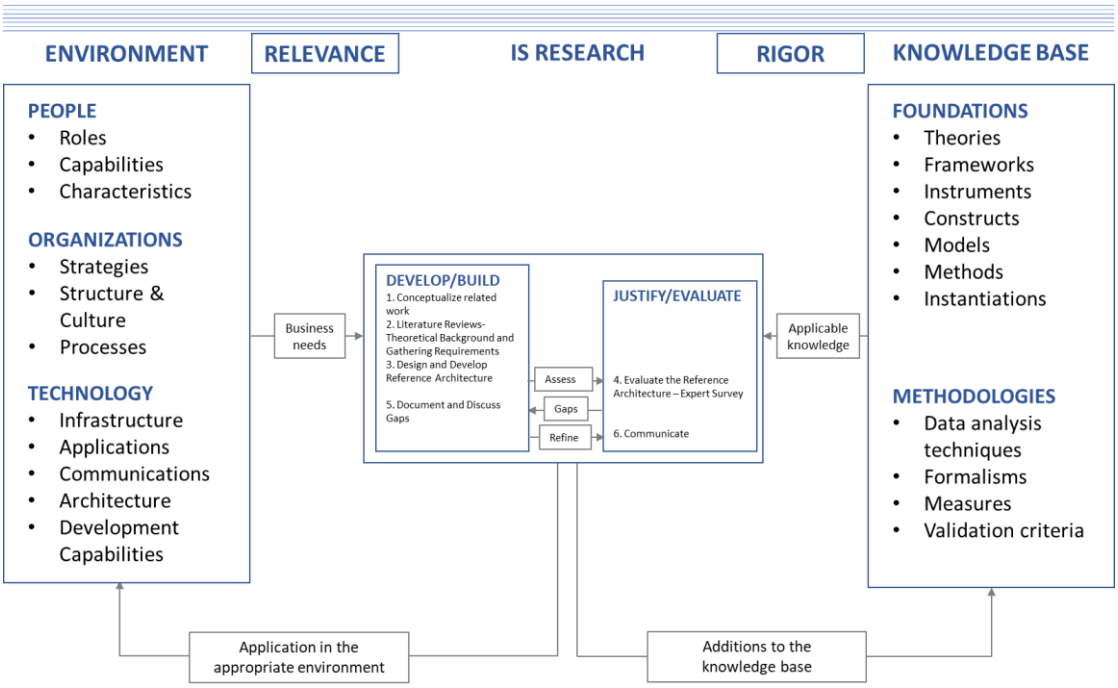

Figure 5: Plan of the research

Definition of the research methodology for the dissertation. As explained above, the research methodology is based on design science research.

1. Conceptualisation - related work. Part of the related work is enterprise planning. The work will give an up-to-date overview of the state of the art of enterprise planning in the process industry, with a focus on scenario planning and integrated enterprise planning. An important part of the work is the area of Artificial Intelligence and how it is used in this context (with a focus on planned scenarios). The status of AI in corporate planning is presented through a review of the current literature. Furthermore, a relevant field is the area of reference modelling and the construction of a reference architecture.

2. The analysis for the purpose of gathering the requirements of an explainable AI system in planning will be done using a two-step approach.

a. Step one – there will be a thorough analysis of the findings in the literature reviews, as well as a thorough analysis of the description of the requirements for a reliable explainable AI system, within the given context (corporate planning) and by using current/recent studies (e.g., Klein et al., 2021; Futia & Vetrò, 2020; Jenzen et al., 2022; Sohrabi et al., 2018)

b. Step two - the findings in the first step will be categorised and processed, to build a theoretical background and for the assembly of requirements for explanations in corporate planning situations.

   The findings of both steps – using additional literature research within the knowledge base (its foundation and methodologies) will be synthesized to create requirements for the reference architecture as being a reference model for explainable AI. In this step, existing architectures will be used as basis templates.

3. Develop a reference architecture as a reference model for designing, building, implementing, and deploying an explainable Artificial Intelligence system. A reference architecture will be developed based on previous findings and by synthesising the requirements from the relevance cycle through the scientific rigour cycle, based on the knowledge base. The resulting reference architecture is then a good starting point for further developments and improvement cycles.

4. The reference architecture will then be evaluated by using the research guidelines and the requirements from the previous steps and through evaluation by participating experts (presentation, discussion and expert survey). Following good research practices, any identified gaps will be documented to improve the reference architecture directly or in the following development cycles.

5. Any gaps will be documented and used for an adjustment directly or will be stored in a backlog for adjustment in further iteration.

6. The reference architecture will be communicated.

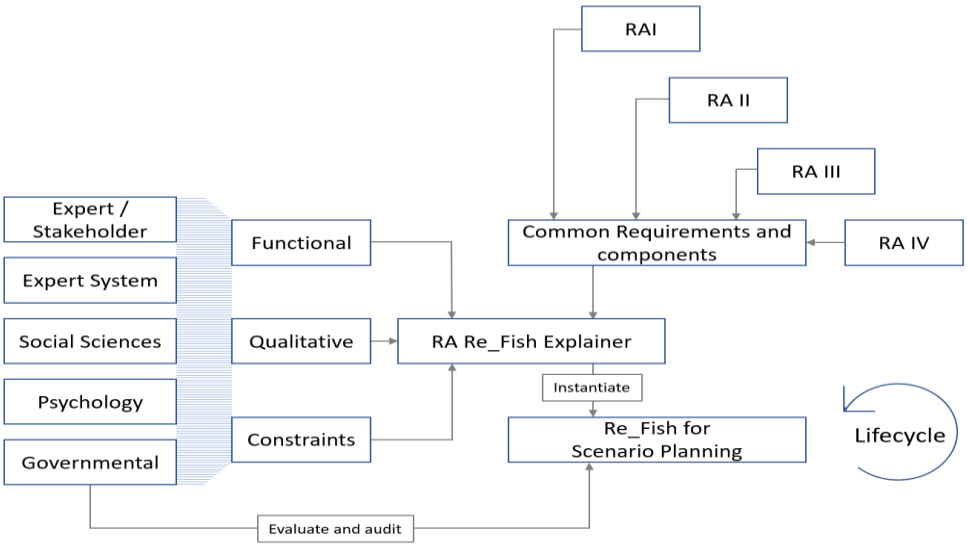

Figure 6: Detailed plan of research

The detailed plan of the research can be seen in figure 6 – from the stakeholder (based on literature and the expert survey) will come requirements – functional, qualitative, and constraint-related. Additional requirements will come through the knowledge base and literature from previous research in the fields of Expert Systems, Social Sciences, Psychology, and Government. From other Reference Architectures (RA I to RA IV) will come common and current architectural requirements, which will lead to an abstraction for building the Reference Architecture for the Re_fish Universal Explainer. The Re_fish for Corporate Planning will be a kind of specific instantiation of the Reference Architecture in that situational context. The review of the requirements, especially those of the governmental stakeholders, will be achieved through evaluation and audits.

1.6 Thesis Structure and Outline

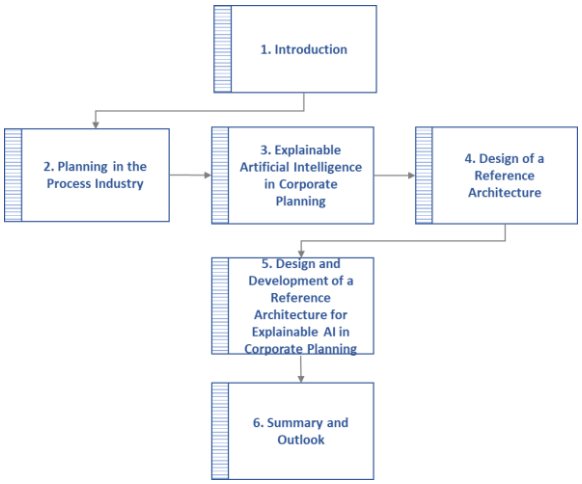

Figure 7: Thesis structure and outline

The research approach is embedded in the structure of the work and the process is illustrated graphically in figure 7.

The introduction in the first chapter is divided into four parts. In Chapter 1.1, due to its importance and relevance for this thesis, the macroeconomic perspective of AI is described, followed by the microeconomic perspective in chapter 1.2. In Chapter 1.3 the motivation and relevance of the topic are described. The motivation is based on the hypothesis that humans need to trust the decision-making process of an AI system when they effectively and efficiently use the system to improve decision-making quality. The pertinence of the topic is visible in the current tendency to implement AI systems in daily life, and the growing relevance of the decisions being made by these systems. In 1.4, the research goal and the research questions are defined. In Chapter 1.5, the research theory and design are explained and set up for the dissertation. In part 1.6, the structure and outline of the dissertation are described.

The dissertation is ordered according to its research design and is divided into four main parts. As it focuses on planning in the process industry, its features are introduced and discussed in chapter 2. After an introduction (Chapter 2.1), the specifics and relevance of the process industry are described (Chapter 2.2). Chapter 2.2.1 introduces to the specifics of the process industry. Chapter 2.2.2 introduces the key trends of the process industry followed by the challenges the process industry is facing nowadays (chapter 2.2.3). Chap-

ter 2.2.4 describes some use cases of AI for the process industry. The planning and decision-making procedures in the process industry are presented in Chapter 2.3, while examining scenario planning (Chapter 2.3.1) and integrated business planning (Chapter 2.3.2). In Chapter 2.3.3, decision types in process industry corporate planning are explained, and in Chapter 2.3.4, the stakeholders of corporate planning in the process industry are described. The way modern information systems support corporate planning in the process industry will be described in Chapter 2.4, whereas in Chapter 2.5, classical decision support systems as well as reporting, business intelligence, predictive and prescriptive analytics, and data science will be differentiated from AI systems (namely, Chapter 2.5.1 Classical Decision Support Systems, 2.5.2 Business Analytics, Predictive and Prescriptive Analytics). In Chapter 2.6, the findings will be presented in condensed form.

In chapter 3, explainable Artificial Intelligence (xAI) is presented in an overview in 3.1, followed by a description of Artificial Intelligence in Chapter 3.2. A deeper look into Machine Learning is given in chapter 3.2.1, followed by a review of Knowledge Based Systems in Chapter 3.2.2. Chapter 3.2.3 introduces Neuro-symbolic AI methods. In Chapter 3.3, explainable AI is defined and explained and brought into the context of the dissertation. To build a trustworthy AI system, explainability must be respected throughout the entire system lifecycle, and not only at the stages of development or production. Therefore, explainability must be central to an AI system's design, implementation, and production. The actuality and state of the art of explainable AI in corporate planning are investigated by a literature. Chapter 3.3.1 Introduces to XAI within machine learning and Deep Learning, followed by Chapter 3.3.2 Knowledge bases XAI and at least Neuro-symbolic XAI (Chapter 3.3.3) Subsequently, Chapter 3.4 shows the importance, relevance, and requirements of ethical, legal, and regulatory requirements for AI, and their impact on explainable AI. The growing field of AI ethics is presented shortly as well as law and regulatory requirements for AI. Chapter 3.5 maps the stakeholders of AI and the stakeholders and their requirements from Chapter 2. Chapter 3.6 closes by summarizing the findings.

Chapter 4 describes the design of a reference architecture for explainable AI systems. After an introduction in Chapter 4.1, the theoretical basis (Chapter 4.2) of reference architectures is introduced in terms of the use of rigorous methods from the knowledge base. The basis for a reference architecture is placed within IT and software architecture and is covered. The methodology of building reference architectures is presented in chapter 4.3 Methodology to Develop Reference Architectures. After describing different possible methods in

4.3.1 “Methods to develop a Reference Architecture” – the selection of the methods to develop the Re-Fish reference architecture is described. Chapters 4.3.2 to 4.3.6 follow the TOGAF and Attributive Architecture Design methodology. In Chapter 4.3.2 the Architecture Vision is introduced. The following chapters are “Establish the Architecture Project” (Chapter 4.3.2.1), “Stakeholders, concerns, and business requirements” (Chapter 4.3.2.2), “Confirm and elaborate Business Goals, Business Drivers and Constraints” (Chapter 4.3.2.3), “Define Scope” (Chapter 4.3.2.4), “Confirm and Elaborate Architecture Principles, including Business Principles” (chapter 4.3.2.5), “Develop Architecture Vision” (Chapter 4.3.2.6). In Chapter 4.3.2.7 the Phase A will be summarized. In Chapter 4.5 consists of a short discussion on the requirements. Chapter 4.6 concludes the chapter by summarising the findings. Chapter 4.3.3 describes “Phase B: Business Architecture”, with the subchapters 4.3.3.1 “Select Reference Models, Viewpoints, and Tools”, 4.3.3.2 “Conduct Formal Stakeholder Review”, 4.3.3.3 “Finalise the Business Architecture” and update the Architecture Definition Document”, 4.3.3.4. is to summarize Phase B. In this chapter the methodology to investigate and gather information about the relevant business processes is done. The Chapter 4.3.4 “Phase C: Information System Architecture” is with its subchapters, 4.3.4.1 “Select Reference Models, Viewpoints, and Tools” and 4.3.4.2 “Summary for Phase C” to develop the application and data architecture of the reference architecture. The methodology to design the technology architecture is described in Chapter 4.3.5 “Technology Architecture”, with its subchapters 4.3.5.1 “Select Reference Model, Viewpoints, and Tools”, 4.3.5.2 “Develop Target Technology Architecture Description” and 4.3.5.3 “Summary of Phase D”. Chapter 4.3.6 describes briefly the phases E to H and Chapter 4.3.7 is summarising the Methodology Chapter and 4.4 closes the whole Chapter 4.

In Chapter 5, Development of a Reference Architecture for Explainable AI in Corporate Planning, covers the development of a reference architecture for a trustworthy explainable AI system, namely Re-fish. After a short overview in 5.1 Introduction, the development of the reference architecture is described through by using a combination of ADD (Attributive Driven Design) and TOGAF ADM methodology in Chapter 5.2 Development of the Re_fish reference architecture. The main sub-chapters include preliminary discussion, purpose and scope, (Chapter 5.2.1). In Chapter 5.2.2 “Architectures of Knowledge Enabled Systems” current architectures of knowledge-based AI systems are investigated regards their explainability and architectural components. Chapter 5.2.3 “Gathering and synthesis of the Requirements” is an overview and summary of all the requirements gathered in the

previous chapters. The architecture of the Re_fish is presented in Chapters 5.2.4 to 5.2.7. Referencing chapter 4, first the business architecture is presented. In the following subchapters the application architecture and the technology architecture are presented. Chapter 5.2.7 presents the overall architecture of Re_fish, which summarises all the previous points of view. The lifecycle management of an AI application and therefore also for Re_fish is presented in Chapter 5.2.8 followed by Chapter 5.2.9 briefly describing the opportunities and solutions etc. (referencing Chapter 4.2.6) In Chapter 5.3, the evaluation of the reference architecture is conducted and Chapter 5.4 summarises and discusses the feedback and documents possible gaps- to be changed in a next iteration of the design of Re_fish. Chapter 5.5 summarises the design and development of the Re_fish Universal Explainer reference architecture.

Chapter 6 concludes the thesis and provides an overall summary of the work and a review of the findings, including prospects and recommendations for further research.

"The kitchen was especially difficult to navigate because so many of its elements would change their relationships to one another moment by moment. […], Melania Housekeeper would constantly move items around, obliging me to start afresh in my learning."
(Ishiguro, Kazuo (2021). Klara and the Sun. Chapter 2)

2 Planning in the Process Industry

2.1 Introduction

There is no standard definition of *process industry*. It serves as an umbrella term for several industries which are crucial in commercially transforming raw materials into finished products. The process industries differ from others in terms of manufacturing characteristics; they use process manufacturing in batches instead of discrete manufacturing. Specific industries include chemical, pharmaceutical, food, and petrochemical production. The processes involved in this transformation typically require physical and chemical changes and, in some cases, biochemical changes. The processes are engineered and take place within process plants. Most of the products have well-defined specifications (recipes).

Process industries can be usefully classified based on the type of feedstock or product involved, for example, petroleum refining, mineral processing, chemical processing, fertilisers, food, and pharmaceuticals (Brennan, 2020). In this work, focus is on chemicals and pharmaceuticals, particularly in the fields of pharmacy and life sciences. Planning and decision-making are significant tasks for a decision-maker in a process industry company. A planning problem typically arises from a gap between the desired state and the current or future state without any intervention. It is important to carefully evaluate all options and make the best decision possible to bridge that gap and achieve the desired outcome. Planning, on strategic, tactical, and operational levels, is essential within the process industry companies. It was Shell in the 1970s that first used scenario planning, a technique within strategic planning (Wack, 1985). The other important area is S&OP – sales and operations planning, as an overarching characteristic of the process industry is its high integration into production networks and connections via complex supply chain networks. When S&OP planning is also connected to financial planning with improved alignment between supply and demand, it is called Integrated Business Planning.

In the next chapter, the process industry is characterised, especially its relevance in Europe (Chapter 2.2. The Process Industry). In Chapter 2.2 the specifics of the process industry

will be discussed (Chapter 2.2.1), key trends (Chapter 2.2.2) and challenges (Chapter 2.2.3) and how AI can support the process industry (Chapter 2.2.4). Chapter 2.3, Planning and Decision Making in the Process Industry, describes the relevance and importance of planning and decision-making within the management process and as a decision problem, in general. However, the focus is on the process industry. In the subchapters 2.3.1. Scenario Planning, the strategic planning approach or technique of scenario planning will be described, in relation to the process industry. Chapter 2.3.2 investigates integrated business planning as another approach covering the range from strategic to operational planning in the process industry. Chapter 2.3.3 will introduce decision making an explanation in planning in the process industry. The stakeholders of planning in the process industry will be investigated on in chapter 2.3.4. Chapter 2.4 is about the information systems to support the planning and decision making. Chapter 2.5 with its subchapters 2.5.1 to 2.5.3 investigates on classical decision support systems as well as on business analytics and reporting. The findings will be summarised in chapter 2.6.

2.2 The Process Industry

The environment and ecosystem provide manifold resources, which if processed, become valuable products for society. Among these are gasoline, metals, polymers (plastics), pharmaceuticals, and food. However, naturally occurring substances require refinement or processing. Since most of these raw materials cannot be found in the same place as processing plants, it is necessary to provide these substances in sufficient quantities, qualities, and at the right time within networks of processing and transport in the right place. This network of process-oriented companies (see below) has significance in the world economy (Brennan, 2020). After briefly describing the macro- and microeconomic perspectives in Chapters 1.1 and 1.2, the following chapters will - highlight the special features of the process industry (2.2.1) - point out the key trends in the process industry and then go into the special features of the process industry. The subchapter concludes with a presentation of AI support in the process industry.

2.2.1 The Specifics of the Process Industry

Companies in the process industry play a key role in transforming raw materials into finished consumables or intermediate products at a commercial scale. To transform the raw materials into products, consumables, or intermediate products for production in a network of downstream companies, within the framework of chemical, physical, or biochemical processes, in addition to specifications (e.g., as recipes), highly complex technical plants, raw materials and utilities are used, which lead to high capital and operating costs. To carry out the processes, a large amount of electricity, fuels, water, cooling, and heating media, etc., is highly dependent on water and energy resources (Brennan, 2020; King, 2019). According to Murzin (2022), the chemical industry converts raw materials (oil, natural gas, air, water, metals and minerals) into more than 70,000 different products. These include products such as fuels, polymers and plastics, basic chemicals, consumer goods and chemicals, products for agriculture, manufacturing, construction, pulp and paper, life sciences, textiles, and other industries (Murzin, 2022).

Since many of the raw materials and intermediate products are highly toxic, to humans and nature, as they are flammable or explosive, and certain production processes may pose high risks to the environment, the process industry is subject to strict regulations during transport, production, and use of the products. With regards to the production of pharmaceuticals, strict quality requirements are placed on the industry. In addition, during production, in addition to the actual marketable products, a large amount of waste may be generated, which in turn can be toxic, flammable, or explosive and must therefore be disposed of accordingly. After using or consuming the products, those must also be disposed of. Thus, the process industry has a great interest in finding solutions for sustainability, be they in the provision of raw materials -- such as production, waste disposal, and also the disposal of products after use, being recycled, as environmentally friendly as possible and to continually optimise and reduce the ecological/environmental impact. The main idea behind this is the circular economy; process industry plants must therefore be ecologically as well as economically sustainable. The benefits of the sales revenue of the products must exceed operating costs and provide an adequate return on capital investment. High cost and a long time of investment in the life sciences industry, expensive to build plants, expensive to operate plants – due to high costs of raw material, personnel, utilities (water, power etc.), insurance, and management-related staff, must be competitive with other companies in the market (Brennan, 2020; Murzin, 2022).

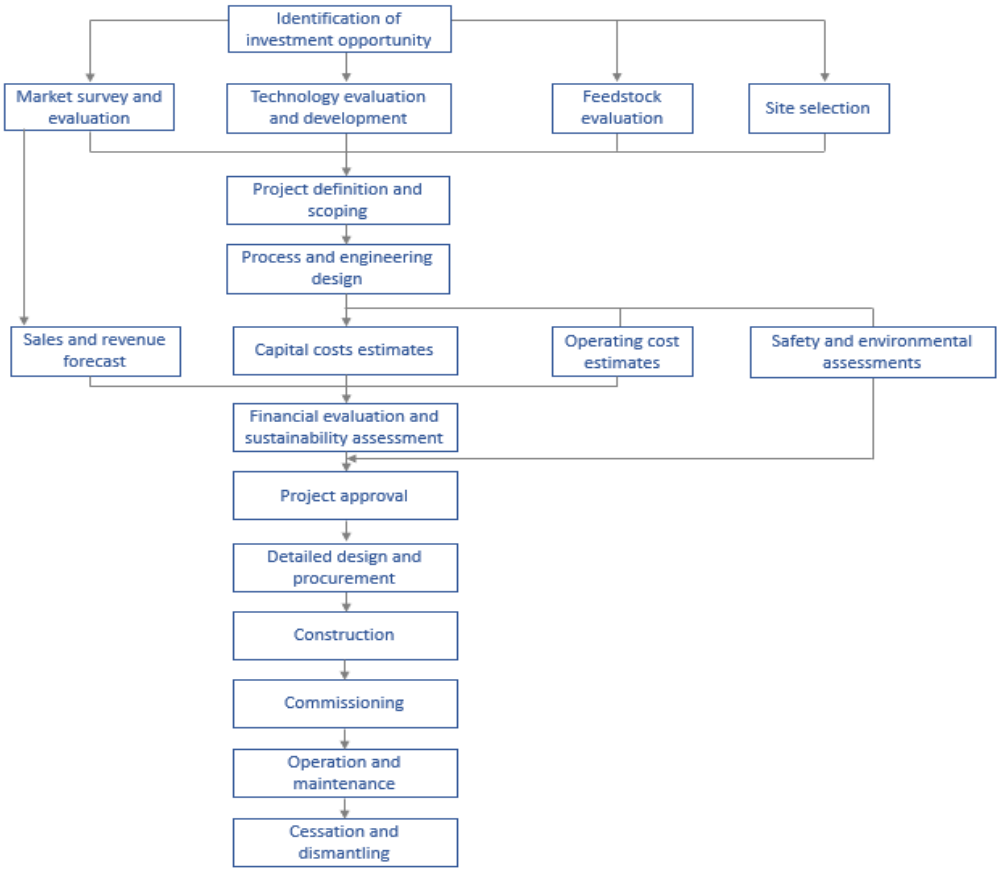

Figure 8: Anatomy of a process industry project – Brennan (2020)

Brennan (2020) describes the typical anatomy of a process industry project (s. figure 8). It includes the identification of an investment opportunity, the evaluation of the markets, evaluation and development of technology, the production capacity, the extent of integration with other manufacturing plants, the storage and transport of raw materials and products, the supply of utilities, and personnel requirements for design, construction, and operation. This will then become the basis for doing more detailed market forecasts, and the process and engineering design can be done to allow capital and operating cost estimates to be made, as well as safety and environmental appraisals, which lead to the financial evaluation and sustainability assessment. The subsequent steps in this framework are project approval, detailed design, construction of the plant, and commissioning. Operation and maintenance follow, and then, once the plant is no longer economically successful, its cessation and dismantling take place (Brennan, 2020). The steps in the framework are done in a logical sequence, but can be iterative, e.g., if the approval to build the plant is not given and its design or another aspect has to be re-evaluated.

Market forecasting is very important to the process industry but also includes imponderables and uncertainties; such issues may have various causes. Brennan (2020) mentions, for example:

- business cycle fluctuations which can be caused by various influences.
- changes in process technology or product development
- changes in industry structure
- changes in international participation in manufacturing
- changes in the balance of supply and demand
- changes to international trade arrangements
- changes in environmental drivers, including government regulation, for example on global warming impacts, materials recycling,
- changes in environmental drivers, including government regulation, for example on global warming impacts, materials recycling.

In addition to the above, macroeconomic parameters, such as growth and competition, are important business influences. Change of growth in a particular country in which a corporation is operating can influence its consumption patterns and capacity for investment in process plants and research and development. Forecasting is one of the key tasks within process industry companies, though it spans a longer timeframe and includes more uncertainties; this technique to improve strategic planning is described in Chapter 2.3.1 - Scenario Planning in the Process Industry. At a tactical level (see Chapter 2.3 - Planning and Decision Making in the Process Industry), there must be an integrated business plan, where the strategic plan is detailed out (given the strategic goals – strategic planning, derived from the strategic scenario analysis, like, e.g. a given ROI (s. below)) so that supply/demand and capacity are aligned and a common consensus plan is also developed, and this is the distinction and difference between sales & operations planning and integrated business planning, aligned with the financial plan and beyond between all functional departments of a company and top management.

For strategic planning, companies use the balanced scorecard to build a value tree and derive key performance indicators leading towards a common objective, e.g., increasing the shareholder value, meaning that the annual operating profit is greater than the total investment outlays. The annual operating profit is annual sales revenue, minus the annual operating costs. The total investment is fixed capital, plus working capital.

$$ROI = \frac{Annual\ operating\ profit}{Total\ investment} \quad \text{(f1)}$$

$$ROI = \frac{Annual\ sales\ revenue - Annual\ operating\ costs}{Fixed\ capital + working\ capital} \quad \text{(f2)}$$

Long term shareholder value will be relevant in chapter 2.3, especially in chapter 2.3.1 and chapters 2.3.2, 2.3.3.

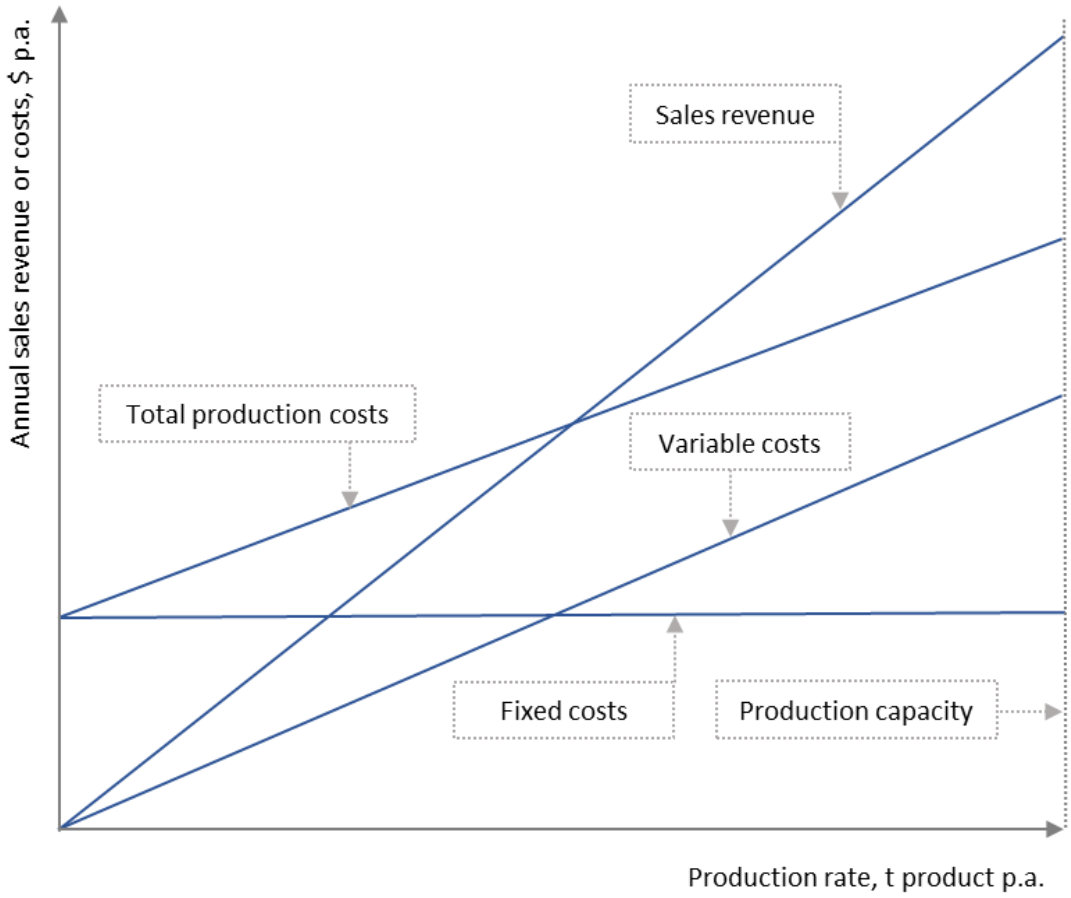

Figure 9: Dependence of sales revenue and production costs on production rate

Figure 9 shows that companies in the process industry have a high volume of fixed costs, as they are an asset-intensive industry. Therefore, the total production costs consist of the variable costs (varying with the production rate) plus the fixed costs (production-rate independent); this means that if the production capacity is not fully used, the full sales revenue potential is not reached and might even cause a negative operating profit.

The complete model is

$$C = \sum R_i r_i + \sum E_j g_j + \frac{(Mm)}{QU} + \frac{(kI)}{QU} \qquad \text{(f3)}$$

C = production cost (€/t product)
R = consumption of raw material I (t raw material/ t product)
r = unit cost of raw material I (€/t raw material)
E = consumption of utility j (e.g., MWh electricity/ t product)
G = unit cost of utility j (e.g., €/MWh)
M = number of employees/ t product
m = average annual cost per employee (€/ person) including payroll overheads
k = factor to account for a number of costs dependent on fixed capital
I = fixed capital investment (€)
P = annual production (t product)
Q = annual production capacity (t product)
U = capacity utilisation (P/Q)

Therefore, to increase the annual operating profit, for instance, it is necessary to increase the annual sales revenue and decrease variable costs (such as the cost of goods sold, SG& A expenses) and optimise fixed capital (reduce fixed capital or optimise capacity utilisation) and reduce working capital (s. f1, f2 and f3).

As mentioned above, no standard definition of the term process industry exists, and many different industries are found within the process industry; it can be seen as an umbrella term for other (sub-) industries. Thus, there is no uniform definition for categorising and assigning specific companies to the various categories of the process industry. One suitable way to allocate companies to different types is to use the statistical evaluation available in most countries. For example, based on the industry overview of the Federal Statistical Office in the Federal Republic of Germany, the industries of agriculture and forestry, fisheries, chemicals, petrochemicals, mining and quarrying of stone and earth, production of foodstuffs and fodder, and pharmaceuticals are categorised as being process industries (Statistisches Bundesamt, 2008). The most important industries (construction, automotive, electronics, and consumer goods) within the process industries are the chemical industry

and the life sciences industry. Providing raw materials for a wide range of products industries and being a significant member in highly integrated networks, the chemicals industry is critical to the global economy. It plays a vital role in developing new materials and technologies that enable sustainable development and protect the environment. Being responsible for developing and commercialising innovative therapies and medical devices (sometimes also differentiated into healthcare) that improve the quality of life and save lives, the life sciences industry is equally important to the global economy as an industry with substantial investments and research, thereby creating high-skilled jobs and fostering innovation.

As already pointed out, one overarching characteristic of the process industry is the high integration into production networks connected via supply chain networks. High regulatory requirements and high demands on quality management characterise the entire industry (SAP, 2009); as a result, they are dependent on resilient supply chains.

Product quality plays an essential role in these upstream relations with other companies, and due to this, quality, quality management, and sustainability are primary criteria.

The process industry is technologically demanding, relies heavily on innovation, and is extensively regulated (REACH, GMP, FDA)[14]. Legal regulations strongly influence it in the environmental sector and as already stated, in terms of the availability and price development of extensively used raw materials, consumed utilities, and significant capital and operating costs. Utilities include electricity, fuels, water, cooling, and heating media, and so forth; these depend heavily on water and energy resources (Brennan, 2020).

Finding 7: Process companies have some special economic features. These result from the production process. The industry is very heterogeneous, but in general this production process is not easy to stop and restart, for example. Production is extremely equipment-intensive and requires large investments. The impact on the environment is also relevant in terms of sustainability and climate protection. Production itself is less labour-dependent than discrete manufacturing. Companies in the research-based life sciences have a com-

[14] Regulation concerning the Registration, Evaluation, Authorisation and Restriction of Chemicals, GMP = Good Manufacturing Practices, FDA = US Food and Drug Administration

plex, extensive and extremely expensive research process that is subject to many regulations - AI could bring significant improvements here, on the one hand in economic terms, but also in terms of curing generally still incurable diseases.

Finding 8: Competition in the process industry sector is very high and has led to continued concentration over the last 30-40 years. Globally, there are currently only three countries (or groups of countries) that achieve significant sales volumes - these are the USA, the EU and, far ahead of the two aforementioned, China.

2.2.2 Key trends of the Process Industry

The German Chemical Industry Association and the consultancy Deloitte conducted a study on the topic of Chemistry 4.0 in 2017. They used the term Industry 4.0, which goes back to a research project of the German government and a resulting high-tech strategy (Kagermann et al., 2011) to strengthen Germany as an industrial location. The name 4.0 refers to the versioning of software systems and the implementation of the fourth industrial revolution through four key drivers - digital revolution or digitalisation, sustainability, climate protection and the closing of material cycles in order to enter into a closed-loop economy. The term "fourth industrial revolution" is immediately criticised on the grounds that the "fourth industrial revolution" is the same technology as the third industrial revolution and is, therefore, only a continuation or further development of this third revolution. Thus, much discussion surrounds the concept of a second phase of digitalisation. We are constantly striving for the next level of technological advancement, and this is no exception. It will be interesting to see what new developments and innovations will emerge in this next phase. It also seems somewhat presumptuous that an industrial revolution is being predicted rather than observed post-hoc. The expectations of this idea were high, and critics already claim that the implementation and realisation of Industry 4.0 have failed.

The aim of the study by the VCI together with Deloitte was to investigate which developments will influence the chemical and pharmaceutical business by 2030 and to derive the tasks for today in order to take advantage of the opportunities (Falter et al., 2017). The study identified digitalisation, sustainability, climate protection and the closing of material cycles as the main drivers. The raw materials of this fourth industrial revolution are now, with increasing importance, data (see above the use of data e.g., in the field of development

and the partnerships between Merck and SAP or Palantir), the recycling of carbon-containing waste, the use of hydrogen from renewable energies in combination with CO2 in the production of basic chemicals. In the area of research, there is decentralisation and the use of large data, as well as joint development with customers. Corporate structures are changing towards more flexible cooperation within the framework of economic networks. Digital business models will emerge and there will be further consolidation. Products are developing in such a way that the chemical industry, for example, is becoming a provider of comprehensive and sustainable solutions, both in terms of the customer and the environment - the spectrum of value creation is expanding. According to Falter et al. (2017), sustainability (in terms of ecology, economy, and social aspects) is becoming a comprehensive guiding principle and concept for the future.

The study identified 30 trends. These were in turn divided into four quadrants based on the categorisation - Incremental vs. Disruptive and Societal-Political and Entrepreneurial-Economic driven. The trends were also divided into small, medium, and large impact. This resulted in 13 trends with a medium impact, 10 trends with a small impact and 7 trends with a large impact. The trends with a large impact are lightweight construction in cars (socially incremental), electromobility, genome editing in medical applications and genome editing as precision breeding (socially disruptive), and finally personalised medicine, industrial biotechnology, and digitalisation of agriculture (entrepreneurial disruptive). It can be seen that a striking number of trends and innovations are taking place in the disruptive area, and this on the basis of advancing digitalisation (and through the use of artificial intelligence, see Chapter 2.2.4) However, these trends pose major challenges for the industry, as their disruptive nature will have a direct impact on process-technologies, product portfolios and thus the entire value creation (Falter, 2017). In the area of process technologies, chemistry can contribute to the coupling of the energy and industrial sectors by, for example, using the overproduction of electricity to produce synthetic raw materials, e.g., synthetic fuel, which can then in turn be used as energy storage when sustainable energy production cannot provide enough base load (see e.g. Chapters 2.2.4 and 5.2.3) and thus significantly reduce the demand for fossil raw materials. The decreasing demand for e.g., fuel-resistant plastics in automotive construction and oil and fuel additives will be replaced by an increasing demand for electric motors, battery technology and lightweight construction materials. It is also possible, according to Falter et al. (2017), that entire value creation structures will change, which in turn will have an influence on customer relationships or may involve completely different business models.

Digitalisation or digital transformation as a sub-area of Industry 4.0 is also seen as a focus in the process industry, with 50% of medium-sized chemical companies planning to invest in digitalisation in the coming years. 30% of SMEs already generate 4% of their turnover with digital business models, i.e., new value-added structures that offer customers a combination of services and products, often through a network of suppliers. A further 40% are planning to introduce such new business models. In addition to these new digital business models, digitisation offers two further categories for growth, innovation, and efficiency gains. For example, through improved transparency and digital processes, as well as, through extensive collection and analysis of process data, along the production process, it is expected that, despite the already traditionally extremely high level of efficiency in the industry, there will be further increases in efficiency due to the type and manner of production, and also through the further automation of processes, for example through the use of AI technology. Furthermore, the collection of internal and external data and their analysis should serve to gain a better understanding of the behaviour of customers and competitors in the markets and thus, for example, become an active company within the framework of forward-looking planning (see 2.3.1 and 2.3.2) and not just act reactively. To this end, the industry is pushing ahead with further developments in the area of predictive maintenance and networked logistics (see e.g., Chapter 2.2.4) as well as in virtual reality applications.

Another core aspect of Industry 4.0 or Chemistry 4.0 is the role of the process industry in the circular sustainable economy. According to the study (Falter, 2027), the process industry must expand its core business to include new business models, such as chemical leasing. A rethink must also take place, in which the focus is no longer on volumes, but on application benefits and value-based pricing. The process industry also has the task of conserving resources by increasing resource efficiency at all stages of the value chain. The service life of products is extended and their resource consumption in use is reduced, and the closure of the cycles is achieved, which leads to ensuring more efficient use of the remaining raw materials through reuse, recycling, energy recovery or biodegradation, as well as in general.

Seven levers can be distinguished with regard to the optimisation of the material cycle. These are re-design, in the sense of a data-supported re-design of the products on the basis of, for example, the evaluation of product usage data, resource-efficient production opti-

misation through the above-mentioned insights and corresponding adjustment of the production processes, modular production or even the use of robots to further automate the material cycle. insights and corresponding adjustment of production processes, modular production or even the use of robots to further increase automation (in the process industry, e.g., in the chemical sector, the share of labour is lower, whereas the utilisation of production facilities (capital) is significantly higher). Automation, and thus possibly replacement or at least augmentation of labour, has a lower effect in this respect (see Chapters 1.1 and 1.2 above). Another lever is take-back, for example, the use of new business models in which the use of the products is recorded in real time by the customer in order to determine usage data on the one hand, but also the correct time for replacement. Recycling, energy recovery and waste disposal are additional levers (Falter et al., 2017).
Following these findings, the study provides a catalogue of twelve recommendations for action that can be clustered into three categories: 1. align strategy, such as anchoring digital and circular economy in the strategy, 2. build resources, in terms of corporate structure and competencies, transform culture and 3. seize opportunities by building and expanding economic networks and using them as platforms, etc.

In summary, it can be said that the process industry is driving the digital transformation, which includes optimising processes in order to be able to carry out further optimisations despite the already very high level of efficiency. For this, the corresponding data must be collected and evaluated in real time. The data in general is seen as a production factor and should be used in the future to optimise relationships with customers but also with the competition. Internally, the aim is to use the data for research and development purposes and to drive forward automation as well. Traditionally, due to manufacturing, the labour factor is used less in companies in the process industry than in discrete manufacturing. Instead, it is an equipment-intensive production. This suggests that automation will have less of an impact in the area of manufacturing (see Chapters 1.1 and 1.2). Closing material cycles is another goal in terms of sustainability and climate protection. In Chapter 2.2.3, it can be seen that the process industry in Europe is already on the right track in this respect, as energy consumption has fallen steadily despite the expansion of production, as have greenhouse gas emissions.

Finding 9: Key trends in the process industry are digitalisation, sustainability, including in complex and networked supply chains, and further process optimisation. This industry is

highly automated due to its production process, but experts suspect that the available data is not yet being used extensively for process optimisation.

2.2.3 Challenges of the Process Industry

The process industry, and in particular the chemical and pharmaceutical sectors, are subject to high regulatory requirements. This is also historically justified, for example, if one looks at the history of the so-called "Schweinfurter Grün", a wall paint that was very frequently used in the 18th century, or the Contergan[15] scandal in the 1960s, in which more than 10,000 children were born with deformities. In the chemical industry, for example, there is the European chemicals regulation REACH ("Registration, Evaluation, Authorisation and Restriction of Chemicals"), which came into force on 1 July 2007 (https://echa.europa.eu/de/regulations/reach/understanding-reach). Additionally, there are different rules for each country on how much of a certain ingredient may or may not be contained in a product. This is particularly important in international trade relations. The EU Commission published its "Chemicals Strategy for Sustainability" in October 2020. Moreover, there is a plethora of other regulations (CLP), Biocidal Products Regulation, Prior Informed Consent (PIC) Regulation, Chemical Agents Directive (CAD) and the Carcinogens, Mutagens or Toxic to Reproduction Directive (CMRD), POPs Regulation, Waste Framework Directive, Drinking Water Directive, etc., ECHA (2023)), compliance with which must be fully recorded and proven from the acquisition of raw materials through production to the consumer. The same applies to the pharmaceutical industry, which also has to prove complete documentation and certification of the production processes (GMP) and document all substances and their quantities, etc. In Germany, the pharmaceutical industry is also subject to the German Medicines Act (AMG), the German Act on Advertising in the Field of Medicine (HWG), the German Ordinance on the Application of Good Manufacturing Practice (GMOP), in den USA die FDA (Food and Drug Administration) etc. Together with the high costs of such verification and certification, the pharmaceutical industry has a number of other special features and challenges, for example, a distinction is made between research-based manufacturers and generic manufacturers. The research-based companies are exposed to high investments in the development of a new drug - for example, the development of a drug takes approx. 15 years - at approx. 450 - 800 m€ and costs. This includes 4

[15] https://de.wikipedia.org/wiki/Contergan-Skandal, accessed 18.06.2023

to 6.5 years for research, drug discovery and preclinical phase- 58% of the costs are incurred in phase I. The success rate for new active substances, from 10,000 active substances in screening, 5,000 to 6,000 are further assessed. In the clinical phase, 5 are still in trials at the beginning, and these clinical trials last between 6 and 7 years and go through a total of three phases. In the end, a maximum of 1-2 active substances are left, of which 1 is approved, which in turn can take up to 2.5 years. After approval, when the active ingredient is launched on the market, it must generate a surplus in addition to the development costs during the remaining term of the patent. So, if it were possible to make significant progress in the area of research through the use of AI, as described in Chapter 1, this would completely revolutionise the development of new drugs. But even so, the search for possibilities for automation is a high priority in this area (Breitenbach & Fischer, 2020).
The market, especially in the pharmaceutical sector, is highly competitive and has been characterised by major waves of consolidation since around the 1970s. In 2008 Merck KGaA acquired Serono, Europe's largest company specialising solely in biotech, for € 10.6 billion, and in 2009 Roche acquired further shares in the American company Genentech for € 46.8 billion. Also, this year, Pfizer acquires the biotech specialist Wyeth for 68 billion US$. Together, the two companies now have 130,000 employees and a total annual turnover of 71 billion US$. Still, in 2009, the American Merck & Co (which operates independently of Merck KGaA) acquires its competitor Schering-Plough etc., for US$ 41 billion. In 2018, the Japanese Takeda Group acquires its rival Shire in its fifth attempt for US$ 64 billion. In addition to these impressive figures, which also reflect the competition in the industry, there is also fierce competition at the global level between the countries USA, China, Europe, etc. This is illustrated below on the basis of a market study.

The importance of the process industry in Germany can be derived when looking at the growing revenue (s. figure 10).

Figure 10: Relevance (sales in €) of the process industry –in Germany (Statista (2023).

Despite the years of financial crisis (2007-2010) and during 2015-2017 and 2019-2021, the overall revenue of the process industry grew continually from 107 m € in 1991 to 227 m € in 2021.

To understand the relevance of the process industry in the world, it is reasonable to look at the significance of the chemical industry within Europe and in relation to the rest of the world. In figure 11, it is shown that Europe was the second-biggest producer of chemicals in the world by 2021.

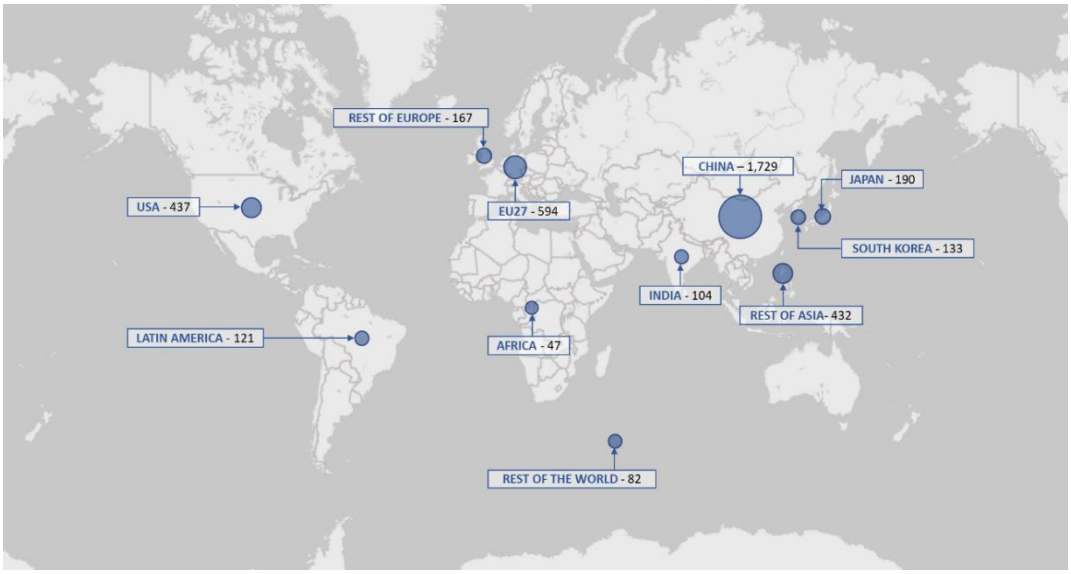

Figure 11: World chemical sales in 2021 (CEFIC (2023))

China dominated the market in 2021 by 1.729 billon €, while Europe was in second place with 594 billion €; the most prominent producers in the world, aside from Europe and China, are the USA (437 billion €), Japan (190 billion €), and South Korea (133 billion €).

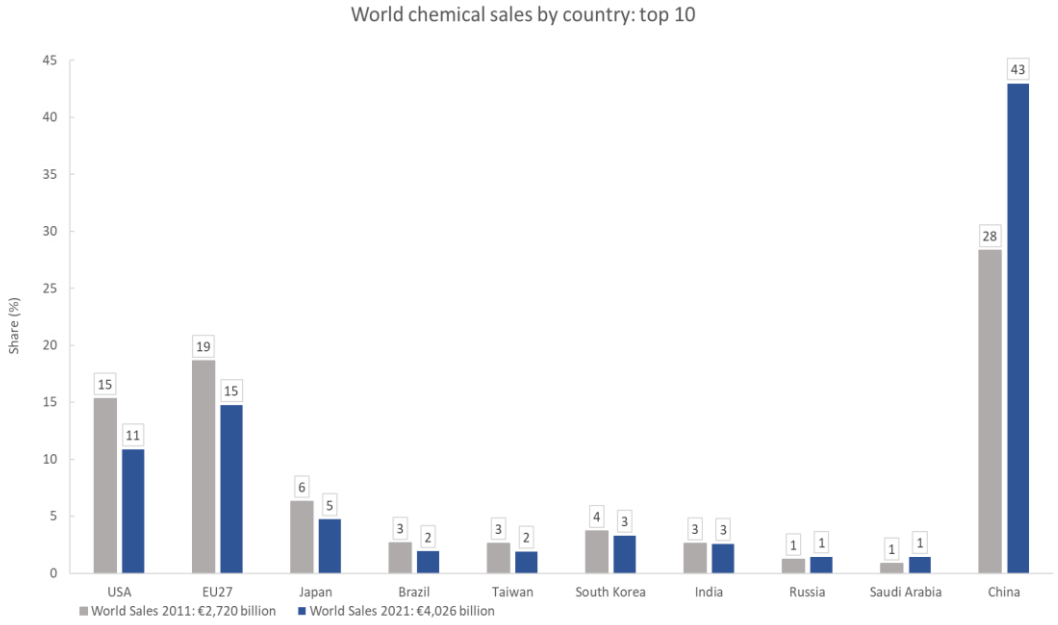

Figure 12: World market share of chemical sales, CEFIC (2023)

In figure 12, the world chemical sales market share is compared between 2011 and 2021 (CEFIC (2023)). It shows that the overall market share of EU27 in sales of chemicals dropped significantly between 2011 and 2021, from 19% to 15%. This is similar to figure 13, which shows that the world market share from 2001 (27%) dropped to 15% in 2021.

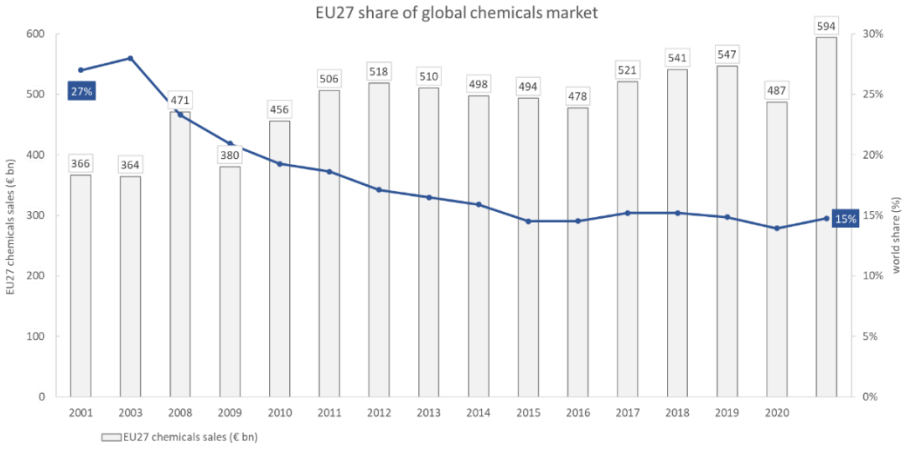

Figure 13: EU World market share of chemical sales from 2001 to 2021 (CEFIC (2023))

While the overall sum of the world market is growing (decreasing in 2001-2003, 2009, 2012-2016, and in 2020) from 366 billion € to 594 billion €, the EU27 market share is declining; EU27 is not keeping up with the market growth pace.

In figure 13, it is shown that the market share of Europe dropped significantly by 2021 compared to 2011 from 19% (2011) to 15% (2021), while China's market share increased dramatically from 28% in 2011 to 43% in 2021.

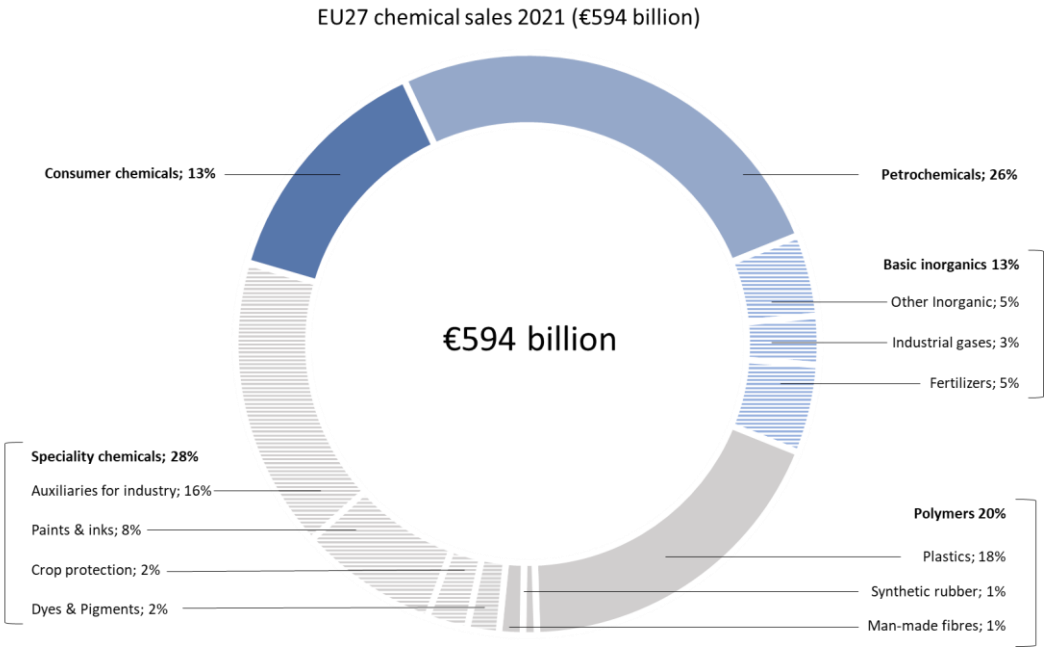

Figure 14: Distribution of sales by 2021 between the different categories of chemical products

In figure 14, it is shown how the chemical sales distribute among the different products of the chemical industry: Specialty chemicals (28%) and Petrochemicals (26%) have the highest volumes of sales.

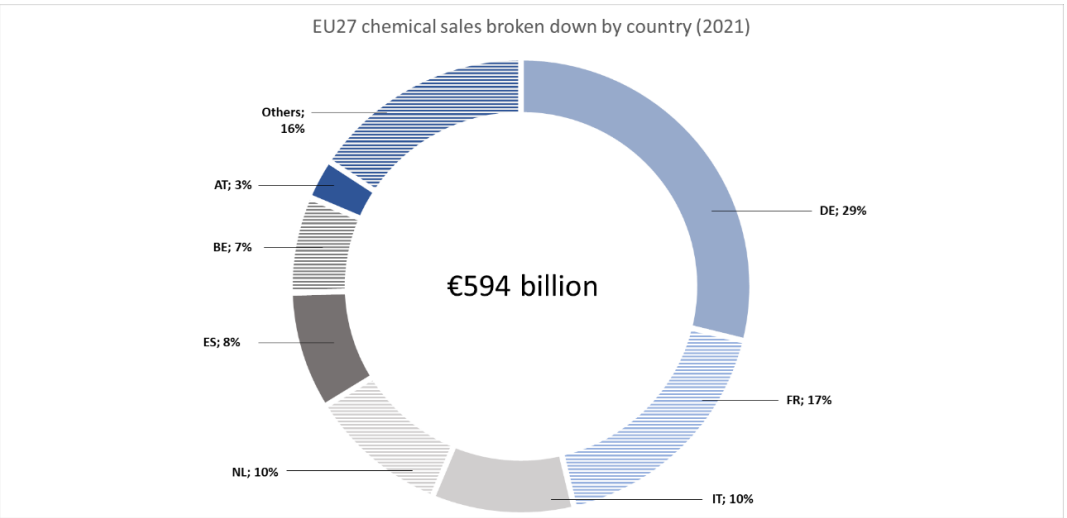

Figure 15: Chemical sales in 2021 broken down by country

Figure 15 shows that two-thirds of the production of chemicals is generated by four member states – Germany (29%), France (17%), Italy (10%), and the Netherlands (10%).

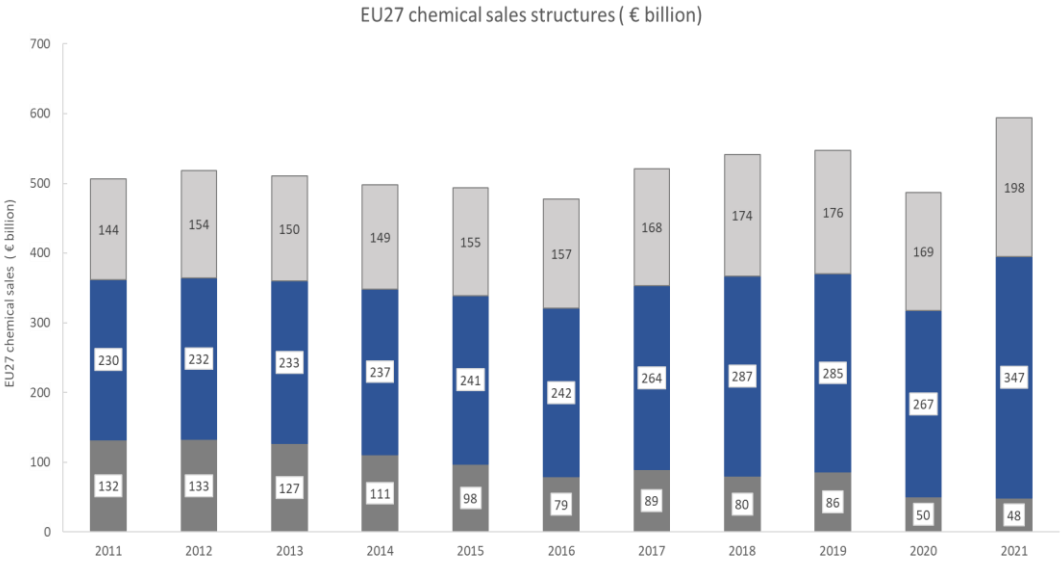

Figure 16: Structure of chemical sales from 2011 to 2021 in Europe

In figure 16, it is shown that home sales are decreasing from 132 to 48, while the intra-EU sales and the foreign sales are growing; however, intra-EU sales are growing at the highest rate (from 230 in 2011, to 347 in 2021)

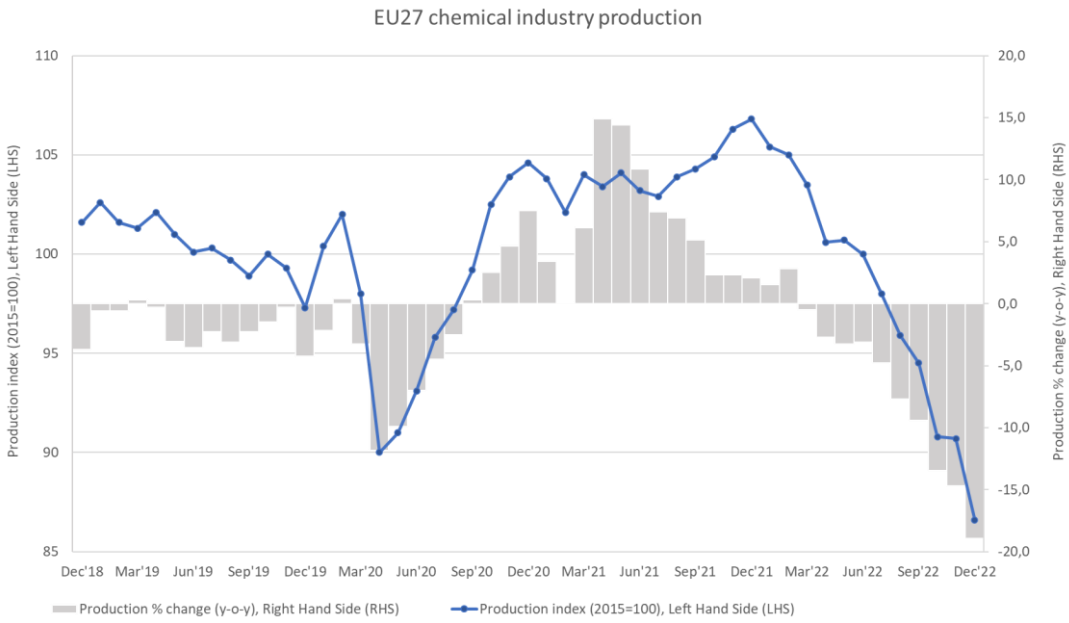

Figure 17: EU27 chemical industry production

Figure 17 depicts EU27 chemical industry production; the coloured line shows that since September 2021, the production index has been declining, and thus output.

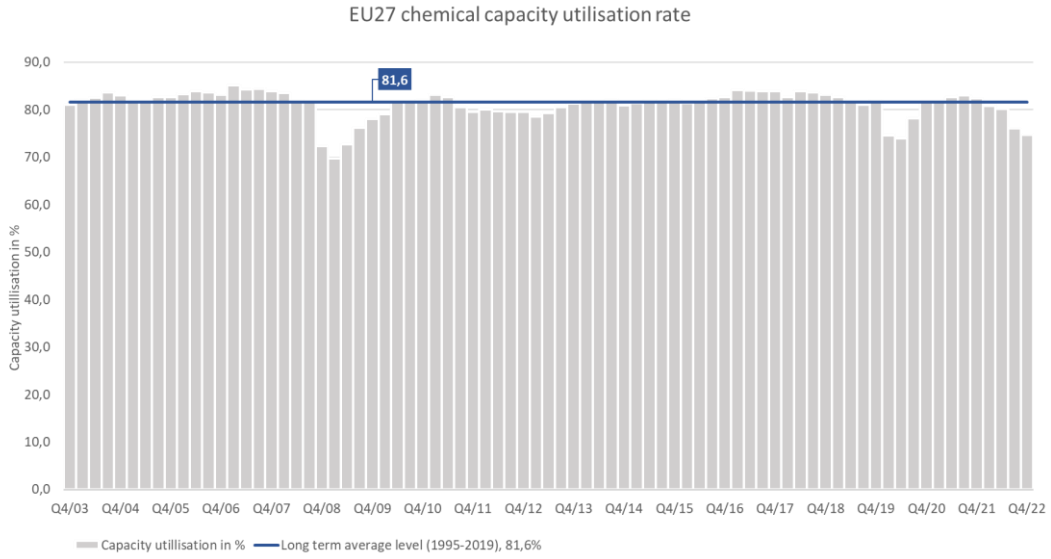

Figure 18: EU27 chemical capacity utilisation rate

In the figure above (s. figure 18), EU27 capacity utilisation below its long-term average indicates significant drops in Q4 2020 and from Q4 2021 until the end of the gathering of figures. This means that the European companies are working significantly below capacities (cf. capacity utilisation, mentioned above)

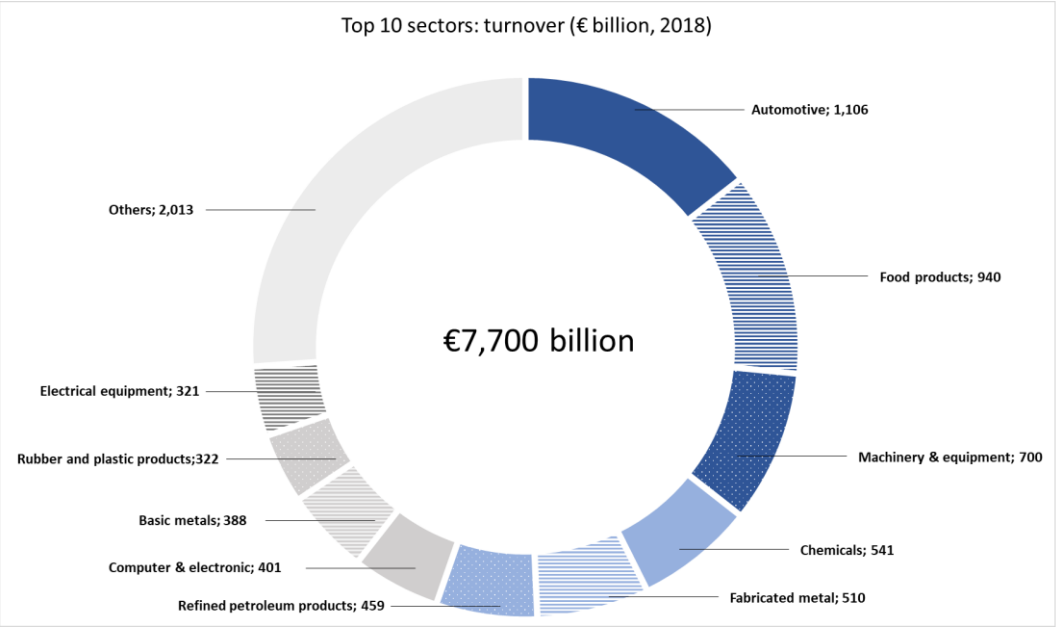

Figure 19: Top 10 sectors: turnover

In the figure 'Top 10 sectors: turnover' (figure 19), it can be seen that the chemical industry is the fourth biggest producer in the EU27 manufacturing sector, behind the automotive, food, and machinery & equipment categories.

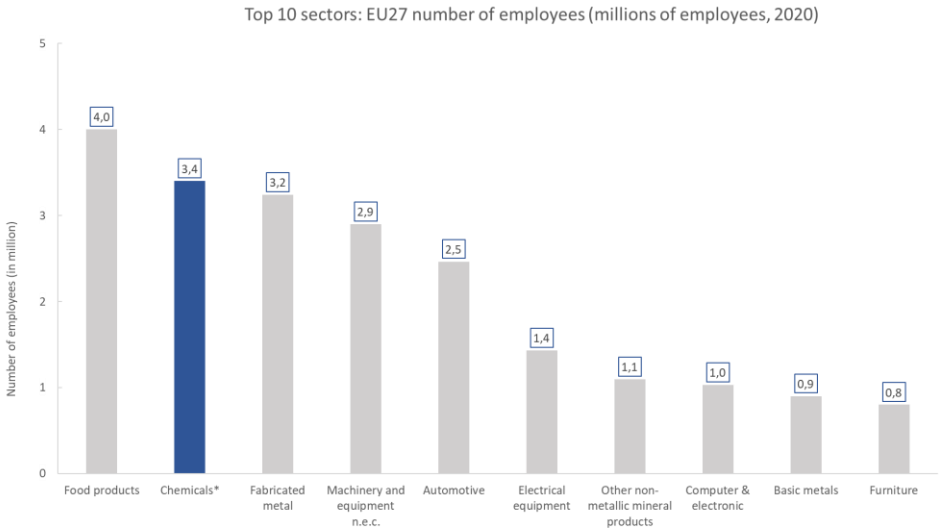

Figure 20: Top 10 sectors: EU27 numbers of employees

In figure 20, chemicals (including pharmaceuticals, rubber & plastic) are shown to have been the second-largest employer after food production in EU27 in 2020, with 3.4 million employees.

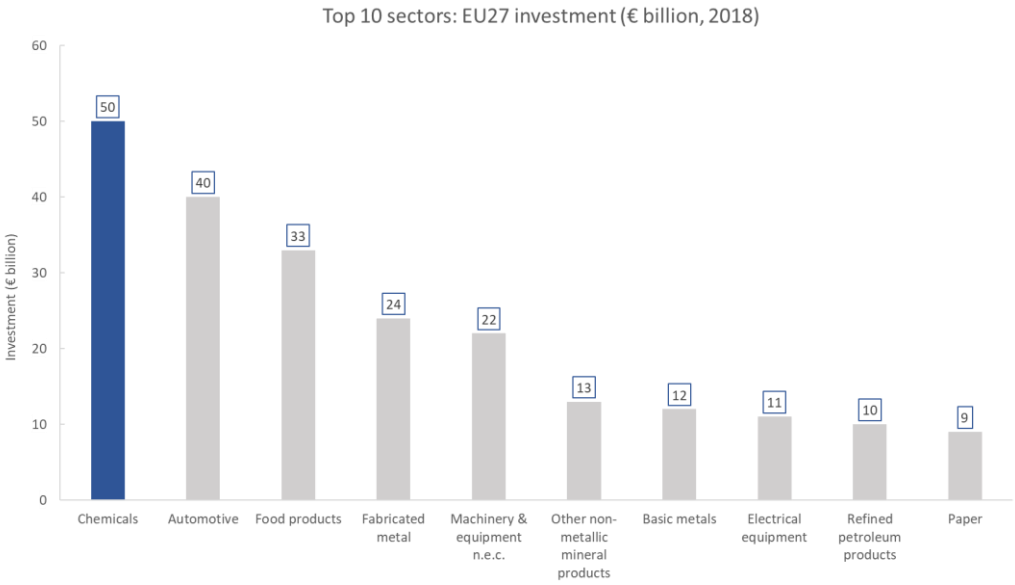

Figure 21: Top 10 sectors: EU27 investment

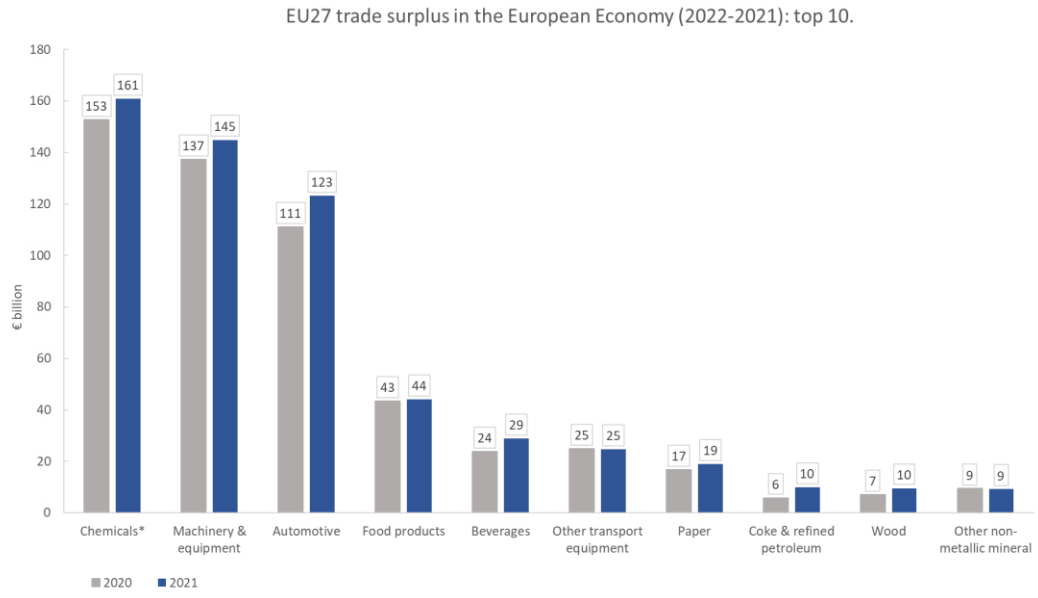

Figure 22: EU27 trade surplus in the European Economy (2020-2021): top 10

In figure 21, it can be seen that the chemical industry (including pharmaceuticals) provided the largest trade surplus in EU27 from 2020- 2021.

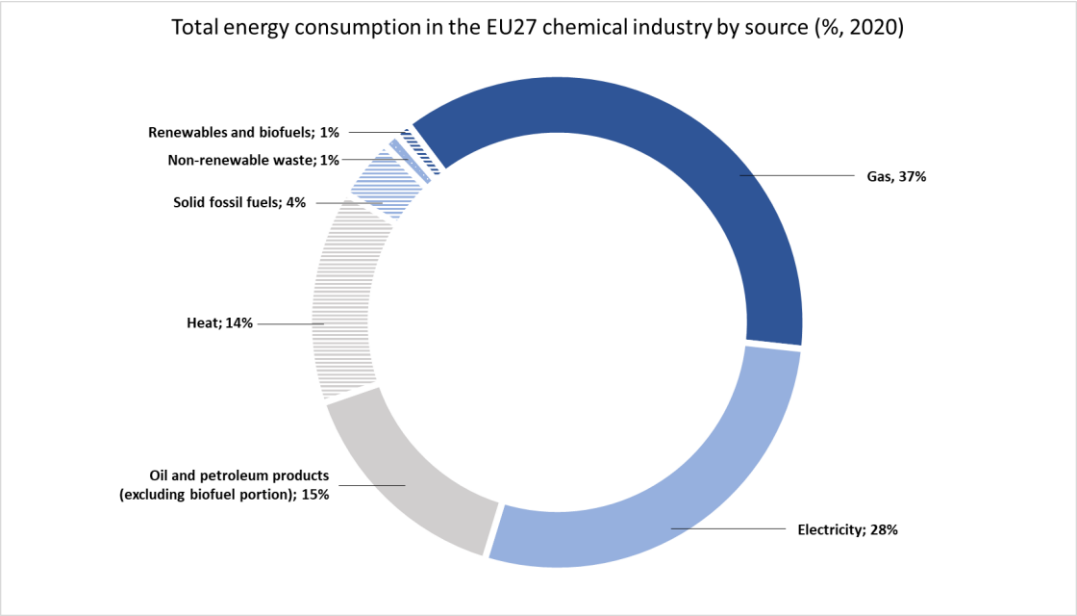

Figure 23: Total energy (589 terawatt hours, 2020) consumption in the EU27

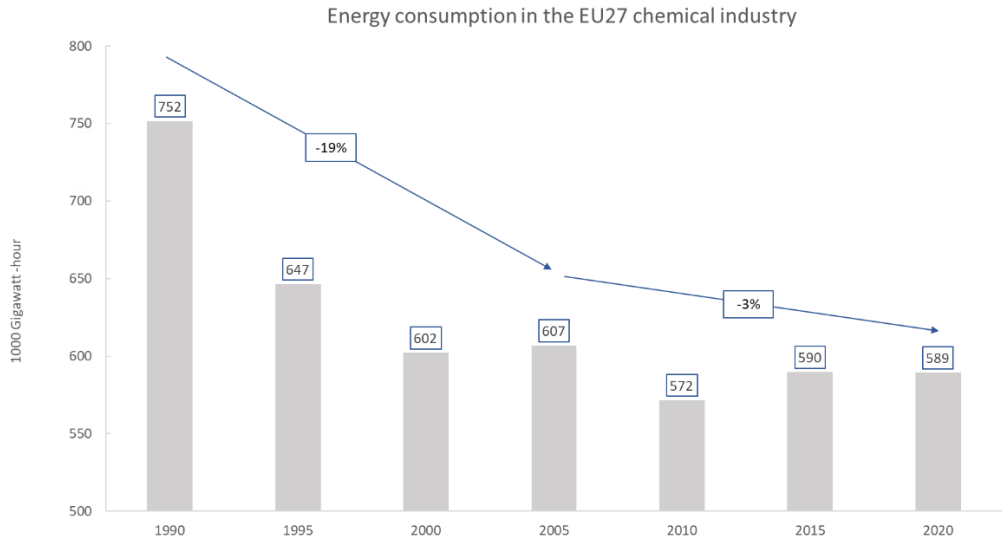

Figure 24: Energy consumption in the EU27 chemical industry

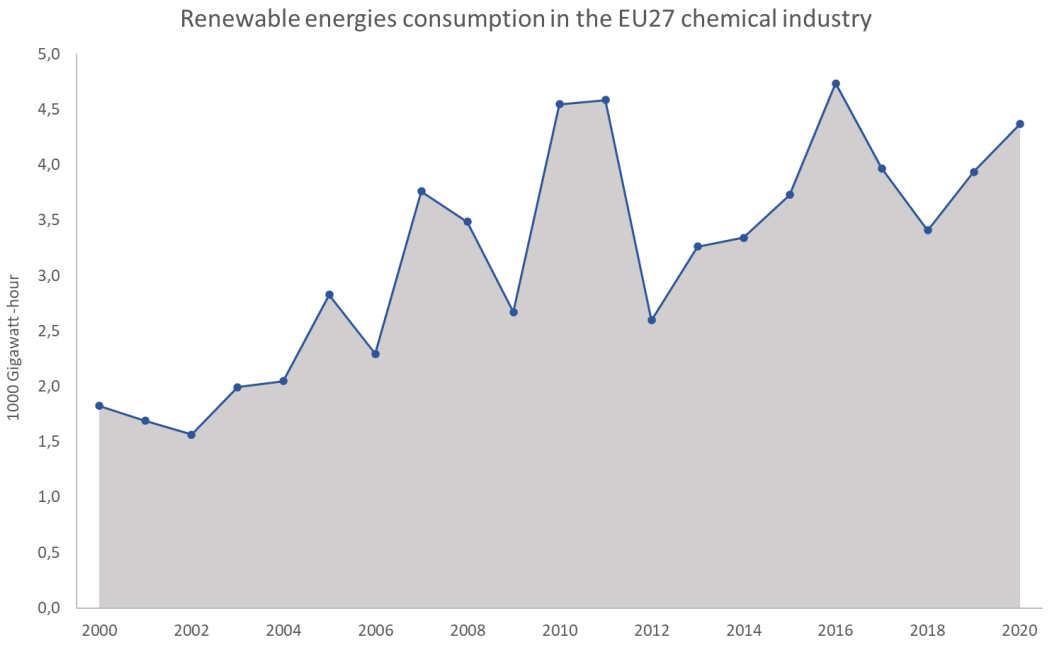

Figure 25: Renewable energy consumption in the EU27 chemical industry

Figures 22 to 25 show the total energy consumption and high dependency on gas as an energy source. Despite this, it can be seen that energy consumption declined from 1990 to 2020 – dropping from 752 tGW to 589 tGW. Between 2000 to 2020, the consumption of renewable energy doubled (figure 24 and figure 25).

Research & Investment

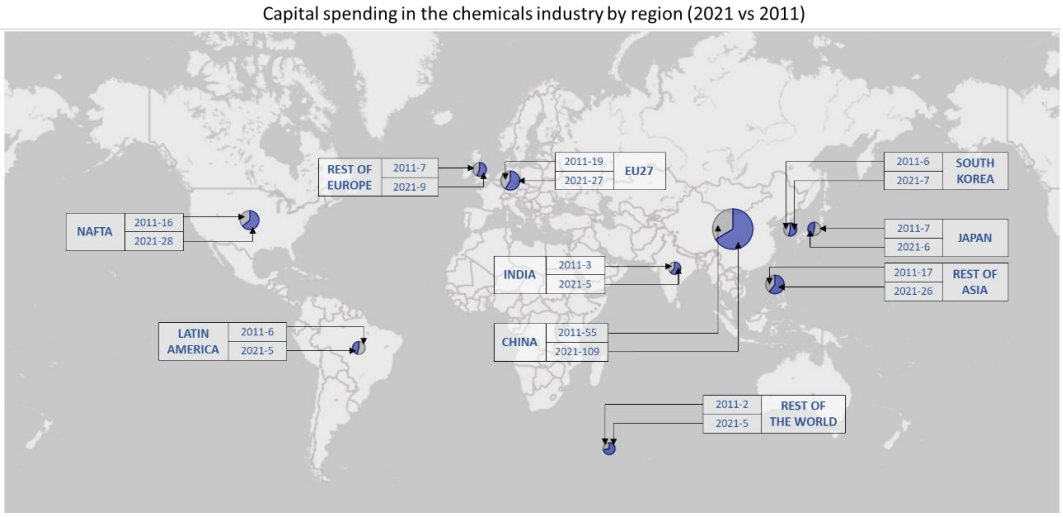

Figure 26: Capital spending in the chemical industry, by region (2011 vs. 2021)

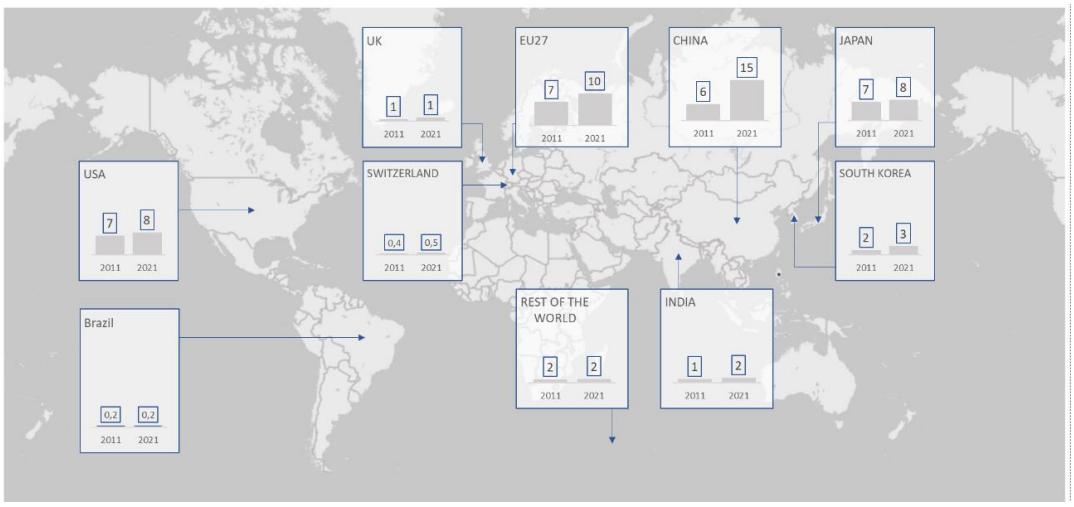

Figure 27: R&I spending in the chemical industry by region (2011 vs. 2021)

In figures 26 and 27, one may discern that China was leading in terms of capital investment in the chemicals industry in 2021, by 109 billion €. As may be derived from figure 27, the EU is the second largest R&I investor in the world, at 10 billion € in 2021 (China was leading at 15 billion €)

Sustainability:

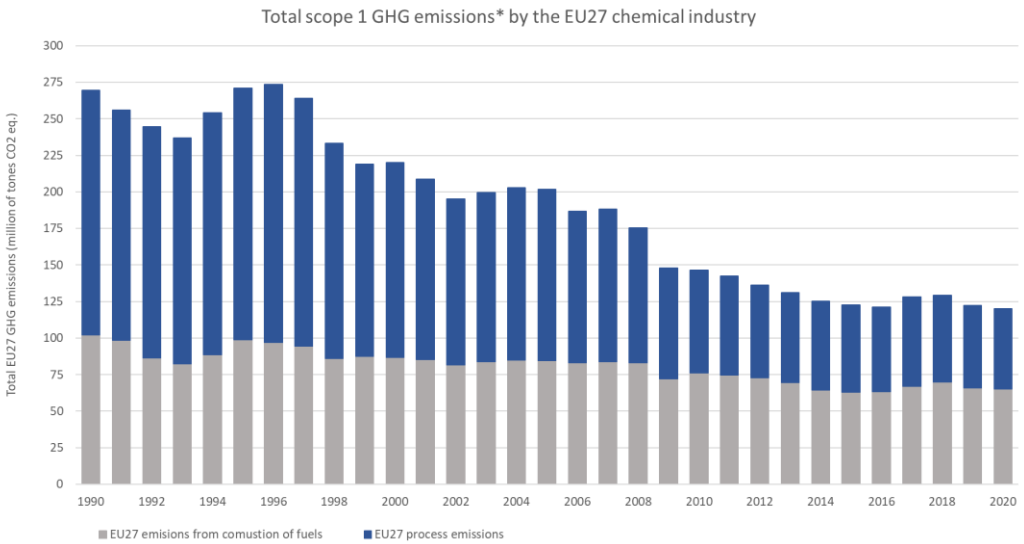

Figure 28: Total scope 1 GHG emissions by the EU27 chemical industry

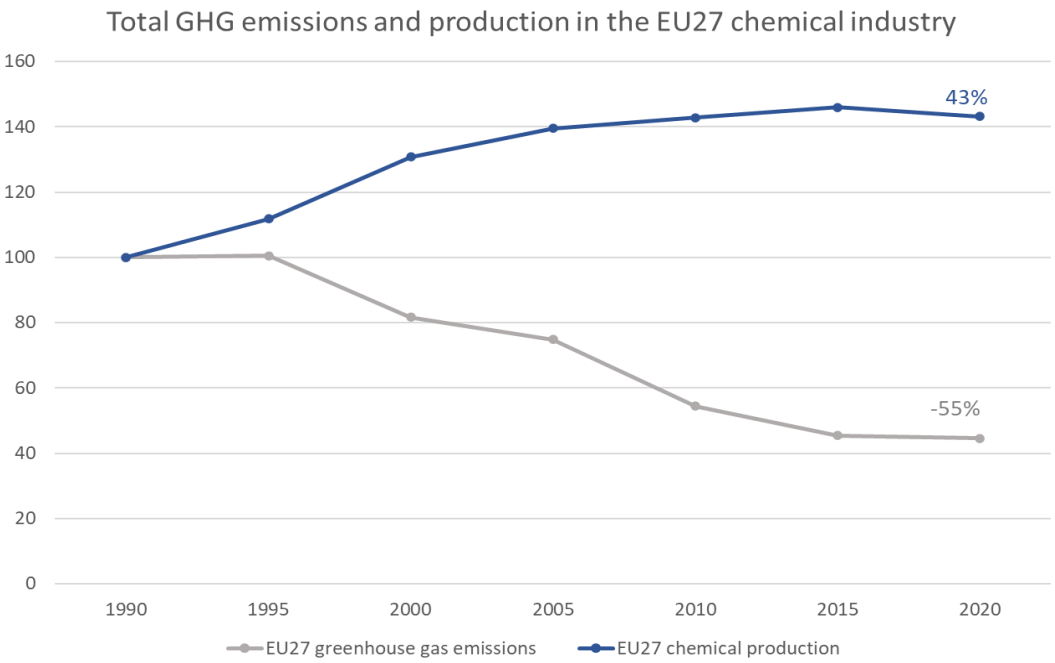

Figure 29: Total GHG emissions and production in the EU27 chemical industry

In figures 28 and 29, one can see that the overall emissions of the chemical industry declined by 55% from the 1990s to 2020, and the emissions of greenhouse gases and production decoupled, compared to a decline of 55% during the same period when production increased by 43%.

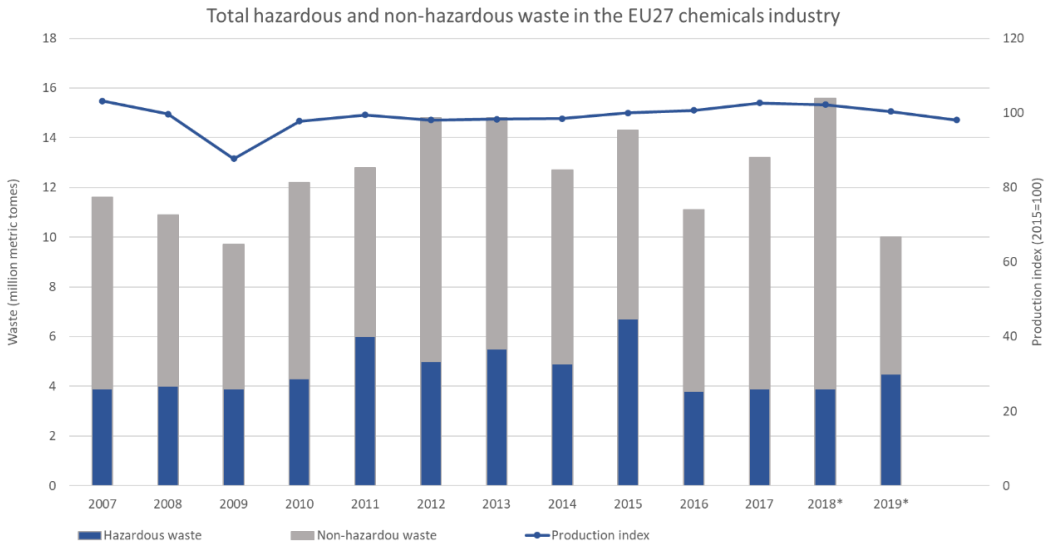

Figure 30: Total hazardous and non-hazardous waste in the EU27 chemical industry

Figure 30 shows that while the production index (2015 = 100) remained almost steady, the production of hazardous and non-hazardous waste declined significantly from 2007 to 2019.

Summary of the market analysis:

The market analysis of the process industry confirmed the image of a major industry in a highly complex dynamic environment. The importance, for example for Germany, can be seen in the growth of the industry since the 1990s. Globally, the process industry is number two in the EU27, ahead of the USA. China is, by far, the market leader and is growing at an enormous pace, so that the market share of the EU27 process industry continues to decline. Within the chemical industry, petrochemicals are (still) the leader, but this is likely to change in the coming years as environmental demands change. Within the EU27 countries, Germany is by far the largest producer in terms of sales, although the share of "domestic" sales has declined over time in favour of sales within the EU27 area. It is interesting to note that capacity utilisation in the chemical industry is below capacity. This is a significant challenge as the process industry is highly equipment intensive and, as shown in the model (see f3), relies on covering high- capacity costs. As the gap between utilisation and potential capacity widens, this can become very detrimental to some of the players in the process industry. The chemical industry is the fourth largest industry in the EU27 and the top two employer. In terms of investment, it is the top investor among the industries, not least because of the often-mentioned intensive use of facilities and also because of the high research expenditures, e.g., in the pharmaceuticals sector. In terms of environmental sustainability, the picture is encouraging. Despite the increase in production, energy consumption has fallen significantly, while at the same time the use of renewable energies has increased. Overall, GHG emissions have more than halved since the 1990s. This is the picture of an industry facing dynamic and accelerated change and the resulting uncertainty at the strategic level within companies. This underlines even more the importance of planning, and in particular scenario planning, which was used for the first time by a company in the process industry (see chapter 2.4). The process industry uses expensive and highly optimised and engineered equipment on a large scale. Therefore, predictive maintenance is a desirable tool for monitoring equipment and reducing costs. Integrated Business Planning - a successor or extension of S&OP - is a necessary methodology to plan efficiently and effectively in complex and highly interconnected supply chains.

Finding 10: Challenges in the process industry ergeben sich, wie bereits oben beschrieben, aus der hoc The challenges in the process industry arise from various aspects. On the one hand, there is the high level of regulation, the fierce competition, which is also reflected in the increased concentration that has taken place since the 1970s. The search for qualified

workers severely restricts the search for locations. There are also challenges posed by the enormous energy requirements and extremely high plant costs, which also have to be maintained over the long term. On the other hand, there are the short time intervals in which, for example in the pharmaceutical industry, sales can be made that cover the development costs.

2.2.4 AI support in the Process Industry

The pharmaceutical industry is currently experiencing a paradigm shift through data exchange and utilisation, as already described above in the context of Chemistry 4.0. Companies are, therefore, increasingly transforming themselves into data science companies. As Breitenbach and Fischer (2020) researched, there are two main factors contributing to the growing importance of data. Firstly, the molecular biological diagnostic methods have become more robust, and secondly, the ability to continuously track health data has significantly improved. This trend is known as the 4D principle, where diagnosis, drugs and devices are interconnected with data in the industry. It is believed that AI will undoubtedly decrease drug development timelines and minimise expenses. When it comes to bioinformatics, data regarding genomes get stored and analysed. On the other hand, computational chemistry creates molecular models that undergo simulations and analysis. The LIMS (Laboratory Information System), in turn, stores all the relevant data, e.g., to prove the above-mentioned GMP. In addition, incorporating it into clinical phases can greatly enhance the process and automate it effectively.

The Five Use Cases (s. e.g., Davenport & Miller, 2022 for more use cases) are briefly presented below as examples to illustrate the possible applications of AI in the process industry.

Use Case I: AI supported image analysis of histological tissue sections

An AI-powered tool can assist in analysing histological tissue sections, particularly in clinical studies evaluating the effectiveness of drugs. (s. above)

Histological imaging involves staining tissue samples taken from patients suspected of having cancer with dyes or dye-labelled antibodies (Kraus et al. 2022). When assessing

tissue samples from patients suspected to have cancer, a pathologist is responsible for examination. However, this process is predominantly done manually. This use case could be applicable in the clinical studies mentioned earlier, specifically in testing the effectiveness of a medication.

AI could help here, in an augmented approach (see Chapters 1.1 and 1.2), by using a model to examine the images for conspicuous patterns. By utilizing this approach, it is possible to not only expedite the process, but also enhance it. The model can acquire knowledge from all available images and patterns, resulting in a substantial improvement of the doctor's abilities. Furthermore, this results in increased knowledge for the professionals in the field (Nagpal et al. 2018). Although the model provides suitable suggestions and decisions, the presence of human input can enhance its recommendations by incorporating personal experiences. This may result in a distinct diagnosis from that of the model, which can then be fed back into the learning pool and potentially used to retrain the model.

By implementing this model, the possibility of missing important irregularities can be greatly decreased. Kraus et al. (2022) see the use of CNN (Convolutional Neural Networks) as possible models here. The approach would be classified as supervised learning since the model would have to learn from images representing patterns that have the respective label cancer = 1 or cancer = 0. Because of the significant impact on the patient's life, the analysis made by the model requires an explanation, e.g., to prevent a misdiagnosis (s. table 2).

Criteria	Description	Requirements/ Evaluation/ Comment
System Type	Decision Support	Augmentation- Human - Machine Case
Type	Anomality detection	
Economic categorization	Augmentation case- enhance human capabilities	Non repitive work- highly skilled
Business Impact	Medium - high	Partially automatise the process- human in the loop
Societal/ Government	Very high	Improve diagnosis quality significant
Criticality/ Impact	Very high	
Data Types	Image data (2- and 3D), high resolution	
Typical AI Model used	Neural Networks- Transformer Networks	Subsymbolic, black box
Stakeholder Group		
Domain Expert	Check plausability of causal relations	Assessments of quality of inidvidual (local) decisions
Developer	Determine confidence, test fainess and biases	Assessability of the model quality- identify and avoid bias in training data
Regulator	Verify compliance	Verification of the comprehensability
User	Reliability, trust	Verification of the result
Suitable Explanation Strategy/ Type	Post- hoc	
Approach	LRP, LIME	Explanations by prototypes and external knowledge base in combination with a neural network

Table 2: Use Case I: AI supported image analysis of histological tissue sections, e.g., drug testing

Kraus et al. (2022) discuss other explanation methods, such as Grad-CAM, Integrated Gradients, or DeepLIFT (refer to Chapter 3.3.1) in relation to this example. This would be utilized by making artificial alterations to the original image to place it in a distinct cate-

gory while maintaining a similar appearance. By presenting various hypothetical representations, both the medical practitioner and the patient can gain a deeper comprehension of the classification of a disease like "cancer".

Use Case II: AI – supported text analysis of medical reports

With the use of text analytics, AI has the potential to aid in the automatic matching of patient reports, providing similar reports to assist in performing a differential diagnosis. This use case is also applicable to the clinical testing phase of a drug, specifically with the use of neural networks as mentioned earlier. In this particular instance, Transformer Networks could be used for the NLP tasks (Otter et al., 2018; Wolf et al., 2019; Nambiar et al., 2020). The referenced model is a type of Deep Learning model called Transformer Networks. They are classified as such due to their numerous layers. Kraus et al. (2022) suggest that these models have the capability to identify latent features. These can be, for example, indirect references or logical conclusions. The medical data sets are used to train the neural network, which is then adapted for the specific application through transfer learning. This process falls under the category of supervised learning, and the AI system is put into productive use only after the completion and testing of the training process (s. table 3).

Criteria	Description	Requirements/ Evaluation/ Comment
System Type	Decision Support	Augmentation- Human - Machine Case
Type	Similarity analysis	
Economic categorization	Augmentation case- enhance human capabilities	Repetitive work
Business Impact	High	Partially automatize the process- human in the loop
Societal/ Government	Very high	Improve diagnosis quality significant
Criticality/ Impact	Very high	
Data Types	Text data - medical reports	
Typical AI Model used		Subsymbolic, black box
	Neural Networks- Transformer Networks	
Stakeholder Group		
Domain Expert	Increase information gain, plausability of causale relationship	Enable decision support through, e.g., substantive justification (local explanation)
Developer	Determine confidence (robustness, stability)	Deeper understanding of system
Regulator	Verify compliance	Verification of the comprehensability
User	Reliability, trust	Verification of the result
Suitable Explanation Strategy/ Type	Post-hoc	
Approach 1	Working with a prototype- to find similarities	Identification and building of a prototype (e.g. temperature, specific symptoms)- then new reports are identified as a specific prototype and therefore identified as "similar". (This is almost same method like AISOP, where the "hiostoric scenarios" are prototypes- AISOP is using a knowledge base)
Approach 2	Working with a knowledge base	The neural net works together with a knowledge base (knowledge graph) and learn the connections based on the specific symptoms ("patterns") and therefore can identify the "similar" patterns of symptoms in the knowledge base (s. AISOP, s. above Approach 1)

Table 3: Use Case II: AI – supported text analysis of medical reports

Use Case III: AI – supported machine or asset condition monitoring (Predictive Maintenance)

As mentioned, the process industry heavily relies on plants and their operations (discussed in Chapter 2, sections 2.1, 2.2, etc.). It is essential to maintain regular plant maintenance to avoid costly downtime of individual devices, machines, or entire plants. This crucial

aspect is necessary to avoid significant financial losses. So, staying on top of maintenance schedules and keeping everything in good working order is important. It is also necessary to understand that in the process industry, things cannot always be put on hold and then started back up again. This is especially true when dealing with physical, chemical, and biological processes. It is a delicate balance that must be maintained to ensure everything runs smoothly. For this reason, effective early warning systems can be highly beneficial in the process industry. These systems can help signal potential machine or plant malfunctions and the need for maintenance, which can ultimately reduce downtime. By avoiding potential issues, companies can ensure that their operations run as smoothly as possible (s. table 4).

Criteria	Description	Requirements/ Evaluation/ Comment
System Type	Decision Support	Augmentation- Human - Machine Case
Type	Anomailty detection for maintenance planning	
Economic categorization	Automisation - replacement	Partially repetetive, partially new approach
Business Impact	High - very high	System failure, production process stop
Societal/ Government	Low	Company internal. High only if failover will have ecological impact etc.
Criticality/ Impact	High	Function safety, economic efficiency
Data Types	Numerical and textual	Sensor data, operational parameters, error codes, machine log data
Typical AI Model used	Bayesian Networks	
	Machine learning based on knowledge graphs	
Stakeholder Group		
Domain Expert	check plasaubility of causal relationships - find cuasal relationshis, determine confidence (robustness, stability), imporve interaction possibilities	Assessment - plausability, statistical evaluation of the models, assessment of an individual decision (local explanation)
Developer	determine confidence (robustness, stability)	Assessment - plausability, statistical evaluation of the models, assessment of an individual decision (local explanation)
Regulator		
User	check plasaubility of causal relationships - find cuasal relationshis, determine confidence (robustness, stability), imporve interaction possibilities	Assessment - plausability, statistical evaluation of the models, assessment of an individual decision (local explanation)
Suitable Explanation Strategy/ Type	Post-hoc	Real time monitoring of systems/ assets
Approach 1	use and fitting of surrogate models (model plausability)	Explanations by prototypes and external knowledge base in combination with a neural network
	Extraction of statistical quality (bayesian statistics)	
Approach 2	Natural language explanation by using knowledge graphs	

Table 4: Use Case III: AI – supported machine/ asset monitoring (Predictive Maintenance)

Use Case IV: AI – supported process control in the process industry

Monitoring the status of process production is crucial for ensuring smooth operations and identifying potential issues before they become significant problems. By detecting the current state of the process, we can derive follow-up processes that help optimize operations and ensure that everything is working optimally. Whether determining the optimal operating sequence or deriving the most effective operational trajectory, production goals are to be achieved while maintaining a safe and efficient workplace.

In the process industry, using systems and models has been a longstanding practice. This is due primarily to the unique production process that is involved in this industry. AI can definitely be useful in both cases mentioned above. On the one hand, it can assist in condition detection. When dealing with complex dynamic systems, it can often be challenging

to determine the system's current status. Many factors are at play, and it can be challenging to track them all at once. However, it is vital to remain vigilant and stay on top of things to ensure the system functions as intended. With careful monitoring and analysis, it is possible to better understand the system's current status and make any necessary adjustments to keep it running smoothly. Depending on the process, this can be done either by variables using sensors to give an up-to-date (real-time) picture of the status or by visual inspection (other possibilities along the senses are of course conceivable). Taking and analysing samples could possibly lead to the destruction of the current product or at least slow down or even stop the production process. An inspection based on sensor data or visual inspection therefore seems to be much more reasonable. The determination of the optimal sequence of subsequent processes is, in turn, a highly complex and sensitive requirement, if one considers the statements made above about the problem of starting and stopping processes. Any critical parameters resulting from the process, such as temperature or pressure, must be given special attention. An autonomous system that is to be used in this environment must be regarded as highly critical (s. table 5).

Criteria	Description	Requirements/ Evaluation/ Comment
System Type	Partially decision support- partially autonomous system	
Type	(1) AI- assisted analysis (state detection - here image analysis) (2) AI-supported feedback control (optimum operating procedure)	
Economic categorization	Automisation- Augmentation	New approach, partially repetitive (system monitoring task)
Business Impact	High - very high	System failure, production process stop
Societal/ Government	High	Possible heavy impact on ecosystem and (regional) society
Criticality/ Impact	Very high	Function safety
Data Types	Numerical data, image data	Sensor data, operational parameters, error codes, machine log data, image data
Typical AI Model used		
	(1) Neural Networks (s. use case 1)	
	(2) Reinforcement learning (model predictive control) - hybrid models	
Stakeholder Group		
Domain Expert	Determine confidence (robustness, stability, vulnerability), check plausibility of causal relationships, improve information and interaction possibilities	Eplainability of individual - local - decision
Developer	Determine confidence (robustness, stability, vulnerability), check plausibility of causal relationships, improve information and interaction possibilities	Single decision explanations and model explanations (local and global)
Regulator	Verification of "comprehensability" and protection concept	Single decision explanations and model explanations (local and global)
User	Similar to domain expert - operator	
Suitable Explanation Strategy/ Type	Ad-hoc Post-hoc	Real time monitoring of systems/ assets
For (1)	LIME	
For (2)	Integration of black box models - hybrid modeling	

Table 5: Use Case IV: AI – supported process control in the process industry

Ensuring that any AI systems used in production areas meet the necessary requirements is crucial. This ensures optimal performance, safety, and compliance with regulations. The need has increased enormously as more and more AI components are installed in robots or production systems. In the European Union, there is the Machinery Directive 2006/42/EC of 17 May 2006, which deals with ensuring that the technical machines used meet the safety requirements. This Machinery Directive was amended on 10 May 2023 (PE-6-2023-INIT). Once the President of the European Parliament and the President of the Council

have signed it, the Regulation will be released in the Official Journal of the European Union. After its publication, it will take 20 days for the Regulation to come into effect. Ensuring that any AI systems used in production areas meet the necessary requirements is essential. This ensures optimal performance, safety, and compliance with regulations. Member States and economic operators will have 42 months before the rules of the new regulation are applied. One of the aims of this amendment is to meet current requirements, for example, according to the EU, more and more machines are being placed on the market which are less dependent on human operators. These machines are used in certain delimited areas for specific tasks but are able to learn and thus perform new actions in the respective context and thus become more autonomous. This creates new requirements for safety.

In the process industry, the regulations are even more comprehensive, since this industry works with hazardous substances, under high pressures, etc. the regulations here are, for example, the SEVESOIII Directive of the Federal Immission Control Act (Twentieth Ordinance on the Implementation of the Federal Immission Control Act). To use AI in the process industry environment, it must be ensured that the system is sufficiently transparent and that the decisions are repeatable, comprehensible in detail and correctable. The German Institute for Standardisation DIN has developed a roadmap for AI standardisation. VDE-AR-E-2842-61-1:2020-07 has already been published and contains a description of the terminology and basic concepts of explainable AI (VDE-AR-E-2842-61-1:2020-07).

The Use Case V: Supply Chain Risk Analysis with SPA and The Use Case VI: Scenario Analysis for Early Warning of Power Failures in the Process Industry

These cases are presented in Chapter 5.2.3, as their system architecture has been incorporated into the development of the Re_fish reference architecture (s. Chapter 5.2.3).

Use Case VI: AI – time series forecasting

Forecasting (s. table 6) is one of the main tasks in planning - based on historical data (possible bias in the data must be taken into account, see Chapters 3.2.1 and 5.2.8), a forecast is calculated in order to anticipate, for example, the future demand for a product, future price development, etc. These requirements are needed both in the area of scenario planning and in the area of tactical S&OP planning. The complexity and possibilities of the methods range from ARIMA (AutoRegressive Integrated Moving Average) to Global Deep Learning Forecasting Models to Neural Basis Expansion Analysis for Interpretable Time Series Forcasting (N-BEATS).

(Manu, 2022; Montero-Manso & Hyndman, 2020; Oreshkin et al., 2020).

Criteria	Description	Requirements/ Evaluation/ Comment
System Type	Decision Support- Augmentation	
Type	Time based Forecasting	
Economic categorization	Automisation- Augmentation	New approach, partially repetitive
Business Impact	Medium- High	Based on forecasting all demand, supply etc. plans will be done
Societal/ Government	Low	Company internal
Criticality/ Impact	Medium	Economic impact on company
Data Types	Numerical data	Historic sales data, current market data, expert adjustments
Typical AI Model used		
	(1) Local model, like auto.arima, TBATS	
	(2) Global methods, Linear Autorgressive, Featurized Linear Autoregressive, Deep Network Autoregressive, Regression Tree Autoregressive	
	(3) N-BEATS	
Stakeholder Group		
Domain Expert	Determine confidence (robustness, stability), check plausibility of causal relationships, improve information	Eplainability of individual - local - decision
Developer	Check plausability of causal relationships - find cuasal relationshis, determine confidence (robustness, stability), imporve interaction possibilities	Assessment - plausability, statistical evaluation of the models, assessment of an individual decision (local explanation)
Regulator		
User	Similar to domain expert - planner	
Suitable Explanation Strategy/ Type	Post-hoc	
For (1)	LIME, SHAP	
For (2)	LIME, SHAP	
For (3)	TimeSHAP, Instance-wise Feature Importance in Time (FIT), Dynamask	

Table 6: Use Case VI: AI – Time Series Forecasting

The special features of the process industry have already been highlighted above. These are, on the one hand, the high proportion of complex production facilities - asset intensive production - and, on the other hand, the networking in highly complex supply chains. Therefore, two topics are of particular importance in the context of corporate planning scenarios - strategic scenario planning (which was already used by Shell in the oil crisis of the 1970s to plan and coordinate capacities in good time, see Schoemaker and van de Hejden (1992) (Schoemaker & van de Hejden, 1992). In Chapter 5.2.3, two use cases or applications are presented for the area of scenario planning with AI applications. In addition, there is also work in other sectors, e.g., in the utilities sector, which deals with the combination of the methods presented in more detail in chapter 3. For example, Eibeck et al. (2020) with their parallel world framework for scenario analysis in knowledge graphs (see also Chapter 5.2.3 AISOP) or Rezaei et al. (2018), A new approach based on scenario planning and prediction method for the estimation of gasoil consumption, in which the prediction results of a neural network and a multilinear regression (MLR) model are compared. Ge et al. (2017), who address the question of the state of the art of machine learning or big data analytics in the process industry and identify the areas of company-wide process monitoring (see Use Case IV) as possible applications (Ge & Chen, 2016 - Plant wide...) or the use in the area and improvement of sustainability efficiency of the energy used (Bakshi & Fiksel, 2003; Hanes & Bakshi, 2015). Or the use of process causality methods knowledge and data based, for fault abnormality detection (Chiang & Braatz, 2003). The use of sensors in the context of Industry 4.0 and their evaluation (e.g., Xu et al. 2014). Yang et al. (2021) deals with intelligent production and the requirements of Industry 4.0

in batch production, which is so typical for the process industry. Yang et al. (2021) describes an intelligent system consisting of a self-learning knowledge base and a "cognitive system". Toorajipour et al. (2020) deal with AI in supply chain management and identify in their survey AI methods that are used in the context of supply chains. Artificial neural networks take the top position, followed by fuzzy logic and models, multi-agent and based systems, for example, to balance demand supply etc. in the context of simulations and thus contribute to better decision-making.

Finding 11: The use of XAI in companies in the process industry naturally depends on the use of AI in the companies. Potential applications have been identified in the areas of scenario planning, sales and operation planning, e.g., forecasting, process control, etc., which, when considering the use of AI in the area of research and development as well as in automated process control, have a significant - positive economic impact in the sense of the economic growth drivers presented in Chapter 1.1. and Chapter 1.2 respectively.

2.3 Planning and Decision-Making in the Process Industry

As mentioned above in chapter 2.2, planning is a central process in the management process and has a critical function in the process industry (s. chapter 2.2, summary). It is even more important in the chemical and life science industry sectors, as both operate within highly interconnected and international supply networks and involve complex and interlinked processes. In these processes, raw materials are transformed into intermediate and finished products through chemical reactions and physical operations.

Therefore, effective planning is essential to ensure that these processes run smoothly, efficiently, and cost-effectively. This requires a thorough understanding of the various process steps, the timing and sequencing of operations, and the interdependencies between process stages. In the chemical industry, planning involves scheduling production batches, allocating resources such as equipment and personnel, and the management of inventory levels. Effective planning can help to minimise downtime, reduce waste, and optimise the use of resources. In the life science sector, planning is critical for developing and producing pharmaceuticals, biologics, and medical devices. This involves coordinating research and

development activities, clinical trials, and regulatory approvals, as well as scheduling manufacturing processes and managing supply chains.

In today's world, modern companies use information systems to carry out highly complex planning processes. One type of these information systems is, aside from ERP (Enterprise Resource Planning), advanced planning and scheduling (APS) systems. Such APS software is widely used in the process industry to optimise planning and scheduling processes. This software can integrate data from various sources (including production schedules, inventory levels, and supply chain information) to generate optimised programs that minimise costs and maximise efficiency. In addition, digital technologies such as artificial intelligence (AI) and machine learning (ML) are becoming increasingly prevalent in the process industry. These technologies can help to improve forecasting accuracy, optimise resource allocation, and enable predictive maintenance, leading to improved efficiency and reduced costs.

Overall, effective planning is critical to the success of the process industry, particularly in the chemical and life science sectors. By using advanced software and digital technologies, companies can optimise planning and scheduling processes, reduce costs, and improve efficiency, ultimately leading to improved profitability and competitiveness.

In the following, the focus is on two planning frameworks: scenario planning and integrated business planning (an extension of sales and operations planning – S&OP). As in the case of the decisions made within sales and operations planning (S&OP), AI and especially xAI can have a significant impact on business operations and financial performance. Within S&OP, AI can be used to analyse large volumes of data, identify trends and patterns, and make predictions about future supply and demand. These predictions can be used to inform decisions about production scheduling, inventory management, and sales forecasting, among other things.

However, it is important to ensure that the decisions made by AI systems are explainable and transparent to stakeholders, including sales teams, operations managers, and executives. This can help to build trust in the AI system, improve decision-making, and facilitate collaboration between different teams.

One approach to XAI in S&OP is using machine learning algorithms that can explain their decisions. These explanations can be in the form of visualisations, charts, or natural lan-

guage descriptions, depending on a user's needs. For example, a machine learning algorithm used for sales forecasting might generate an explanation for its prediction, based on factors such as historical sales data, market trends, and product promotions. Another approach to XAI in S&OP is to use inherently interpretable models, such as decision trees or linear regression models. These models can be easier to understand and explain than more complex types like neural networks, which can be more opaque thus more difficult to interpret. Overall, XAI is an essential consideration in S&OP, as it can help to improve the transparency, accountability, and trustworthiness of AI systems being used in that context. By making AI more explainable, businesses can ensure that the decisions made by these systems align with their strategic objectives.

Planning is one of the main tasks in the management cycle of a company, the reason for building a plan is a potential deviation of a given system state from its desired state, as perceived by the planner. This deviation is regarded as needing optimisation or may be considered no longer acceptable. This is a problem "[...] which can therefore be regarded as a deviation of a current or expected state from a desired state described by goals. We also speak of a decision problem"(Klein & Scholl, 2011, p.1) Decisions must be made to solve the problem – in this case, in order to eliminate the deviation from the desired state (Klein & Scholl, 2011).

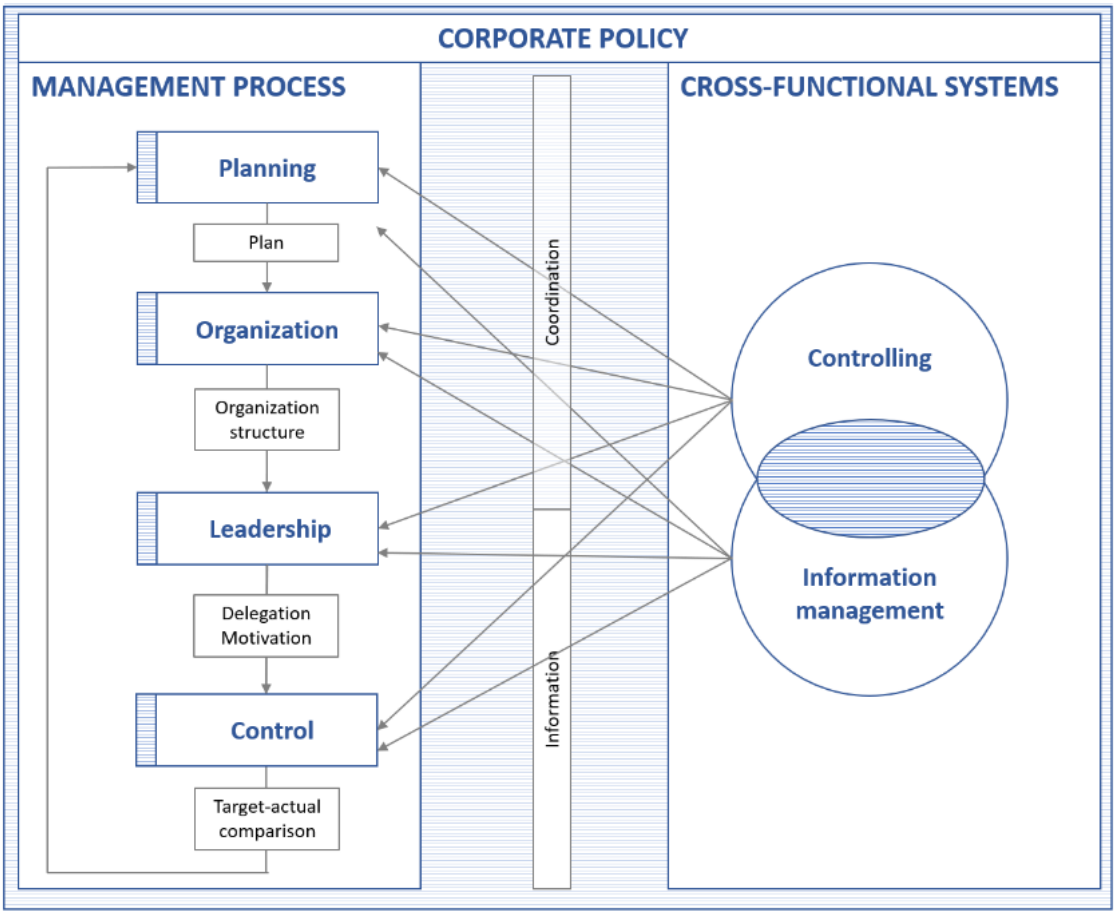

Figure 31: Planning process

In figure 31, one can see that the management process includes planning, organisation, leadership, and control. The supporting systems are for controlling and information management.

In Chapter 2.3.1, scenario planning as a method or tool within strategic planning will be introduced. Chapter 2.3.2 describes integrated business planning as an extension or enhancement of sales and operations planning (S&OP); in Chapter 2.3.3, the entire planning process will be synthesised and described, as well as where AI (and especially XAI) can be of help in PI planning processes.

As mentioned above, planning concerns information-processing. Therefore, the term 'information' must be described. Information may be defined as judicial knowledge relevant to a decision (Klein & Scholl, 2011). In the planning or planning process, information on various sub-areas is required. For example, information is needed about the state of the

problem, the goals, the various alternative courses of action, different environmental developments, and interdependencies between actions and their results (Approach XAI). Information relating to facts that cannot be influenced (present or future) is referred to as data. The data-character of information is therefore situational, personal, and problem-dependent (Klein & Scholl, 2011).

One of the methods or tools within the strategic planning process (s. Chapter 2.3.1) is scenario planning. As described above, companies in the process industry are part of a highly complex network of companies. They are particularly vulnerable when these networks or their connections break down. Therefore, process industry companies try to use tools such as scenario planning, in addition to risk analysis, to make their supply chain as resilient as possible.

2.3.1 Scenario Planning in the Process Industry

The idea behind scenario planning is that by identifying fundamental trends and uncertainties, a manager can construct a series of scenarios that might help "to compensate for unusual errors in decision making – overconfidence and tunnel vision" (Schoemaker, 1995). Often, managers or decision-makers made wrong decisions in the past because they had not anticipated possible scenarios. They all made a kind of myopic statement; the list is long, e.g., Ken Olsen or Thomas Watson.[16] Scenario planning is a disciplined method for imagining possible futures that companies have applied to numerous issues. Schoemaker states that Royal Dutch/Shell has used scenarios since the early 1970s to generate and evaluate its strategic options. Since then, Shell has been seen consistently better in its oil forecasts than other major oil companies. Shell was also one of the first companies to see the overcapacity in the tanker business and Europe's petrochemicals (Schoemaker, 1995).

Scenario planning, or the scenario technique, is a strategic planning tool. It is beneficial because it systematically analyses alternative developments, breaks them down into individual steps, and asks for the appropriate alternative courses of action (Mössner, 1982).

[16] Ken Olsen, the founder of Digital Equipment, is said to have predicted that there is no reason why anyone should have a computer in their home. Thomas Watson, who headed IBM, claimed that the world market for computers was no more than five computers - Watson was considered one of the best salesmen of his time. https://www.watson.ch/digital/microsoft/207532210-5-beruehmte-zitate-ueber-die-zukunft-die-alle-frei-erfunden-sind , accessed 18.06.2023

This technique addresses the insufficiency of using only historical or planned data in a completely changed environment, e.g., one wherein a current model is not appropriate anymore.[17] By using scenarios, both the uncertainties and the future orientation of the planning are considered. This is becoming more necessary, as in recent decades, the world market dynamics, and the complexity of supply networks (especially in the process industry) have grown enormously. This and the higher velocity of new, changing constraints in society and technology on a global scale have led to a vast number of discontinuities and other uncertainties that require quick reactions by management. These disturbances and their impact must be taken into consideration in strategic or corporate foresight. Besides these external factors, certain internal factors could lead to suboptimal strategic planning, e.g., if future opportunities, innovations, and trends are not anticipated. Successful companies, by using scenario planning, are starting to launch programs and initiatives to prevent threats early, even when they have a long duration. This is mainly because such companies use future-oriented scenario planning or techniques based on several quantitative and qualitative methods (Kahn & Wiener, 1969). The idea of this technique is that alternative pictures (scenarios) of the future will be constructed, using succeeding events and branching chains, which may also provide a basis for strategic management planning (Welge & Eulerich, 2017). The definition of a scenario from an economics perspective is such: "A scenario is to be understood as a description of a possible future situation in which potential developments of all environmental factors and internal factors relevant to the company as well as the factor interdependencies are considered" (Welge & Eulerich, 2017).

A combination of different scenarios leads to a series of possible future developments, which can be used for decision preparation or decision-making. The objective of the scenario technique is grounded preparation for strategic approaches to upcoming eventualities. Scenarios are also used within the management of crisis, discontinuity, and risk.[18]

[17] This is (somehow) similar to the conceptual drift, when used in supervised machine learning and the model is not appropriate anymore for the current environment/ecosystem.

[18] s. chapter 3.2.2 and 5.2.2 - AISOP and SPA

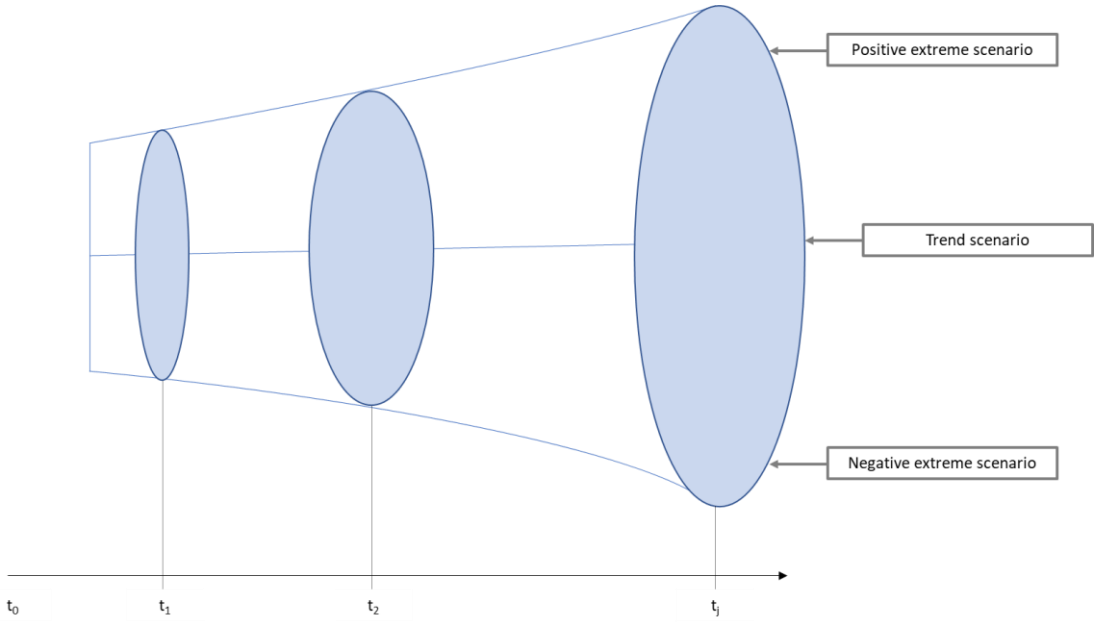

Figure 32: Funnel model of scenario planning technique

Typical result of the scenario technique is the funnel model shown in figure 32; here, it is shown with three "possible" scenarios: positive, trend, and negative.

Within scenario planning, AI/ XAI systems can be of help in forecasting. AISOP (s. Chapter 3.2.2 and 5.2.2) is a knowledge-based system for scenario planning; SPA is a risk planning system. Schoemaker (Schoemaker, 1995) emphasises the importance of forecasts and categorises them on the basis of the degree of prediction uncertainty, and how complex the planned (or predicted) issue is. Forecasts are important in scenario planning, as well as for integrated business planning. Uncertainty concerns the degree of available knowledge about the target variable.[19] As humans may display overconfidence when they do forecasts, uncertainty can be defined as the level of disagreement among forecasters, or the doubts of a single forecaster, regarding the correct value of an unknown interest: "We are too sure of our single view about the future and fail to consider alternative views sufficiently" (Fischhoff et al., 1977). One of the reasons for this overconfidence is that one may suffer from the inability to envision all of the possible pathways (Schoemaker, 1995). Other reasons may include the following:

[19] Bounded rationality – s. above

- Illusion of control

  People harbour erroneous feelings of control, which get stronger as they attempt to predict the future.

- Information distortion

  Bias occurs because information may not represent the actual situation; people tend to overestimate the information closest at hand. (This can only be overcome by consulting available data.)

- Risk perception

  Regarding risk perception, people dread the risks they have a poor understanding of, or which they have no control over. Furthermore, people react to saliently presented risks and may overweigh them, instead of those presented otherwise (car accidents and plane accidents vs. cancer, etc) (Kahneman & Tversky, 1974, 1982).

Another important aspect emphasised by Schoemaker (1995) is so-called *complexity*, which he defines as the number of variables and how deeply they interact in a desired prediction task. As Schoemaker (1995) points out, there is extensive literature on research into heuristics and biases and how often these mechanisms affect a decision-maker's uncertainty estimates; however, they do not address the issue of interrelatedness. He points out that people as decision-makers are only able to aggregate additively, rather than being able to understand interrelationships- or even causality (Schoemaker, 1995; Pearl, 2018). It is typical of human behaviour that people, and in this case -- decision-makers, tend to (or even need to) simplify the world that surrounds them through cognitive tools such as associative networks, scripts, schemas, frameworks, and mental models. Additionally, whenever new information is discovered, people tend to insert it into an existing frame quasi-associatively, without moving said existing frame. This filters the information, which may then be completely wrong, because the frame has in fact changed. Therefore, in addition to new information, the frame must also be permanently checked and adapted, if necessary (Russo & Schoemaker, 2016).

Complexity has two other dimensions: it concerns cross-sectional complexity, namely how data is connected at a certain point, and on the other hand, there is dynamic complexity, in which time-elements take on the role of feedback loops (Russo & Schoemaker, 2016).

Scenario planning is a technique within strategic planning, and supports a decision maker in developing different scenarios, thereby preventing common biases and ‘gut-feeling’ decision-making. The positive aspect of scenario planning is that it is possible to predict future developments and adjust one’s own behaviour and decision-making. It can be used as an ongoing technique to evaluate a corporate strategy. Scenario planning is very time consuming and does not deliver entirely perfect predictions. Without a tool, there is a problem of currency and complexity, as it might not be possible to evaluate all possible developments in a timely manner. In mid-length and longer run terms, there is a higher likelihood of disruptive events. Information systems, especially AI models, can be of great support within scenario planning. AI can help to gather current information (currency and provenance) and provide this information within the decision-making process. AI models can also be of use in evaluating the risks of a specific decision, and in gauging scenario probabilities. However, because humans follow a number of biases in their behaviour, it is all the more important that in addition to AI models for supporting decisions, these are also explained in a way that is understandable to users. For this purpose, methods of explainable AI are applied (see chapter 3). Chapter 3 also examines the two systems AISOP and SPA and analyses their architectures, particularly with regards to explainability.

Scenario planning is usually carried out by a strategic planner and analyst on behalf of the management board/board of directors. The management board also initiates the entire strategic planning cycle and monitors the process holistically. The specifications of the management board are, for example, the KPIs, the selection of business areas (product - market combinations), etc. The development of a vision as a guiding star for the entire company are also tasks of these stakeholders. Above these stakeholders are the owners (or the board of directors) and the local, global, etc. society. The auditor is usually a watchdog appointed by the regulator, in this case e.g., the legislator, to ensure compliance with the rules.

2.3.2 Integrated Business Planning in the Process Industry

Integrated Business Planning is considered an improvement on Sales & Operations Planning (S&OP) (s. Hsieh & Hsu, 2012; Willms & Brandenburg, 2019). The term Sales & Operations Planning was first used by Dick Ling in his book "Orchestrating Success" in the 1980s (Ling & Goddard, 1988). At the time, another concept was predominant and well known, namely Manufacturing Resource Planning (in short, MRP II). While MRP II was

focused on a single manufacturing plant, S&OP was seen as the overarching starting process for a business.

Sales & Operations Planning is seen as a forward-looking process, with a minimum horizon of around eighteen months or six quarters, integrating and aligning strategic and tactical views and decisions, and directing operational planning and overall execution, as a process for integrated decision making (see figure 33). As S&OP can be considered an integrated decision-making process, it must be the driver of tactical and operational planning and execution; the financial perspective within S&OP is its support of the business plan. To ensure that decisions will be made beyond the end of a given year, the planning horizon must be at least eighteen months. Coldrick et al. (2003) define the operational planning as being the day-to-day execution of an operational plan. Tactical planning is about delivering the year's budget and the strategic plan for performance in future years. As an integrated form of decision making, it ultimately enables a business to monitor and update its strategies by using tactical planning and reporting, on a monthly basis, using the operating plan.

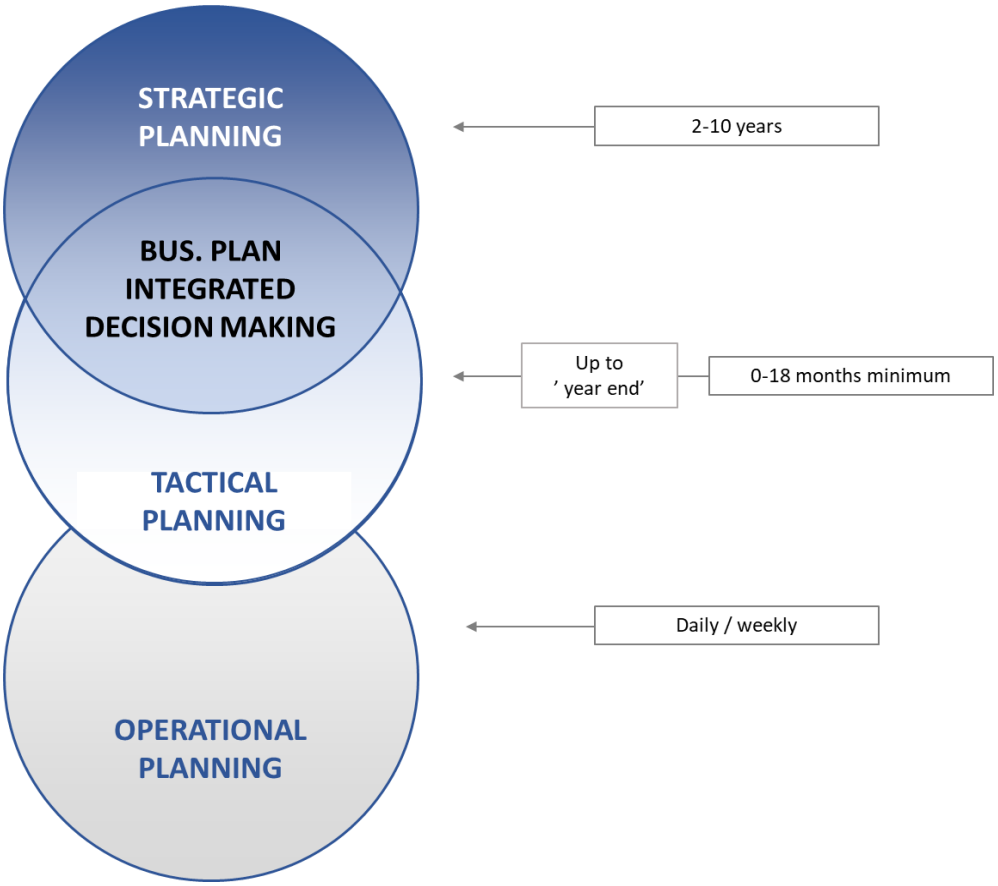

Figure 33: S&OP Planning- Coldrick et al. (2003)

The main idea of S&OP is to integrate business, sales, and production planning. When the plans are created in silos and not aligned, it may lead to massive disconnect, department-optimised plans, and many disputes between sales, marketing, and manufacturing. A typical situation is mentioned by Coldrick et al. (2003) The finance department raises an initiative to improve working capital by reducing inventory levels (time purchased inventory is held until it is transferred into cash). When this initiative is not aligned with the marketing and sales departments, it could lead to customer service failures; when sales, marketing, and manufacturing initiate customer service improvements, it leads towards reduced working capital. One of the major findings at the beginning of implementing S&OP was that inventory and customer service resulted from the plan, while first and foremost supply and demand were the drivers.

S&OP subsequently became a logistics matter for supply chain managers, who are measured by their volumes; it was the goal to get single volume numbers. Sales, marketing, and finance were more interested in a range of numbers. They started doing more of their own financial scenario planning, and without being linked to finances, volume forecasting became less of a priority than the financial forecasting, as sales, marketing, and general management were measured on financial results, while manufacturing and the supply chain were measured on operational targets, based on volume predictions. As a result, any number provided by S&OP was overridden by the budget (Coldrick et al., 2003).

The revolutionary idea behind the S&OP, according to Coldrick et al. (2003), was that once a month, after forecasting the demand in the demand plan and the reconciliation of supply and demand, figures would be aligned with sales, production, and inventory. After this alignment, a pre-S&OP meeting would be set up, during which the aligned plans would be agreed upon, and another meeting would be prepared with department heads and C-Level managers, for overall alignment. After that meeting, a reconciliation of volumes might be carried out with financials and a check against the budget in a respective period of time.

With the growth of markets, globalisation, and so forth, the Group's environment became more complex, and with it, the S&OP planning process. Sales, marketing, and finance of the legal entities should be controlled regionally. The many sales and marketing units were interfaced with many procurement units. This is relevant, among other things, to the process industry, whose operations are integrated in highly complex networks.

The S&OP process then developed increasingly into what is now referred to as Integrated Business Planning.

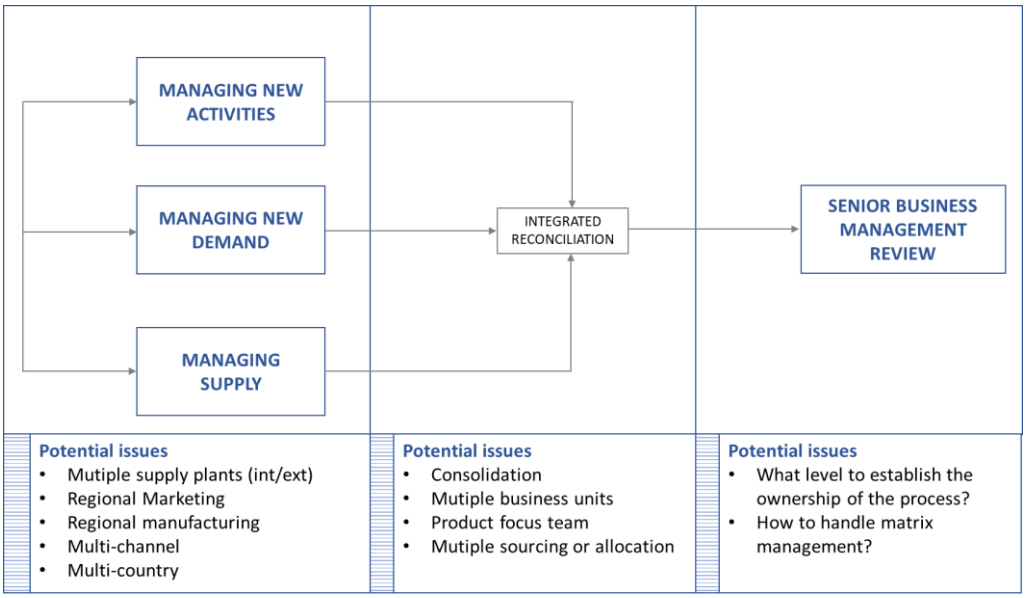

Figure 34: Change from S&OP planning towards IBP- Coldrick et al. (2003)

On the basis of figure 34, one can observe the change in the direction of an integrated decision-making process. For example, the coordination may be seen between the various functions after involving the finance department from the outset and changing the coordination from a volume-related to a business-related direction. In addition, especially in highly innovative sectors, innovations were thus managed, using the Stage Gate model. In the new product launches, not only the share within which new products generate a positive cash flow was considered, but also the entire life cycle, and thus possible cannibalisation effects, etc. A demand plan which is sustainable over eighteen months can only be achieved when the plans for functions are coordinated, volume and value are integrated, and finances as well as the supply chain are committed to the plan. While S&OP was done at the SKU level, Integrated Business Planning normally starts at a higher aggregated level.

The process of integrated usability planning envisages that the supply side of the company does not take the lead in the S&OP process but proceeds in an integrated manner. The establishment of a continuous coordination process is the most important step in S&OP, towards Integrated Business Planning. This can be seen in figure 35.

Figure 35: Reconciliation within integrated business planning- Coldrick et al. (2003)

The idea of the Integrated Business Planning process is to fit the fragmented elements of the value chain into one which is regionally integrated. The challenge of a multinational S&OP process is to define where the steps of an integrated business planning process (new activities, demand, supply, integrated reconciliation, and senior management review) have to take place. While new activity launches and demand management in the fast-moving consumer goods industries are usually handled in countries, new activity direction is managed regionally or globally; supply is managed regionally, and reconciliation processes and senior management reviews are carried out in the countries and region. In pharmaceutical and chemical companies, the decisions are by management, and taken globally. Figure 36 shows the decision framework of a household goods company and the balancing of the different decision variables in a distributed environment – country, region and global.

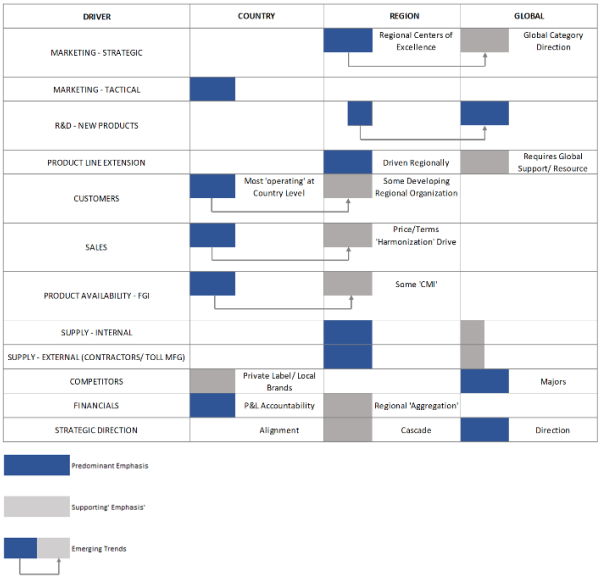

Figure 36: Sample of a decision framework for a household goods company – Coldrick (2003)

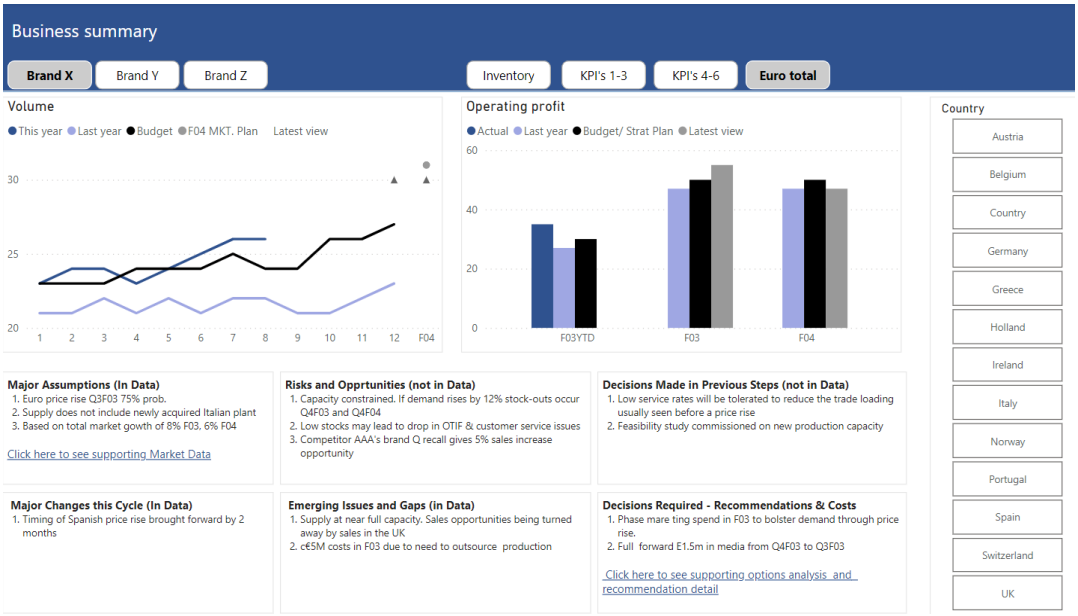

Figure 37: Sample of a decision dashboard sample with explanations

Figure 37 shows a typical decision support dashboard[20] and how it could look. It depicts major KPIs related to the overarching goal of the plan. Also, there are assumptions in the data and major changes which are relevant during the current cycle, as well as their explanations. Coldrick et al. (2003) point out that such decision support systems can be augmented by modern information technology. The statistical forecast models are so complex that small changes have corresponding effects, and must be seen as a black box, due to the lack of explanation. Typical questions asked by management include the following:

- What major assumptions is this forecast built on?
- What changes to assumptions have been built into this forecast since the last cycle?
- What issues and gaps should I know about?
- What are the risks and opportunities surrounding this recent view?
- What decisions have already been taken but are not yet reflected in this view?
- What decisions should we be taking now?

A statistical forecast in the context of an S&OP process is an important instrument; however, a high-level adjustment often has a greater effect than a change in a forecast at the SKU level. Further improvement of the forecast accuracy at the SKU level leads to an illusion of accuracy, which is not understood by management because the forecast lacks high-level assumptions (Coldrick et al., 2003).

[20] Such a decision board can be seen as a good example for an implementation of Re_fish (s. chapter 5.2.7)

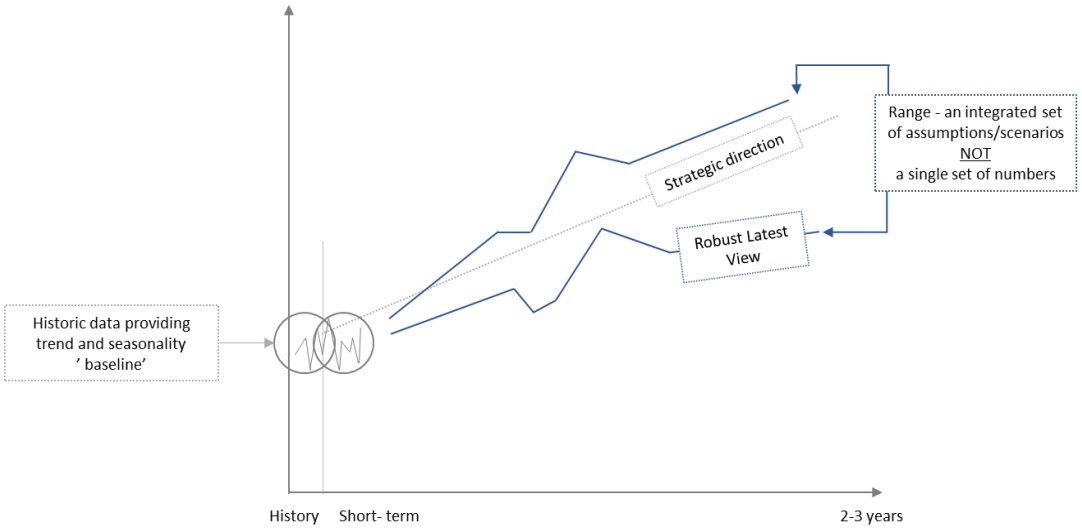

Figure 38: Recognising inherent uncertainty (s. scenario planning) – Coldrick et al. (2003)

The further into the future the forecast goes, the more uncertain it is the forecasts are therefore forecast scenarios used to make assumptions about parameters (external and internal). Figure 38 is a good illustration of the uncertainty of planning and decisions made in the present based on assumptions about the future. The use of different scenarios requires different inputs from the business and an understanding of how and when to use the results, taking into account uncertainties; this shows the maturity of an organisation in using integrated business planning. Integrated business planning is a technology-based approach to managing a company's future-oriented activities, i.e., forecasting, planning and budgeting. It enables each business unit to plan meaningfully and provide the figures for company-wide planning, budget analysis and reporting: The sales department plans sales, the marketing department plans marketing, the manufacturing units plan manufacturing. Integrated business planning makes it easy to take the necessary information from the individual departmental plans and immediately consolidate it into a company-wide view. Integrated corporate planning has several advantages. One is that an integrated approach supports a highly participative, collaborative, action-oriented style of planning and budgeting that builds on short, frequent planning sprints. This promotes more accurate plans as refinements are made at shorter intervals, allowing greater flexibility in responding to market or competitive changes. An ongoing, collaborative dialogue on target achievement brings together finance, business unit managers and executives to promote better alignment and buy-in. When other departments in a company see the budget as the finance department's

work and not their own, it is harder to achieve accountability. And that can easily happen if you spend a lot of time rolling up and consolidating spreadsheets (Coldrick et al. 2003).

Kugel (2023), from Ventana Research, points out that an important benefit of integrated business planning is that plans can be more relevant. As well, corporations that have short planning cycles are able to update their plans more frequently. Ventana Research found out that under one-half of organisations state that their workforce plans remain relevant over their whole planning period, and only the low number of 45% of demand plans remain relevant, even after being updated every month or at least every quarter. In terms of financial plans, Ventana found out that only 29% of the budgets remain relevant during the planning period (yearly). In effect, the organisation (especially the departmental leaders) start improvising. There may also be a lack of coordination between business units and departments (organisational adjustment s. 2.3.3 – the strategic planning process). Therefore, integrated business planning is highly beneficial to the senior leadership team, as it achieves a closer strategic alignment across the whole corporation, achieved by dint of a high participation and collaborative process that combines operational as well as financial elements (Kugel, 2023).

The main idea behind Integrated business planning is to integrate all plans within a corporation, to deploy the business strategy and drive business management. Scenarios are also part of the concept to optimise the business plans and the performance. Integrated Business Planning processes are enabled by technology – that is, supported by information systems. Depending on the complexity, it might be even necessary to use applications to plan and simulate decision alternatives (Markin et al., 2021). However, AI systems can add value to even this process. This can be by using AI models to forecast the demand and supply planning – or by using specific models within the plan. To get the expected results, it is necessary to provide the user (stakeholder) with current understandable and trustworthy explanations (s. Chapter 3).

2.3.3 Decision Making and Explanations in Planning in the Process Industry

In Chapter 2.3, planning and decision making was introduced as one of the main tasks in the management cycle of a company. The reason for the planning consists of the deviation of the system state under consideration from a desired state, as perceived by the planner. This deviation is considered in need of optimisation or as no longer acceptable. This is

referred to as a problem "[...] which can therefore be regarded as a deviation of a current or expected state from a desired state described by goals. It can also be described as a decision problem". Decisions must be made to solve the problem, e.g., to eliminate the deviation from the desired state (s. Chapter 2.3) (Klein & Scholl, 2011; Chakraborti et al., 2020).

In figure 30 (s. Chapter 2.3), the management process consisting of planning, organisation, leadership, and control is shown, which is supported by the cross functional systems that are controlling and information management.

The task of planning is to develop and provide measures (and alternative measures) to solve the decision problem and close the problem gap – the difference between the initial state and the target state. According to Klein/Scholl (2011), the main features of planning are as follows:

- Goal oriented

  A goal must be defined in advance, which describes a desired state and thus shows the gap in relation to the current state.

- Design oriented

  Planning serves the planner(s) as an instrument to shape the future state, according to the ideas of the planner(s).

- Future oriented

  Planning is a more forward-looking and uncertain process, as future developments are difficult to predict.

- Rational process

  Planning is a rational process, but is also subject to the imperfection of information -- Bounded Rationality

- Information processing process

  Planning involves collecting, storing, selecting, processing, and transmitting information.

- Subjective process

Planning is a subjective process that is reflected in the selection of the planning object, the objectives, the planning method, and the evaluation of the results.

Planning can therefore be seen as a fundamentally systematic and rational process based on incomplete information to solve decision-making problems, taking into account factual objectives (Klein & Scholl, 2011). A planning or decision-making problem can be described, according to Wild (1982), based on the following criteria:

- the time range of the planning/decision problem
- the duration of the problem solution and, if necessary, the possibility of modification
- the extent of uncertain environmental influences and dynamics of the environment
- information needs for problem-solving
- the grade of innovations

Wild categorises a planning or decision-making problem in accordance with the following criteria (Klein & Scholl, 2011; Wild, 1982):

- based on the time range of the planning/decision problem
- the duration of the problem solution and, if necessary, the possibility of modification
- the extent of uncertain environmental influences and dynamics of the environment
- information needs for problem solving
- the grade of innovations.

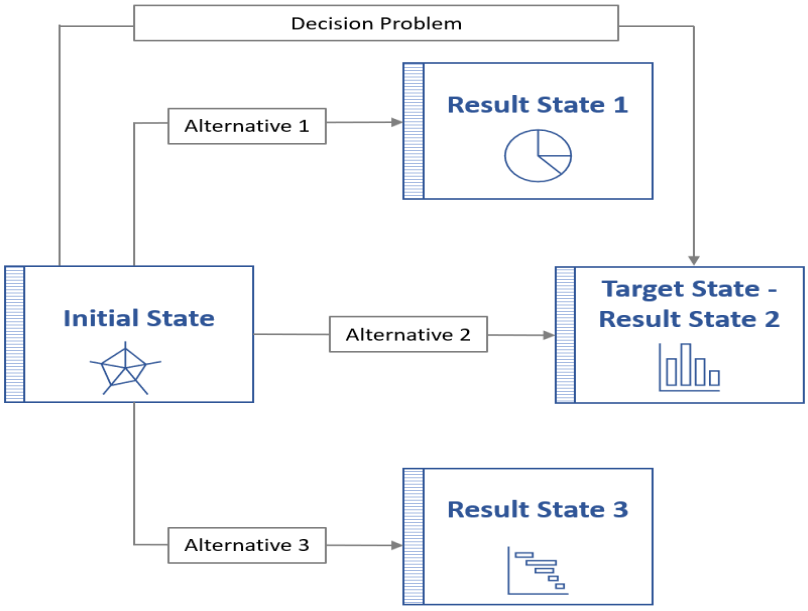

Figure 39: Planning concepts and definitions - based on (Klein & Scholl, 2011)

As already mentioned, planning can be categorised according to different criteria and mapped to the management process; however, it is necessary to define terms concerning planning (s. figure 39).

Initial state:

According to Klein and Scholl (2011), the initial state is a previous or future fact of a system that the planner(s) cannot influence. These facts are therefore referred to as predetermined or estimable data. Due to the uncertainty of the planning, several possible but different target states can be considered in the planning, called scenarios (scenarios or scenario states).

Problem / Decision Problem:

The problem can be described as the difference between a current or predicted initial state, perceived as unsatisfactory or unacceptable, and a desired or desired target state. This difference solves tension which the decision-maker or planner tries to eliminate. It is important to note that problems are not real but purely subjective constructs. The difference, or the tension caused by it, can be remedied by solving the problem in which the initial state is converted into the ideal final state (Berens & Delfmann, 2004).

Alternative courses of action:

The action alternatives describe the various design options that can be used or utilised by the planner as measures to achieve the desired target state. These measures affect the variables, namely the system facts that can be influenced by the planner. There are also different interdependencies between the variables of a system (Klein & Scholl, 2017).

Action results:

The various results of the measures serve to assess the different alternative courses of action about their contribution to achieving objectives.

Target State:

Goals and targets describe the desired target state. These objectives and goals can, in turn, be related to each other and, e.g., compete with each other.

Plan:

The result of the plan is one or more systems of problem-solving measures, which contain the definition of the problem, the objectives, the interdependencies, the results of the action, and also "instructions for the implementation and control of the execution of the plan" (Klein & Scholl, 2011).

As mentioned above, planning pertains to information processing, and therefore, the term information must be defined. Information is purpose-oriented or decision-relevant knowledge. In the planning or planning process, information on various sub-areas is required. For example, information is needed about the state of the problem, the goals, the various alternative courses of action, different environmental developments, and interdependencies between actions and their results. Data is information relating to facts that cannot be influenced whether present or future. The data character of information is therefore situational, personal, and problem dependent (Klein & Scholl, 2011).

Planning is a decision problem related to complex natural systems and it is therefore necessary to reduce the complexity of this problem using a model.

Enterprises and phenomena in economics are described as being systems. Systems are elements (objects) which are linked by types of relations. Models can be homomorphic or isomorphic, dependent on structural identity (Kastens & Kleine Büning, 2021; Stachowiak, 1983; Berens & Delfmann, 2004).

The following describes the planning process based on Kaplan and Norton's (2008) Strategic Management Cycle - from a strategic level -- strategic planning, down to the operational level -- and is shown in figure 40. The whole process aims to develop a strategic and operational plan. The process starts with developing the strategy (1) – defining the company's mission, values, and vision. Subsequently, a strategic analysis is conducted, e.g., by using the scenario planning technique (s. Chapter 2.3.1). The next step is to define a strategy map (or balanced scorecard) by using the measures and targets for the strategic objectives. It is used to integrate the four perspectives of the Balanced Scorecard (s. figure 41). The defined strategic plan and its artifact, the strategy map, is used for communication and alignment of the organisation. The next step 3 includes the building of the financial plan and the start of the sales forecast. With these steps, the planning process moves from the strategic level towards the operational level.

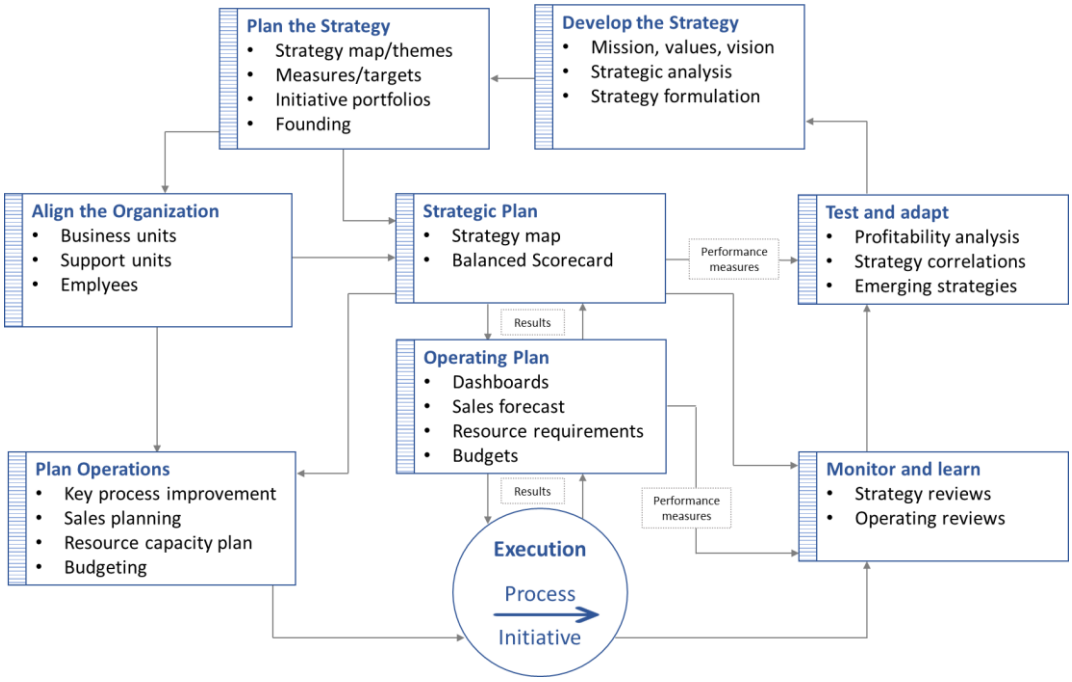

Figure 40: Strategy Management Cycle of Kaplan and Norton (2008)

On the tactical level, the integrated business plan is used to integrate and align all plans (s. Chapter 2.3.2)

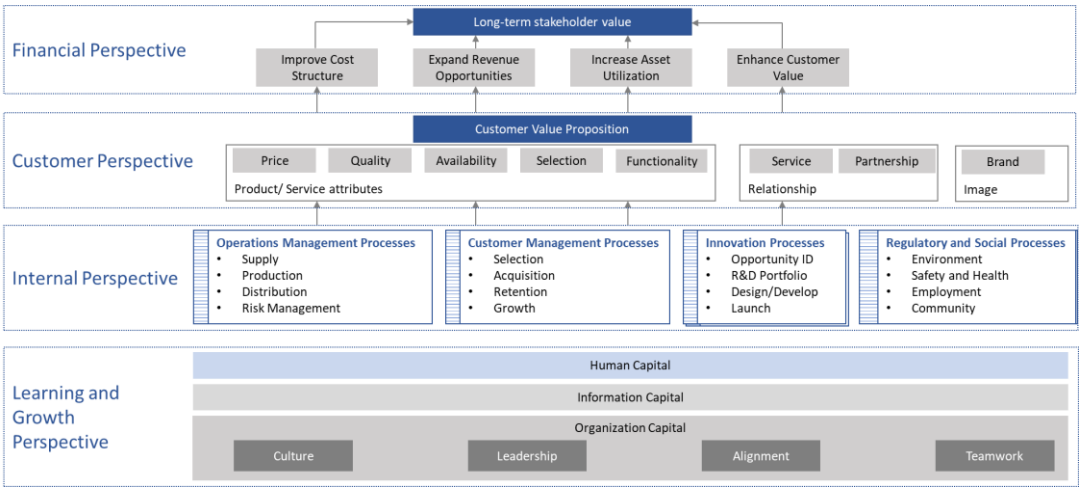

Figure 41: Strategy Map – Balanced Scorecard by Kaplan and Norton (1992)

In figure 42, the whole process is shown on a timeline starting from the right with a two-year perspective ahead, and then moving to the left with a perspective on the past (reporting) on combining the above described strategic and operational planning process with monitor and learn, and test and adapt (left side of the graphics).

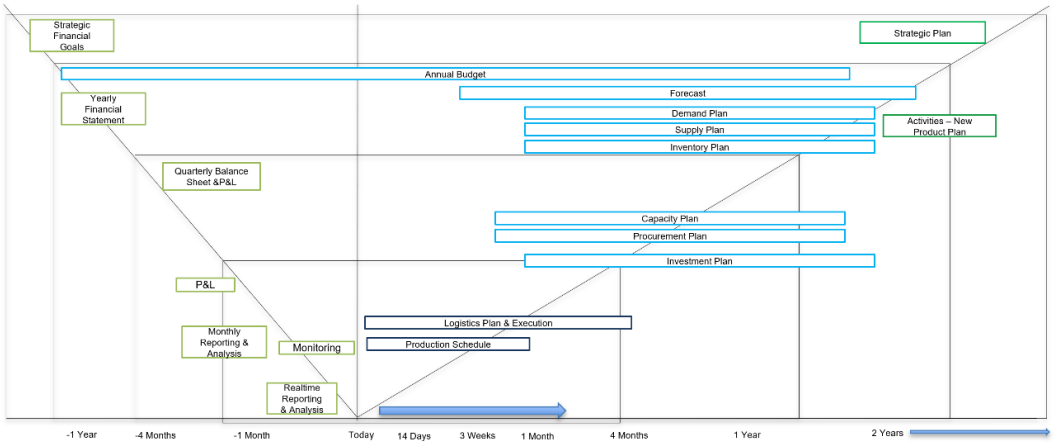

Figure 42: Planning and reporting and monitoring on a timescale

The main goal of a management process can be different from company to company. Most of the time it concerns increasing, shareholder value. For example, an improvement of cost structure can increase long-term shareholder value (reduction of costs of goods sold, process innovation within the production process, and therefore reduced production cost via the scale effect or learning effect, the reduction of R&D costs). The other area is improving asset utilization–optimising fixed assets and reducing working capital. There is also the

option to expand revenue opportunities and enhance the customer value, as relevant areas on how to improve the long-term shareholder value.

When it comes to goal setting, one of the main objectives of a life science company is to increase shareholder value. Figure 43 shows a strategy map with the four perspectives of the balanced scorecard: financial, customer, internal (process), and learning and growth. The shareholder value can be increased by improving the cost structure, e.g., reducing the COGS/COS, and/or reducing the SG&A costs. Another mechanism can be to increase revenue and improve the margin, for instance by reducing the R&D expense in percentage of revenue. Process industries, especially chemical and life science companies, are a mature and highly industrialised and automatised industry sector and use a high volume of equipment and machinery. Therefore, another way to increase shareholder value is by improving capital efficiency – and this means optimising fixed assts and reducing working capital.

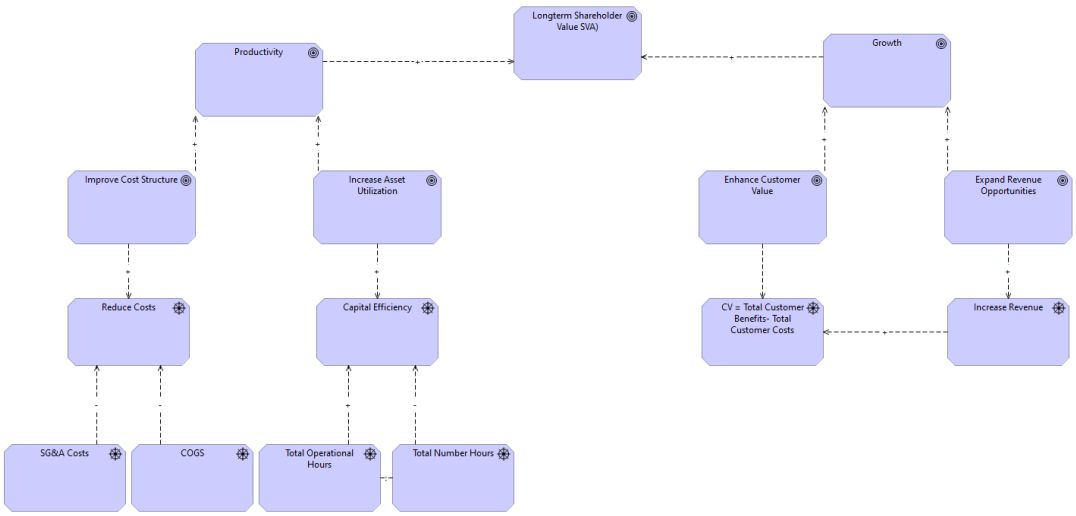

Figure 43: Strategy map with drivers

Figure 43 shows an example of a strategy map with drivers. The main objective in this case is to maximise long-term shareholder value (SVA):

SVA = $NOPAT - CC$ (f5)

Total Cost =$SG\&A + COGS$ (f6)

Asset Utilisation = $\frac{Total\ Operational\ Hours}{Total\ Number\ of\ Hours}$ (f7)

Following the structure of the strategic management process (s. figure 40), the main parts of the planning are:

1. Develop Strategy
    - Input
        - Company mission, values, and vision
    - Main process steps within “Develop Strategy”
        - Strategic analysis – scenario planning technique
    - Deliverables
        - Scenario planning
    - Succeeding process
        - Plan the strategy.
    - Stakeholders
        - Principal/ board of directors
        - Strategic planner
        - Strategic analyst
2. Plan Strategy
    - Input
        - Company mission, values, and vision
        - Selected scenario
    - Main process steps within “Plan Strategy”
        - Strategic analysis – scenario planning technique
    - Deliverables
        - Strategy map
        - Balanced scorecard
    - Succeeding Process
        - Align the organisation
        - Strategic plan
    - Stakeholders

- Principal/ board of directors
- Strategic planner
- Strategic analyst

3. Align the Organisation

- Input
  - Strategic initiatives
  - Select scenario
  - Balanced scorecard - metrics
- Main process steps within “Plan Strategy”
  - Align and structure organisation
- Deliverables
  - Organisational structure
- Succeeding process
  - Plan operations
- Stakeholders
  - Strategic planner
  - Board of directors/managers

4. Plan Operations

- Input
  - Selected scenario
  - Strategy map
  - Balanced scorecard - metrics
- Main process steps within “Plan Strategy”
  - Integrated business planning
- Deliverables
  - Financial budget/plan
  - Sales forecast/demand plan
  - Resource and requirements
  - Supply plan

    - Inventory plan
    - Consensus plan
- Succeeding process
    - Monitor and learn.
- Stakeholders
    - Demand-, supply-, production-, inventory-, financial planner
    - Strategic planner
    - Board of directors/managers

In this process from strategic management and planning towards tactical level – stakeholders are involved (as we can see above – and see below) – Some of these stakeholders are the decision-makers (managers) we mentioned at the beginning of this thesis (see chapter 1). Within the planning process, each of these stakeholders carries out a decision-making process if they are entrusted with one. The model of Simon (2019) [21] gives an overview of how such a decision process works. An AI system that wants to support a decision-maker in making decisions must support the sub-processes of intelligence, design and choice, or be able to explain them if the decision is made automatically. This model follows a four-phase approach. The process first included three major phases, namely intelligence, design, and choice. At a later stage, he added a fourth phase – implementation.

[21] Simon (2019) later was awarded with the Noble Prize for this theory.

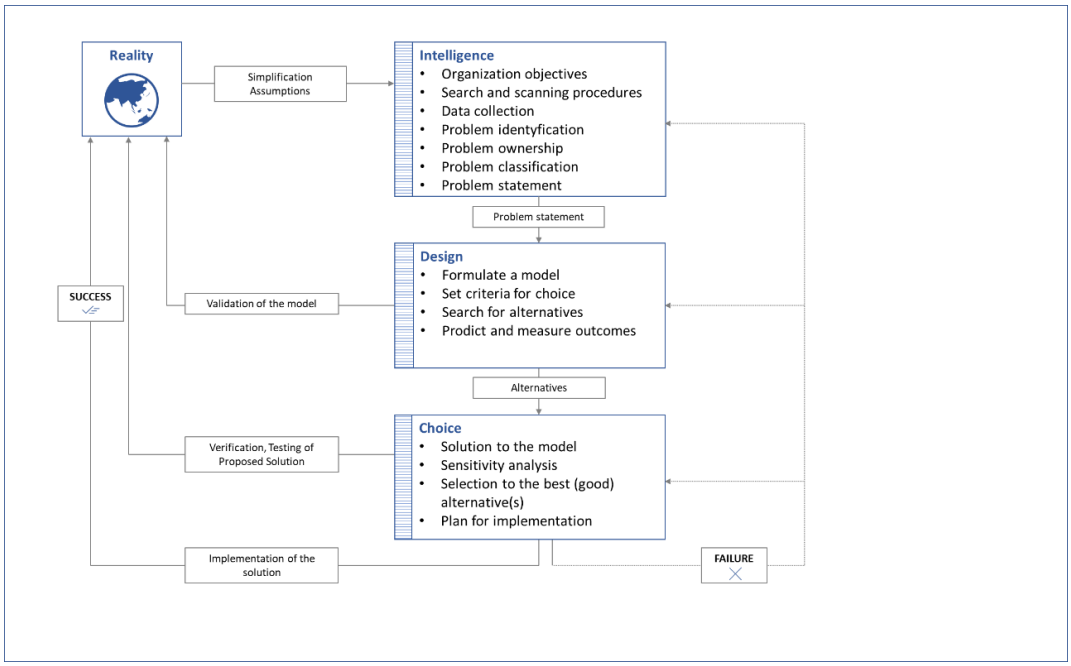

Figure 44: Decision making modelling process – (Simon, 1977; Sharda et al., 2020)

In figure 44, it is shown that there is a continuous flow of activity: intelligence - design - choice. A feedback loop can go back to the previous phase at every stage. Building a model is an essential part of the decision-making process. The feedback loops show the often-non-linear decision-making process, from problem discovery to solution via decision-making. The process starts with the intelligence phase, in which the problem is identified and defined. In the design phase, a model representing the problem is built. To simplify the model, construction assumptions are made.

In the Choice phase, a proposed solution for the model is selected and the solution is then tested. Finally, if the solution makes sense, it is implemented. Successful implementation also results in solving the real problem. A failure leads one to return to an earlier phase in the process, which may occur at any point, as mentioned in the beginning (re. feedback loops). Explanations are especially important during the “intelligence” phase of decision making, when the decision maker seeks the amount, quality, and timeliness of information, to be able to decide. Especially in this phase, models and systems of AI are able to provide significant value to the decision maker. Simon discovered that the limited availability of information causes a significant deterioration of decision quality and effect (Simon, 2019).

(This and the following in anticipation of Chapter 3.3 (see Chapter 3.3), in order to clarify the connections between decisions, explanations and XAI) In order to enhance the understanding of the connections between decisions, explanations, and Explainable AI (XAI), a new research approach called Explainable AI Planning (XAIP) has been developed. The focus of XAIP is to provide explainability in complex planning situations where users interact with AI technologies (especially e.g., with robotics or autonomous vehicles)[22]. This approach is aimed at enhancing trust among end users. Although XAIP mostly applies to robotics, particularly models of autonomous agents moving in environments, it can also be adapted to XAI in business planning with some modifications. Chakraborti et al. (2020) distinguish between the end user (stakeholder, business user, or planner), the domain expert, and the developer, in their approach. An AI model could implement Simon's (1977) decision cycle sub-processes as capabilities for a sub-area of planning, such as scenario planning or tactical planning. For instance, one capability of the AI agent could be to determine the optimal scenario from available options based on strategic KPIs to achieve objectives (intelligence - design - choice).:

$$\delta_{\pi}: C \times\rightarrow S \times \mathbb{R} \qquad \text{(f8)}$$

With C being a set of capabilities of the agent or being available to the agent, S as the set of States and the real number is the cost of making the transition. There is now the planning Algorithm A: Π × τ→ π The planning algorithm solves Π subject to τ (=optimal planning, s.o.)

The plan will now be $\pi = \langle a_i, a_i, \dots, a_n \rangle a_i \in A$, which transforms the current state (any state) I ∈ S of the agent (model) to its goal G ∈ S with

$$\delta\pi(\pi, I) = \langle G, \sum_{a_i \in \pi} c_i \quad \rangle \qquad \text{(f9)}$$

With c(π) being the plan cost

Now in a planning problem, the explainee asks the explainer the questions and the explainer will answer with:

Q.: "Why π?" *or* "$Why\ not\ \pi'$?"

[22] S. also PPDL, MAPL etc.

A.: An explanation Ɛ such that the explainee can verify

$$A: \Pi \times \tau \nrightarrow ; or \qquad \text{(f10)}$$

$$A: \Pi \times \tau \rightarrow \pi' with\ \pi \equiv \pi'\ or\ \pi > \pi' \qquad \text{(f11)}$$

Ɛ, the explanations, can now be classified into three different categories - explanations regarding the algorithm, the model or the plan. Not all types of explanations are relevant for all users (stakeholders), resulting in a matrix of 3 user groups and 4 explanation categories (algorithm-based, model-based with inference and model reconciliation and plan-based explanations). Explanations related to the steps, i.e., algorithm-based questions or answers, and the current status/state of the model or the steps of the plan, are relevant for the developers (see the module called "Tracker" in Chapter 5.2.5 as a solution to this). Model-based explanations are relevant for the end user (business user, planner, etc.) as well as for the domain designer (domain expert).

A user has less computational power than the AI model, so the gap in "understanding" A^H the user can be used to

$$A: \Pi \times \tau \rightarrow \pi \quad \text{and } A^H: \Pi \times \tau \nrightarrow \pi\ ; \qquad \text{(f12)}$$

Therefore, the user (domain expert or business user) seeks for an explanation to close the gap – "inferential reconciliation":

$$A^H: \Pi \times \tau \quad \overset{?}{\rightarrow} \pi \qquad \text{(f13)}$$

To achieve this the user might ask questions like:

Q1: "Why is this action in this plan or why a Ɛ π?"

The explanation here is a causal link chain (s. below – causal inference) or

Q2: "Why not this other plan π'?" (contrastive)

This means that a desired goal (contrastive foil) cannot be achieved but is seen by the user as a constraint - and an explanation is produced to identify an exemplary plan that satisfies these constraints and so show why and how the calculated plan is better. The explanation must therefore reconcile the result of the model and the user's mental model.

$$A: \Pi \times \tau \rightarrow \pi \qquad \text{(f14)}$$

$$\Pi^H + Ɛ \rightarrow \Pi^{H^*} \text{ so that } A: \Pi^{H^*} \times \tau \rightarrow \pi \qquad \text{(f15)}$$

The planning-based explanations show a complete representation of the respective plan as an explanation. In the following, however, we will first describe typical decision variables that can occur in the context of corporate planning (Chakraborti et al., 2020).

As shown in 2.3.2 “Integrated Business Planning”, the term tactical sales and operations planning (S&OP) and the term integrated business planning may be understood to mean cross-functional integrated tactical planning, in a company whose intention is to integrate all different plans (procurement, production, demand, distribution, and financial or budgeting) to get a single plan. The time horizon of the integrated business planning spans from three to eighteen months, so it also covers and supports the annual business planning process at a product family level. The following focus is on the classic four plans within integrated planning, namely procurement, production, distribution, and sales. There will be no differentiation between S&OP and integrated business planning. The financial plan will only be respected where necessary to limit the scope. To integrate the different plans, IT systems must be aligned entirely and support this approach. The integrated solutions will search for an optimal solution and make automated decisions. Particularly when AI models are used, the decision-maker must understand the decisions being made. In their research, Pereira et al. (2020) emphasize that future advanced planning systems must better and more proactively support planning processes because of growing complexity; the authors developed a framework in which they investigate and define which decision parameters (strategic, tactical, operational) have to be made, and on which level, in integrated business planning. The framework is shown in figure 45.

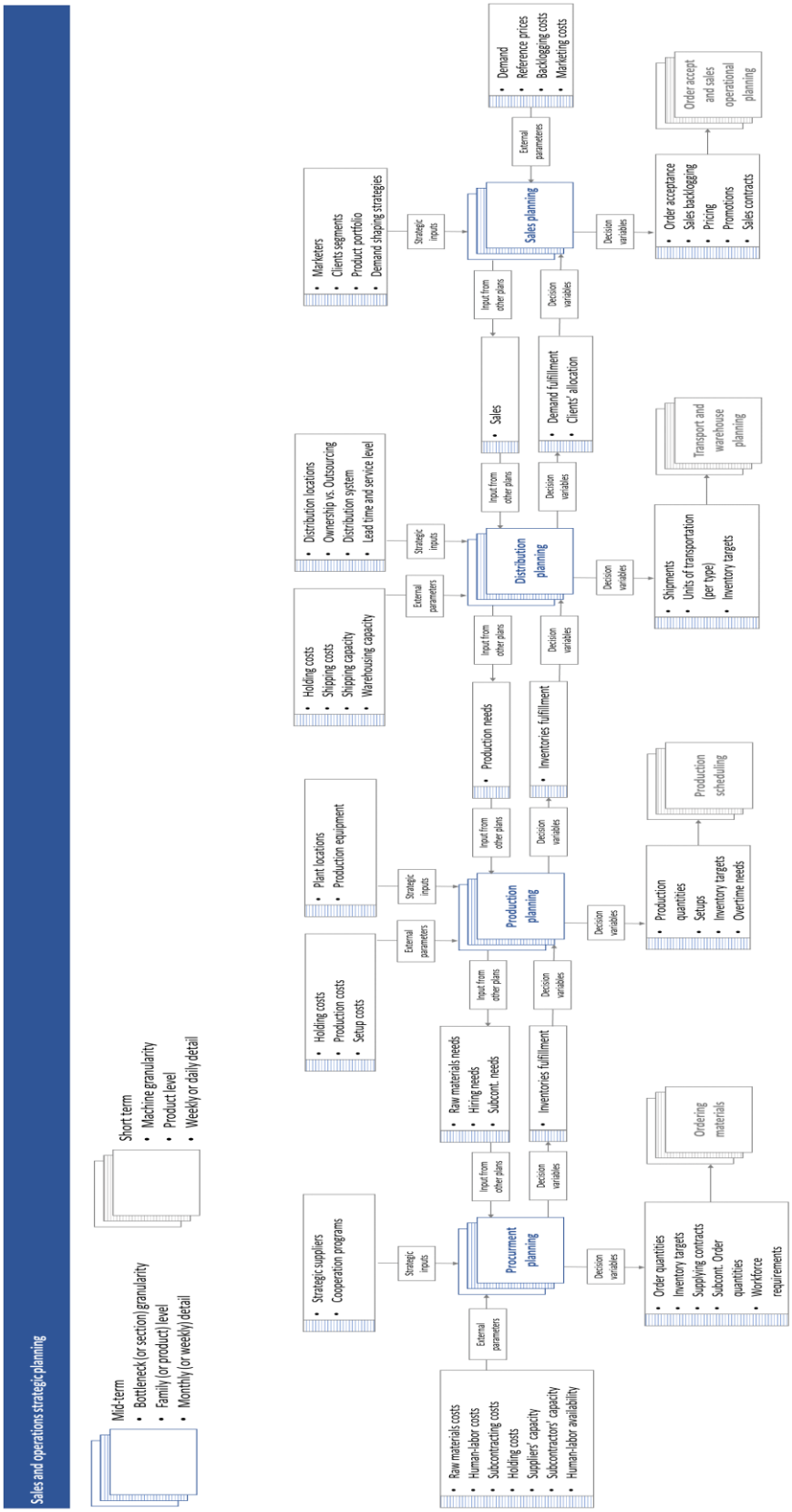

Figure 45: Holistic framework for tactical sales and operations planning- Pereira et al. (2020)

The framework includes the strategic planning (s. above “1. Develop Strategy” and “2. Plan the Strategy”) and on the tactical level “Procurement Planning”, “Production Planning”, Distribution Planning” and “Sales Planning”. The framework differs between decisions (to be made by a specific stakeholder), external parameters, and strategic inputs. On the strategic and tactical level, the granularity of the planning is on bottleneck (for a specific section), family (product family or at least product level) and monthly level (or weekly). On the operational level (ordering materials, production scheduling, transport and warehouse planning and order acceptance and sales operations) the planning granularity is on the machine level, product level and weekly level which take over the planning at the tactical level. The focus of this work is on the mid-term decision variables, external parameters, strategic inputs, and inputs from other plans and on sales planning. For sales planning (s. figure 46), e.g., the identified strategic inputs, which cannot be decided by the planner or stakeholders at the tactical level, these parameters are:

- Markets
- Clients' segments
- Product portfolio
- Demand shaping strategies.

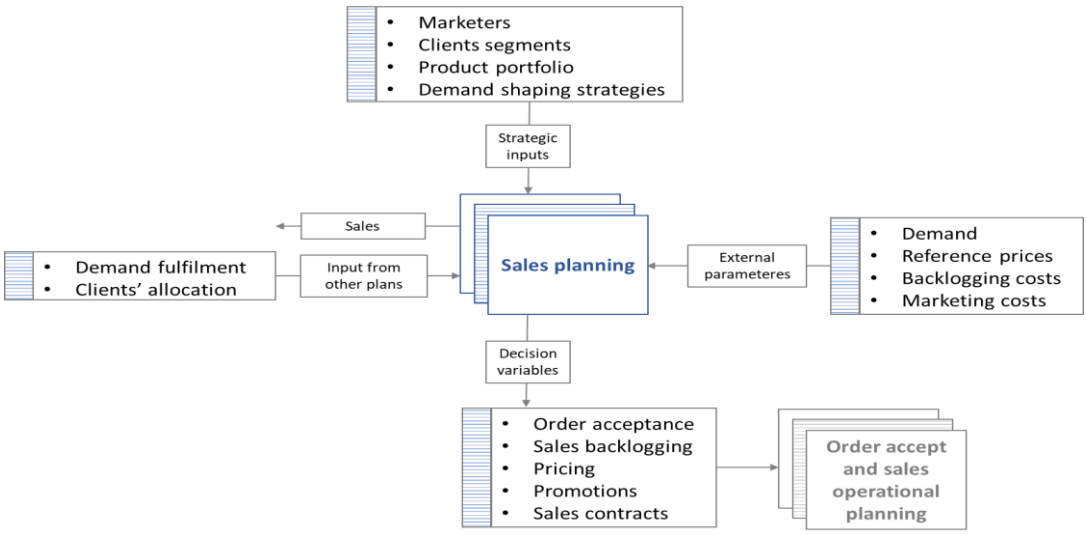

Figure 46: Sales Planning- Pereira et al. (2020)

External parameters are the demand from the market, partially known in advance or predicted based on past sales. Particular knowledge about the market can be seen as expertise.

Here, AI can help by analysing current information about the market and anticipating future demand or demand changes for a specific product. Here, it can be significantly advantageous for a company to rely on a hybrid AI model – with comprehensive market data comprising a knowledge base which is built automatically, adjusted, and enriched by experts. Such a system can detect even "weak signals" of market change and adjust a forecast accordingly (s. Chapter 2.3.1 "Scenario Planning"). Also, reference processes, backlogging, and marketing costs are external in this decision context. To make a decision about the "order acceptance and sales operational planning", other decisions need to have been made about "order acceptance" (meaning those concerning orders, which can already be seen), "sales backlogging", and "pricing" to evaluate the forecast, planned "promotions", and already-signed "sales contracts". That is, before decisions about "order acceptance" and "sales backlogging" are made, others must be made beforehand. These decisions occur when demand fulfilment is not given or when capacity is insufficient (as extension of capacity is a decision which is made on a strategic level as a kind of feedback loop and is therefore a strategic input given to a current plan and cannot be changed). The "order acceptance" is the strategy to accept and fulfil the most crucial customer orders, even accepting penalties when it is not possible to not accept the order. Any decision must take into consideration what is better for the company. "Sales backlogging" is about postponing the sales order given by customers, with the risk that they are not accepting postponement.

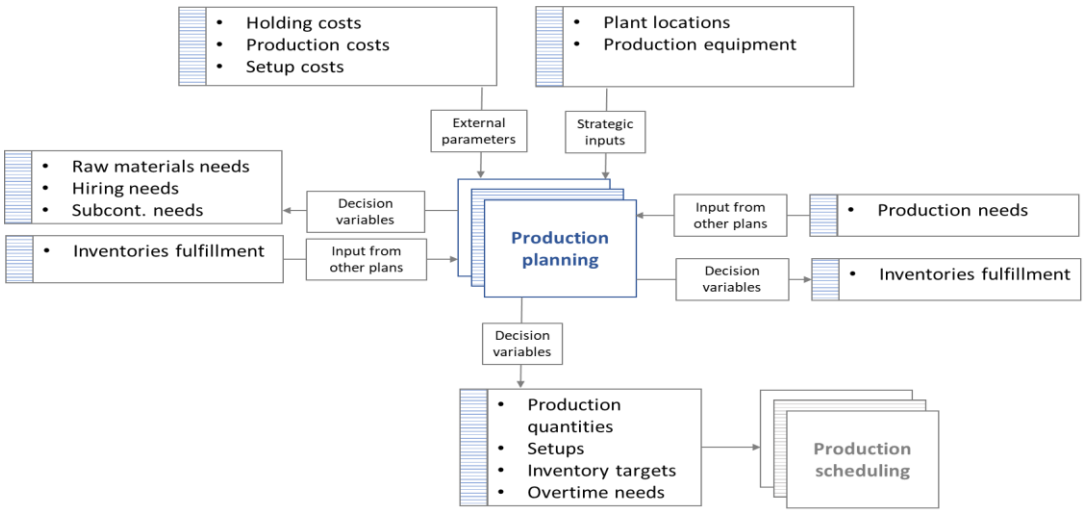

Figure 47: Production planning- Pereira et al. (2020)

External parameters are the holding costs, production costs and setup costs, which are also included here as parameters in the model. The plan defines the production quantity of the

product. If seasonality is relevant, the tactical or mid-term production planning (s. figure 47) must be able to cope with these demand variations. According to Pereira, three strategies can be distinguished. First. Additional quantities can be produced in advance to be able to meet the peak in demand later. secondly. Additional capacities can be made available temporarily through additional workers. Thirdly. Through subcontractors, external suppliers can temporarily provide additional capacity to meet the additional demand. These decisions can be modelled in the variables inventory targets, overtime needs, hiring needs, subcontracting needs. Subcontracting and hiring needs are purchasing activities that require explicit market research by the purchasing department. Raw material requirements can be communicated to the purchasing department directly on the basis of production quantities. Inventory fulfilment translates the requirements of the fulfilment rate of the production requirements. The key decisions in the production area are therefore the production quantities, production to stock and the associated decisions on inventory levels.

Based on the planning horizon, production planning is a multi-period planning problem.(s. figure 47) The planning period in the literature ranges from weeks to 2 years. The planning can be broken down into weekly or monthly time buckets. Daily time buckets are usually linked to operational planning. Usually, the production planner plans at the product level or at least at the aggregated product family level. So, as to reduce the complexity of the model. In a large supply chain network with several production sites, production planning also results in a multi-location planning problem.

Tactical procurement planning (s. figure 48) aims to acquire the most cost-efficient purchasing plan for the required resource from the market in order to meet the needs of production. The production needs, which consist of raw material needs and hiring and subcontracting needs, are the main decision parameters of purchasing planning. Tactical purchasing planning is included in the model as external strategic parameters. These are the strategic suppliers and any existing cooperation programmes.

There are also a couple of external parameters. External costs, for example, for raw materials, human labour, subcontracting and holding and market availability. Such as supplier

capacity, subcontractor capacity and human labour availability. The procurement plan defines the number of raw materials and final product, which will be ordered from the market and the decisions are the order quantity and subcontracting order quantity. It may be necessary that the raw materials have to be stored, therefore inventory targets have to be given. Depending on the industry, it may be necessary to conclude contracts with suppliers. In the labour market, human labour needs make it necessary to bind the required quantity of workers to the company through contracts.

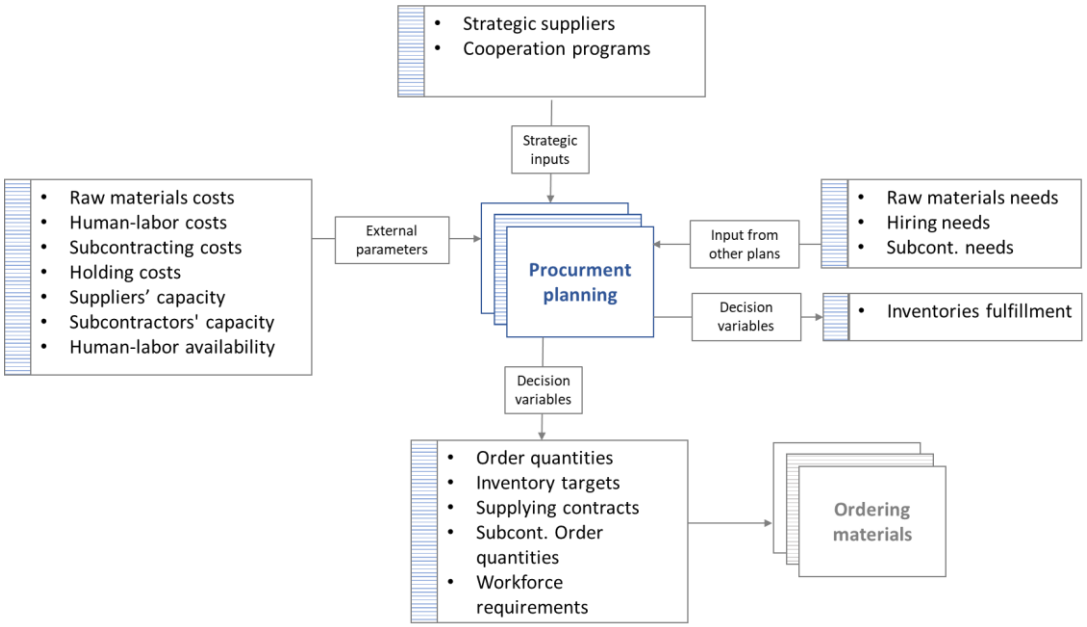

Figure 48: Procurement planning - Pereira et al. (2020)

The decision parameters include order quantities and inventory levels, as well as supplying contracts, subcontractor order quantities and workforce requirements. In many cases, companies stockpile raw materials to compensate for lacks in supply chain capacity, such as the availability of raw materials. Inventory targets are calculated based on production requirements.

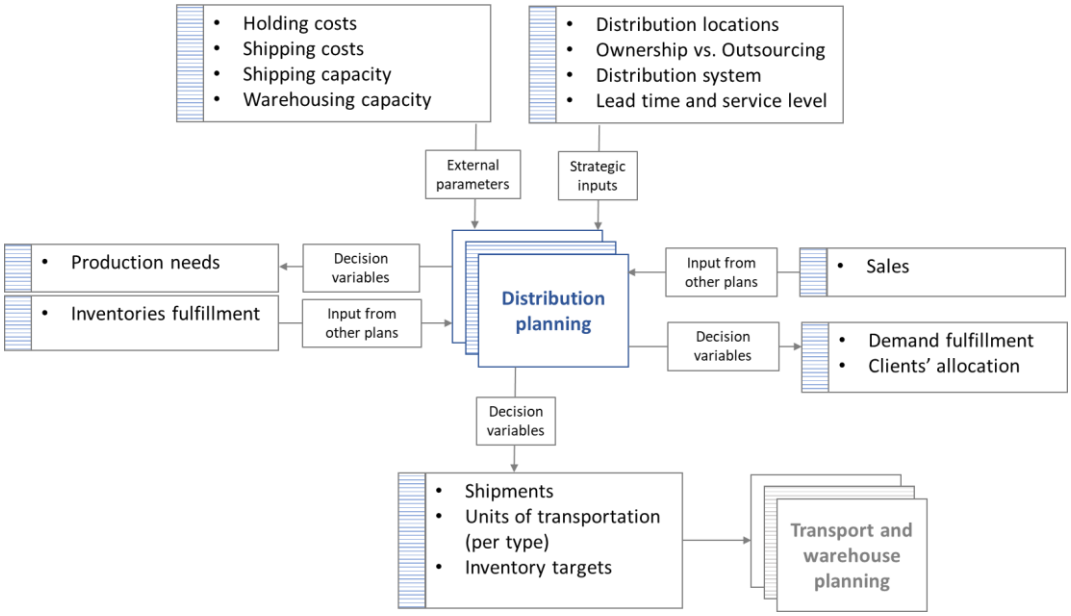

Figure 49: Distribution planning - Pereira et al. (2020)

Tactical - mid-term distribution is to bridge the gap between the production and the customer. Its goal is to fulfil the demand by considering transportation and warehousing capacity with the lowest costs. External information respective parameters are- holding costs, shipping costs, shipping capacity and warehousing capacity - and from the strategic planning distribution (s. figure 49) locations, distribution system and intended lead time and service level are inputs to the model. If the demand is not fulfilled after the planning, the distribution planner must negotiate with the sales planner. The distribution strategy is about the balancing between ownership or outsourcing via 3PL. Decision variables are demand fulfilment and client allocation, the shipping quantities are to ship the number of finished products in a given environment like the design of the supply chain network, the distribution system. Main goal is to optimize efficiency and costs.

Distribution planning also comprises transportation requirements and transportation modes, like e.g., route planning.

The whole model of strategic management and planning- scenario planning and integrated corporate planning was used to illustrate the various decision variables. this serves to identify the stakeholders of corporate planning and their requirements.

2.3.4 Stakeholders in Corporate Planning in the Process Industry

The stakeholders, which in (software) architecture are seen “an individual, team, or organisation (or classes thereof), with interests in (or concerns relative to) a system.” (Lankhorst, 2017), of the corporate planning process in the process industry are shown in the tables 7-10 below.

Stakeholders of corporate planning are:

- Society, supranational/ global or regional government, regulators- auditors
- Owner
- Board of directors/executive managers
- Strategic Plan:
    - strategic planner
- Demand Plan:
    - demand planner (marketing planner)
    - distribution planner
- Supply Plan:
    - production planner,
    - inventory planner

If one follows the presentation by Bejger/Elster (2020), one recognises that further stakeholders can be outside the company, such as the owner of the company, the regulator, the auditors, the regional or country-related social society (e.g., society in Germany) and the supranational society (e.g., the European Union).

ID	Stakeholder Group	Stakeholder	Domain	Type	Decision	Input for plan	Input from plan	Deliverable	Level	Industry remarks	Impacted Stakeholder	Description	Decision specification	AI Method	Explanation Method	Level of Causal hierarchy	Explanation Type
SI1			Sales Planning	Strategic Input	Markets		Strategic plan	Sales plan			Demand/ Sales Planner	Markets					
SI2			Sales Planning	Strategic Input	Clients segments		Strategic plan	Sales plan			Demand/ Sales Planner	Clients segments					
SI3			Sales Planning	Strategic Input	Products portfolio		Strategic plan	Sales plan			Demand/ Sales Planner	Products portfolio					
SI4			Sales Planning	Strategic Input	Demand shaping strategies		Strategic plan	Sales plan			Demand/ Sales Planner	Demand shaping strategies					
EP1			Sales Planning	External parameters	Demand	Distribution planning	Sales Planning Forecast	Sales plan			Demand/ Sales Planner	Demand					
EP2			Sales Planning	External parameters	Reference prices	Distribution planning		Sales plan			Distribution Planner	Reference prices					
EP3			Sales Planning	External parameters	Backlogging costs	Distribution planning		Sales plan			Distribution Planner	Backlogging costs					
EP4			Sales Planning	External parameters	Marketing costs	Distribution planning		Sales plan			Distribution Planner	Marketing costs					
DE1	Business User	Demand planner	Sales Planning	Decisions	Order acceptance	Sales operational planning		Sales plan	At customer level (partial aggregation at regional level, global aggregation)	Possible strategy in food industry	Customer	When capacity is insufficient, it is the planning to decide, which impacts the supply chain. It can also be used to decide on the overall customer value. Can be combined with DE2	Minmze the cost and negative impact, when capacity is insufficient.	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (F-IT), Dynamask (2) LIME, SHAP	1,2,3	Association, Intervention, Counterfactual
DE2	Business User	Demand planner	Sales Planning	Decisions	Sales backlogging	Sales operational planning		Sales plan	At customer level (partial aggregation at regional level, global aggregation)	Not possible in food industry - process industry with harvesting seasonality	Customer	Postponement of sales orders- can be combined with DE1	Minmze the cost and negative impact, when capacity is insufficient.	Forecasting Use Case VII	TimeSHAP, Instance-wise Feature Importance in Time (F-IT), Dynamask	1,2,3	Association, Intervention, Counterfactual
	Business User	Demand planner	Sales Planning	Decisions	Pricing	Sales operational planning		Sales plan	Pricing can be used to optimize capacity. When possible (contracting, esp. b2b), the price can change dynamically from period to period. Demand leakage between product and markets can be worth to consider.	Less valid in b2b supply networks.	Customer	Decision is how to set up pricing, for specific product and/ or certain amount and time	Optimize pricing- product and time related, so that the maximum profit is reached for a specific period	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (F-IT), Dynamask (2) LIME, SHAP	3	Counterfactuals
DE4	Business User	Demand planner	Sales Planning	Decisions	Promotions	Sales operational planning		Sales plan	Promotions, specific for products and specific in duration can be also a method to optimz capacity. This strategy is more valuable, if the company owns the whole supply chain to the point of sales. Less valid in b2b supply networks	Less valid in b2b supply networks.	Customer	Decision is how to set up pricing, for specific product and/ or certain amount and time. Consider- new activities (new product launches) and product lifcycle	Optimze promotion- product and time related, so that the maximum profit is reached, consider new product launches and profitability in mid term.	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (F-IT), Dynamask (2) LIME, SHAP	3	Counterfactuals
DE5	Business User	Demand planner	Sales Planning	Decisions	Sales contracts	Sales operational planning		Sales plan	In a b2b environment sales contract are usual. However, they block a certain amount of capacity, which could be used for more profitable orders.	Usual in b2b environment, like process industry	Customer	Decide the optimized mixture of sales contracts and spot business (higher risk)	Check'on historic data, which customer, depending on customer value with no contract could be offered a contract and which existing contract should be cancelled	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (F-IT), Dynamask (2) LIME, SHAP	3	Counterfactuals
SI5	Business User	Production/ Supply planner	Production Planning	Strategic Input	Plant locations		Strategic plan	Production plan			Production Supply planner	Plant locations					
SI6	Business User	Production/ Supply planner	Production Planning	Strategic Input	Production equipment		Strategic plan	Production plan			Production Supply planner	Production equipment					
EP5	Business User	Production/ Supply planner	Production Planning	External parameter	Holding costs			Production plan			Production Supply planner	Holding costs					
EP6	Business User	Production/ Supply planner	Production Planning	External parameter	Production costs			Production plan			Production Supply planner	Production costs					
EP7	Business User	Production/ Supply planner	Production Planning	External parameter	Setup costs			Production plan			Production Supply planner	Setup costs					

Table 7: Stakeholder Map A – Model and Decision Variables - Part I

ID	Stakeholder Group	Stakeholder	Domain	Type	Decision	Input for plan	Inputs from plan	P-Interface	Level	Industry remarks	Involved Stakeholder	Description	Decision specification	Method	Explanation Method	Level of Causal Hierarchy	Explained by Type
S11	Business User	Strategic Planner	Sales Planning	Strategic Input/ Decision	Markets		Strategic plan	Sales plan			Demand/ Sales Planner	Markets		(1) Forecasting with Knowledge Base (Knowledge Graph) (2) Scenario Application (AISOP, SPA etc.)	(1) TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask, Neuro-symbolic (2) LIME, SHAP, Neuro-symbolic model	1,2,3	Association, Intervention, Counterfactual
S12	Business User	Strategic Planner	Sales Planning	Strategic Input/ Decision	Clients segments		Strategic plan	Sales plan			Demand/ Sales Planner	Clients segments		(1) Forecasting with Knowledge Base (Knowledge Graph) (2) Scenario Application (AISOP, SPA etc.)	(1) TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask, Neuro-symbolic (2) LIME, SHAP, Neuro-symbolic model	1,2,3	Association, Intervention, Counterfactual
S13	Business User	Strategic Planner	Sales Planning	Strategic Input/ Decision	Products portfolio		Strategic plan	Sales plan			Demand/ Sales Planner	Products portfolio		(1) Forecasting with Knowledge Base (Knowledge Graph) (2) Scenario Application (AISOP, SPA etc.)	(1) TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask, Neuro-symbolic (2) LIME, SHAP, Neuro-symbolic model	1,2,3	Association, Intervention, Counterfactual
S14	Business User	Strategic Planner	Sales Planning	Strategic Input/ Decision	Demand shaping strategies		Strategic plan	Sales plan			Demand/ Sales Planner	Demand shaping strategies		(1) Forecasting with Knowledge Base (Knowledge Graph) (2) Scenario Application (AISOP, SPA etc.)	(1) TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask, Neuro-symbolic (2) LIME, SHAP, Neuro-symbolic model	1,2,3	Association, Intervention, Counterfactual
EP1	Business User	Strategic Planner	Sales Planning	External parameters	Demand	Distribution planning	Sales Planning Forecast	Sales plan			Demand/ Sales Planner	Demand					
EP2	Business User	Strategic Planner	Sales Planning	External parameters	Reference prices	Distribution planning		Sales plan			Distribution Planner	Reference prices					
EP3	Business User	Strategic Planner	Sales Planning	External parameters	Backlogging costs	Distribution planning		Sales plan			Distribution Planner	Backlogging costs					
EP4	Business User	Strategic Planner	Sales Planning	External parameters	Marketing costs	Distribution planning		Sales plan			Distribution Planner	Marketing costs					
DE1	Business User	Demand planner	Sales Planning	Decisions	Order acceptance	Sales operational planning		Sales plan	At customer level (partial aggregation at regional level, global aggregation)	Possible strategy in food industry	Customer	When capacity is insufficient, it is the planning to decide, which impacts the supply chain. It can also be used to weigh profits against overall customer value. Can be combined with DE2	Mimize the cost and negative impact when capacity is insufficient.	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	1,2,3	Association, Intervention, Counterfactual
DE2	Business User	Demand planner	Sales Planning	Decisions	Sales backlogging	Sales operational planning		Sales plan	At customer level (partial aggregation at regional level, global aggregation)	Not possible in food industry - process industry with harvesting season/ally	Customer	Postponement of sales orders- can be combined with DE1	Mimize the cost and negative impact when capacity is insufficient.	Forecasting Use Case VII	TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask	1,2,3	Association, Intervention, Counterfactual
	Business User	Demand planner	Sales Planning	Decisions	Pricing	Sales operational planning		Sales plan	Pricing can be used to optimize capacity. When possible (contracting, esp. b2b), the price can change dynamically from period to period. Demand leakage between product and markets can be worth to consider	Less valid in b2b supply networks.	Customer	Decision is how to set up pricing for specific product and/ or certain amount and time	Optimize pricing- product and time related, so that the maximum profit is reached, for a specific period	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	5	Counterfactuals
DE4	Business User	Demand planner	Sales Planning	Decisions	Promotions	Sales operational planning		Sales plan	Promotions, specific for products and specific in duration can be also a method to optimize capacity. This is very, very valuable, if the company owns the whole supply chain to the point of sales. Less valid in b2b supply networks	Less valid in b2b supply networks.	Customer	Decision is how to set up pricing for specific product and/ or certain amount and time. Consider - new activities (new product launch), existing product lifecycle	Optimize promotion- product and time related, so that the maximum profit is reached, consider new product introduction and profitability in mid term.	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	5	Counterfactuals
DE5	Business User	Demand planner	Sales Planning	Decisions	Sales contracts	Sales operational planning		Sales plan	Unit in b2b environment, like contract are usual. However, they block a certain amount of capacity, which could be used for more profitable orders.	Usual in b2b environment, like process industry	Customer	Decide the optimized mixture of sales contracts and spot business (higher risk)	Check on historic data, which customer, depending on customer value with no contract could be offered a contract and which existing contract should be canceled	(1) Forecasting Use Case VII, (2) Reinforcement Learning Net	(1) TimeSHAP, instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	5	Counterfactuals
S15	Business User	Production/ Supply planner	Production Planning	Strategic Input	Plant locations		Strategic plan	Production plan			Production/ Supply planner	Plant locations					
S16	Business User	Production/ Supply planner	Production Planning	Strategic Input	Production equipment		Strategic plan	Production plan			Production/ Supply planner	Production equipment					
EP5	Business User	Supply planner	Production Planning	External parameter	Holding costs			Production plan			Production/ Supply planner	Holding costs					
EP6	Business User	Production/ Supply planner	Production Planning	External parameter	Production costs			Production plan			Production/ Supply planner	Production costs					
EP7	Business User	Production/ Supply planner	Production Planning	External parameter	Setup costs			Production plan			Production/ Supply planner	Setup costs					

Table 8: Stakeholder Map A – Model and Decision Variables - Part II

Table 9: Stakeholder Map A – Model and Decision Variables - Part III

ID	Stakeholder Group	Stakeholder	Domain	Type	Decision	Input for plan	Input from plan	Deliverable	Level	Industry remarks	Impacted Stakeholder	Description	Decision/ specification	AI Method	Explanation Method	Level of Causal hierarchy	Explanation Type
DE19	Business User	Procurement planner	Procurement Planning	Decisions	Workforce requirements	Ordering materials		Procurement plan	Multi- item level, multi supplier and origin			Amount of workers to hire or to dismiss	Costs, availability	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	3	Counterfactuals
EP8	Business User	Procurement planner	Procurement Planning	External parameter	Raw material costs			Procurement plan				Raw material costs					
EP9	Business User	Procurement planner	Procurement Planning	External parameter	Human-labor costs			Procurement plan				Human-labor costs					
EP10	Business User	Procurement planner	Procurement Planning	External parameter	Subcontracting costs			Procurement plan				Subcontracting costs					
EP11	Business User	Procurement planner	Procurement Planning	External parameter	Holding costs			Procurement plan				Holding costs					
EP12	Business User	Procurement planner	Procurement Planning	External parameter	Supplier's capacity			Procurement plan				Supplier's capacity					
EP13	Business User	Procurement planner	Procurement Planning	External parameter	Subcontractor capacity			Procurement plan				Subcontractor capacity					
EP14	Business User	Procurement planner	Procurement Planning	External parameter	Human-labor availability			Procurement plan				Human-labor availability					
SI9	Business User	Distribution planner	Distribution Planning	Strategic Input	Distribution locations	Transport and warehouse planning		Distribution plan				Bridging the gap between production and the clients - fulfillment of estimated demand considering transportation and warehousing capacity while minimizing costs					
SI10	Business User	Distribution planner	Distribution Planning	Strategic Input	Ownership vs. outsourcing	Transport and warehouse planning		Distribution plan				Transportation system ownership or outsourcing to 3rd party					
SI11	Business User	Distribution planner	Distribution Planning	Strategic Input	Distribution system	Transport and warehouse planning		Distribution plan				Distribution system					
SI12	Business User	Distribution planner	Distribution Planning	Strategic Input	Lead time and service level	Transport and warehouse planning		Distribution plan				Lead time and service level					
IP5	Business User	Distribution planner	Distribution Planning	Input from other plan	Sales		Sales plan	Distribution plan				Sales					
DE20	Business User	Distribution planner	Distribution Planning	Decisions	Demand fulfillment	Transport and warehouse planning		Distribution plan			Customer	If the (sales) plan is not fulfilled completely it must be again negotiated with the sales team	Customer order, fullfilment, delivery time	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	3	Counterfactuals
DE21	Business User	Distribution planner	Distribution Planning	Decisions	Client's allocation	Transport and warehouse planning		Distribution plan			Customer	Allocation to specific distribution centers	Customer order, fullfilment, delivery time	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	3	Counterfactuals
DE22	Business User	Distribution planner	Distribution Planning	Decisions	Shipments	Transport and warehouse planning		Distribution plan					Customer order, fullfilment, delivery time	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	3	Counterfactuals
DE23	Business User	Distribution planner	Distribution Planning	Input from other plan	Units of transportation (per type)	Transport and warehouse planning		Distribution plan		Shelf life of pharmaceutical products		Referring to the amount of final products needed to be transported	Decision about stock level and also regards or under consideration of capacity (shipment or storage)- Transshipment flows			1,2,3	Association, Intervention, Counterfactual
DE24	Business User	Distribution planner	Distribution Planning	Decisions	Inventory targets	Transport and warehouse planning		Distribution plan		Shelf life of pharmaceutical products		Inventory level of distibution centers and retail units- Distribution center as being a consolidation point- capacity of shipment and storage	Decision about stock level and also regards or under consideration of capacity (shipment or storage)	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	3	Counterfactuals
DE25	Business User	Distribution planner	Distribution Planning	Decisions	Production needs	Transport and warehouse planning		Distribution plan				Production needs	Costs, availability	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP, Instance-wise Feature Importance in Time (FIT), Dynamask (2) LIME, SHAP	3	Counterfactuals

ID	Stakeholder Group	Stakeholder	Domain	Type	Decision	Input for plan	Input from plan	Deliverable	Level	Industry remarks	Impacted Stakeholder	Description	Decision specification	AI Method	Explanation Method	Level of Causal Hierarchy	Explanation Type
IP6	Business User	Distribution planner	Distribution Planning	Input from other plan	Inventories fulfilment	Transport and warehouse planning		Distribution plan				Inventories fulfilment					
EP15	Business User	Distribution planner	Distribution Planning	External parameter	Holding costs	Transport and warehouse planning		Distribution plan				Holding costs					
EP16	Business User	Distribution planner	Distribution Planning	External parameter	Shipping costs	Transport and warehouse planning		Distribution plan				Shipping costs					
EP17	Business User	Distribution planner	Distribution Planning	External parameter	Shipping capacity	Transport and warehouse planning		Distribution plan				Shipping capacity					
EP18	Business User	Distribution planner	Distribution Planning	External parameter	Warehousing capacity	Transport and warehouse planning		Distribution plan				Warehousing capacity					
SI13	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Strategic input	Strategic- tactical financial goals	Financial Plan		Financial Plan/ Budget			All other planner, management board, board of directors	Input from strategic planning, balanced scorecard					
SI14	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Strategic input	Strategic scenario - scenario parameters selected	Financial Plan	KPI - Balanced Scorecard	Financial Plan/ Budget			All other planner, management board, board of directors	Input from scenario analyss and selected scenario					
SI15	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Strategic input	Capex projects planned	Financial Plan	Capex plan	Financial Plan/ Budget		s. business model of process industry - chapter 2	All other planner, management board, board of directors	Input from strategic planning, investment projects and intiatives					
SI16	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Strategic input	M&A projects planed	Financial Plan	M&A plan	Financial Plan/ Budget		s. business model of process industry - chapter 3	All other planner, management board, board of directors	Input from strategic planning- M&A intiatives					
IP7	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Input from other plan	New products planes. hist high level sales plan	Financial Plan	New products plan	Financial Plan/ Budget		s. business model of process industry - chapter 4	All other planner, management board, board of directors	Input from R&D and marketing, sales about introducing new products within tactical time span					
IP8	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Input from other plan	Historic plan with assumptions and adjustments	Financial Plan	Hestoric plan	Financial Plan/ Budget			All other planner, management board, board of directors	Use historic plan with adjustments derived from the strategic plan					
IP9	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Input from other plan	Headcount etc	Financial Plan		Financial Plan/ Budget			All other planner, management board, board of directors	Use historic plan with adjustments derived from the strategic plan, aligned with structural planning					
DE26	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Decisions	Identify and close gaps (alternatives)	All other		Financial Plan/ Budget			All other planner, management board, board of directors	Identify any gaps between the KPI specifications and the results from first financial plan	Identify gaps and decide on possible strategies to close	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP. Instance-wise Feature Importance in Time (FIT). Dynamask (2) LIME. SHAP	3	Counterfactuals
DE27	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Decisions	Forecast and simulate Financial Plan	All other		Financial Plan/ Budget			All other planner, management board, board of directors	Forecast and adjust plan according to assumptions and strategic plan	Simulate and decide on appropriate forecast	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP. Instance-wise Feature Importance in Time (FIT). Dynamask (2) LIME. SHAP	3	Counterfactuals
DE28	Business User	Finance/ Financial Planer	Financial Plan/ Budget	Decisions	Decide on Financial Plan Budget for Integrated Business Plan (S&OP)	All other		Financial Plan/ Budget			All other planner, management board, board of directors	Decide on one version of financial plan and provide it to the other plans (integrated business planning)	Decide on financial plan/ budget to start within integrated business planning (consensus plan)	(1) Forecasting Use Case VII (2) Reinforcement Learning Net	(1) TimeSHAP. Instance-wise Feature Importance in Time (FIT). Dynamask (2) LIME. SHAP	3	Counterfactuals
EP19	Business User	Finance/ Financial Planer	Financial Plan/ Budget	External parameter	Cots - structural costs. which are not changeable in tactical perspective	All other		Financial Plan/ Budget			All other planner, management board, board of directors	All costs that are not to be changed within tactical time span					
EP20	Business User	Finance/ Financial Planer	Financial Plan/ Budget	External parameter	Market costs. taxes etc.	All other		Financial Plan/ Budget			All other planner, management board, board of directors	External parameters not to be changed within scope					

Table 10: Stakeholder Map A – Model and Decision Variables - Part IV

The table 7-10 shows stakeholder map- The Id of the stakeholder, the stakeholder group, stakeholder (this will be needed later for the requirements derived from usage of (X)AI). The domain, what kind of type the respected stakeholder requirement is, e.g., strategic input, external parameter etc. The decision, which the stakeholder is supposed to do, input from/ for plan, the deliverable, which level the decision has to be made, e.g., product vs. product group etc., the impacted stakeholder, a general description the specification of the decision and the proposed explanation type- when the decision is made by AI. The idea is here that the AI has to provide the decision like the human and therefore must give the explanation in the same manner as the human.

2.4 Information Systems to Support Planning and Decision-Making in the Process Industry

As already pointed out, companies in the process industry are highly integrated in complex supply networks. Such a supply chain or network consists of many suppliers, manufacturing plants (those which are owned or subcontracted plants), distribution centres, and customer sites.

When the planning starts, it could be a straightforward way to search for alternatives, though this could lead to certain issues. For instance, there could be conflicting objectives and ambiguous preferences among the different alternatives. An example could be providing customer service with high levels of stock, while at the same time minimising inventory to optimise working capital. Both objectives cannot be reached at the same time. Such issues are known to be multi-objective decision problems, which can be solved by setting the objective to a minimum or maximum satisfaction level for each goal, except for one that will be optimised, e.g., to optimise customer service at a set (satisfactory level) of inventory.

Objectives can be calculated with prices or scores and be optimisable. The issue when doing so is the selection and usage of the weights, as they could influence the optimisation problem.

Advanced Planning Systems are being used to deal with planning problems such as those just mentioned and are able to provide all of the necessary functionality for solving such

multi-objective planning problems while calculating large numbers of different alternatives. Planning systems include data about the future and therefore deal with uncertainty (s. Chapter 2.3.1 Scenario Planning). Even when forecast models estimate the data, there is still an error – even more so when the surrounding ecosystem has changed and is not reflected in the model (concept drift) or the historical data is not useable anymore (or is biased, etc.).

APS systems are using data which is based on an ERP (Enterprise Resource Planning) system; there are three main points which can characterise these systems:

1. They provide an integral planning of the entire supply chain, at least from the suppliers up to the customers of a single enterprise, or even of a more comprehensive network of enterprises.

2. They provide a true optimisation by properly defining alternatives, objectives, and constraints for the various planning problems and by using optimising methods of planning, either exact ones or heuristics (see, e.g., Fleischmann & Meyr, 2003)

3. They are a hierarchical-planning systems (see e.g., Schneeweiss, 2003)

The first generation of APS, like SAP APO (Advanced Planning and Optimisation), are now being replaced by a more modular approach. Parts of the APO solution are now provided by the S/4HANA ERP system (e.g., ATP or aATP – available to promise and advanced available to promise), and parts are provided by a cloud-based solution SAP IBP (product name is “Integrated Business Planning”). The new ERP solution S/4HANA, as well as the planning solution, are enhanced by Artificial Intelligence capabilities to automatise the whole planning process, or to provide better forecasting methods to deal with the issues mentioned above (Markin, 2021).

2.5 Classical Decision Support Systems, Business Analytics, Data Science and Reporting

Business Intelligence (short BI) is an umbrella term including and combining (IT) architectures, tools, databases, analytical tools, applications, and methodologies. There are also other buzzwords used with BI, such as Business Performance Management (BPM), which is used by software vendors to differentiate their offers from those of competitors. The main objective of BI is to enable stakeholders (users) to obtain interactive access, in real-

time, if necessary (depending on the requirements), to data which may be manipulated to provide business managers and analysts the means to carry out appropriate evaluations.

The idea of Business Intelligence (short BI, first time used by Luhn (1958) for a document-centred business intelligence - BI system) is that decision-makers obtain valuable insights by analysing historical and current data, situations, and performances, which enable them to make better and more informed decisions. From an end-to-end perspective, the BI process is based on acquiring data, transforming it into information, then into decisions, and finally bringing about actions. Business Intelligence therefore supports the "intelligence" phase in the decision model from Simon (2019) (s. Chapter 2.3.3).

The above processes should prevent managers from making decisions based on 'gut feelings' or on decisions made successfully in other situations that might not fit the current one. Alternately, the managers may see decision-making as a skill which must be acquired after years of studying, or trial and error using intuition, and not a process which is based on information. It is more important to emphasise methodical, thoughtful, analytical decision-making rather than flashiness and interpersonal communication skills.

Since the 1960/ 1970ties systems to support decision-making were built and called DSS (Decision Support Systems). Since then, these are typically built to support the solution of a certain problem or to evaluate an opportunity.

While a BI system is built to monitor situations and identify problems or opportunities, it is up to the user to further investigate the specific problem and apply analytical methods. BI systems provide models and data access; DSSs have their own databases built to solve a specific problem or set of problems (Sharda et al, 2019).

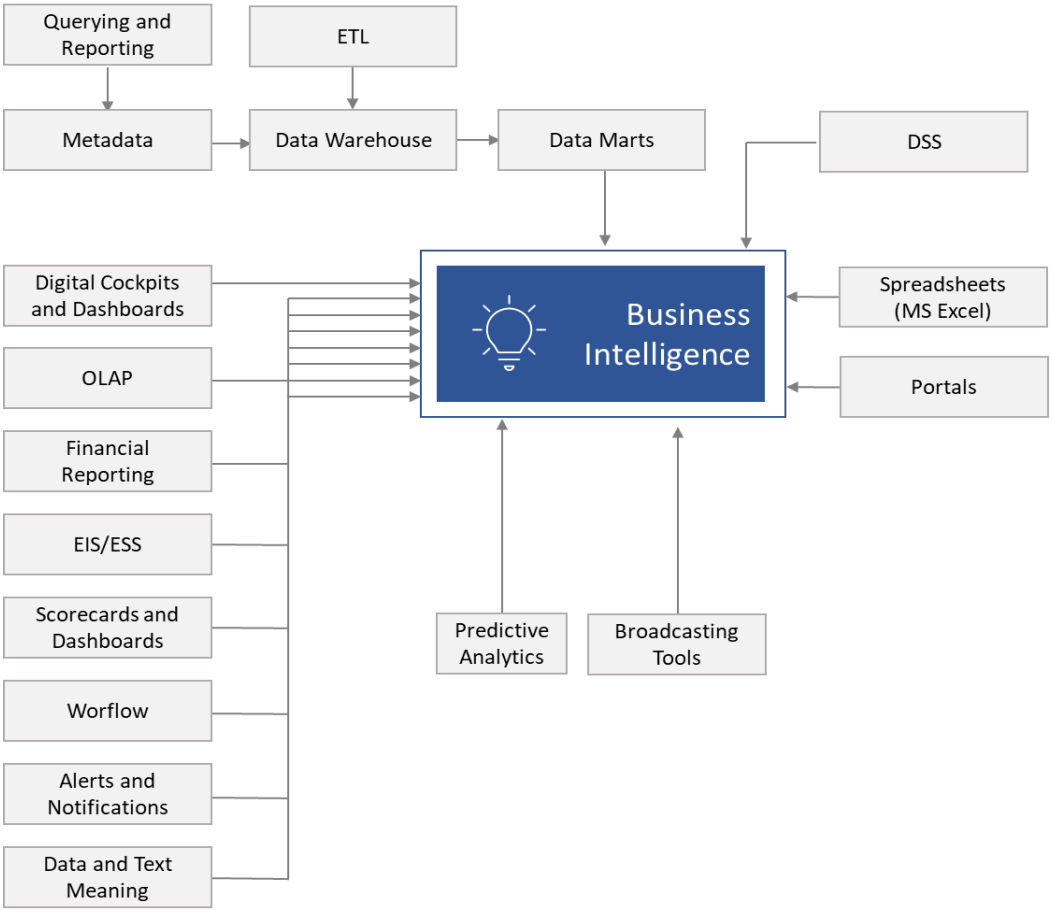

Figure 50: Business Intelligence and its neighbouring disciplines- Sharda et al. (2019)

In figure 50 business intelligence and its neighbouring disciplines are shown. It presents the Data Warehouse concept with ETL (Extract Transform and Load) the Data Marts, etc. EIS/ESS – Executive Information Systems, Support Systems, Data and Text Mining, a predecessor of Data Science and the DSS (Decision Support Systems).

2.5.1 Classical Decision Support Systems

While BI systems are somehow problem-agnostic systems, which can be used for various solutions, a DSS system is a system built to support a specific solution. A BI system is usually dependent on a database, like a data warehouse, while a DSS is using its own database. Formally, a DSS is an approach (or methodology) for supporting decision making. Therefore, following Sharda et al. (2019, p.16/17), "[A DSS] …uses an interactive, flexi-

ble, adaptable computer-based information system (CBIS) especially developed for supporting the solution to a specific unstructured management problem. It uses data, provides an easy user interface, and can incorporate the decision maker's own insights. In addition, a DSS includes models and is developed (possibly by end users) through an interactive and iterative process. It can support all phases of decision making and may include a knowledge component. Finally, a DSS can be used by a single user or can be Web based for use by many people at several locations."

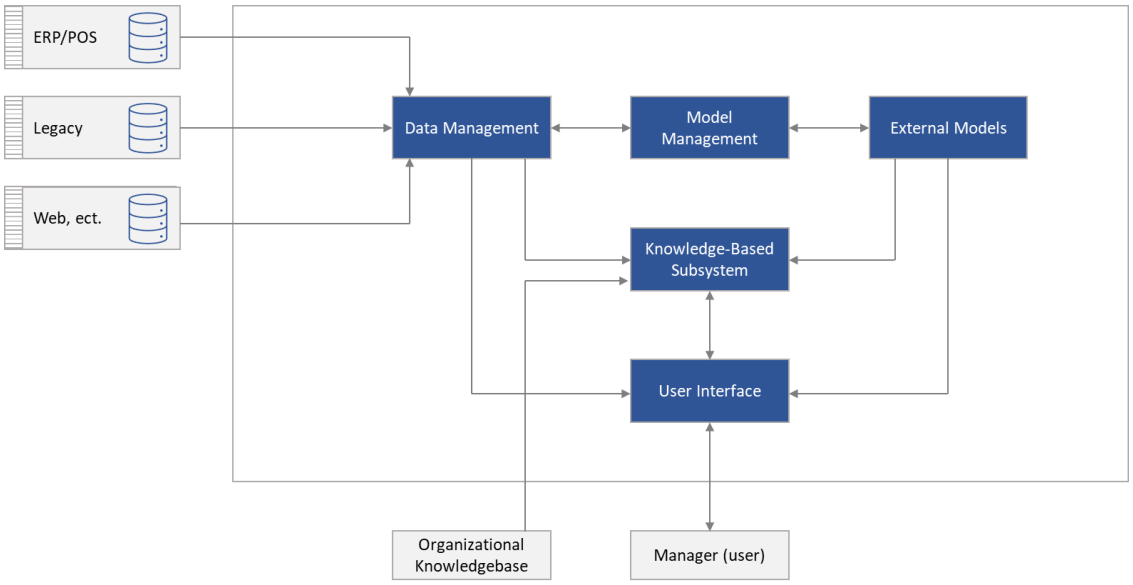

Figure 51: Typical architecture of a DSS system

Figure 51 shows a typical architecture of a "classic" DSS system. The system consists of a data management component to gather data from ERP/POS, legacy systems, or web data. It also includes model management and external models, which are used as a decision-making model. There are also knowledge-based subsystems and a user interface to engineer and work with the knowledge component, for instance in order to enrich the existing knowledge with new insights from experts. This knowledge base can be combined and connected with the organisational knowledge base of the company. Managers, being the users, work with the DSS via a user interface, and can use the system within the decision-making process.

The DSSs are the predecessors of the knowledge-enabled hybrid systems, which will be introduced in Chapter 3.3.2 (s. chapter 3.3.2 “The Hybrid Approach”) to provide explanations for users by using a (graph-based) knowledge base.

2.5.2 Business Analytics, Predictive and Prescriptive Analytics

The word analytics has somehow replaced the word Business Intelligence, and includes other terms, e.g., decision making, etc. In the literature, therefore, BI has largely been replaced by analytics. Sharda et al (2019) provide the definition of the Institute for Operations Research and Management Science (INFORMS), which defines (Business) Analytics being that which represents the combination of computer technology, management science techniques, and statistics to solve real problems. In figure 52, it is shown that Sharda et al (2019) consider Business Analytics an umbrella term for descriptive analytics or classic BI, namely looking backward and answering the questions “what happened?”, “what is happening?” by using business reporting, dashboards, scorecards, and data warehousing. Predictive analytics is to answer forward-looking questions, like “what will happen?” or “Why will it happen?”. To provide answers, prescriptive analytics is using data and text mining, web/media mining, and forecasting technologies. At last, prescriptive analytics is answering the questions as to “What should I do?” and “Why should I do this?”

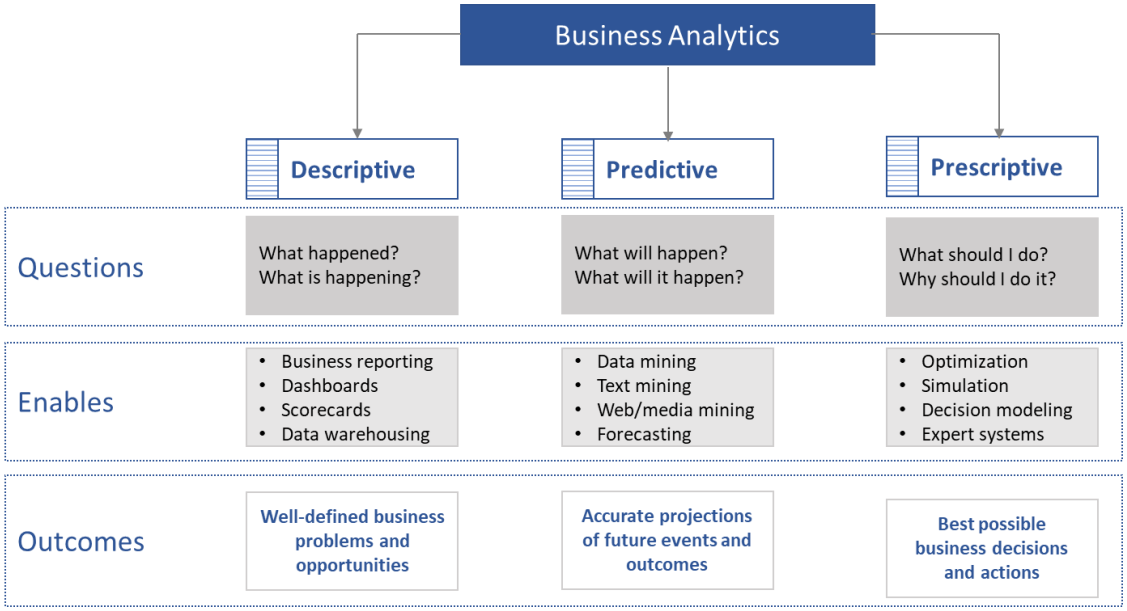

Figure 52: Business Analytics

The main objective of prescriptive analytics is to provide a recommendation for a specific action within a particular problem space. These recommendations result from solving optimisation problems and can either be presented to the user (decision maker) in a report or can be used directly in a system for automatic decision-making. Sharda et al. (2019) also refer to these kinds of analytical systems as normative analytics.

2.5.3 Data Science

The most current term for analytics is Data Science (D. J. Patil of LinkedIn is sometimes credited with creating the term "Data Science", O'Neil & Schutt (2014)). When someone is doing Business Intelligence, it pertains more to doing descriptive or reporting analytics, while in contrast, a data scientist is responsible for predictive analysis and statistical analysis and uses more of the advanced analytical tools and algorithms. Data scientists may have a deeper knowledge about programming (Python, R) or statistical knowledge – and are sometimes lacking knowledge from the business domain, which a more business-intelligence-focused user might have.

2.6 Summary

The findings regarding planning are that it is a critical aspect in the process industry, particularly in the chemical and life science sectors. Both industries play significant roles in the global economy and involve complex, interconnected processes, in which raw materials are transformed into intermediate and finished products through a series of chemical reactions and physical operations.

To allow these processes to proceed without disruptions and remain smooth, efficient, and effective, it is necessary to have a thorough understanding of the different steps, their timing and sequences, as well as the interdependencies existing within the whole process. Planning in the chemical industry involves the scheduling of production batches, the allocation of resources such as equipment and personnel, and the management of inventory levels. Effective planning will help to minimise downtime, reduce waste, and optimise the use of all resources involved. Planning in the life science sector is critical for developing and producing pharmaceuticals, biologics, and medical devices (sometimes seen as a separate branch, being healthcare). Planning in the life science industry includes coordinating

research and development activities, clinical trials, and regulatory approvals, as well as scheduling manufacturing processes and managing supply chains. Besides effective supply chain planning, Advanced Planning, and scheduling (APS) software is widely used in the process industry for optimisation, as this software can integrate data from various sources, including production schedules, inventory levels, and supply chain information, to generate optimised programs that minimise costs and maximise efficiency. Artificial Intelligence in the process industry will be used to improve forecasting accuracy, optimise resource allocation, and enable predictive maintenance, leading to improved efficiency and reduced costs. Effective planning is critical to the success of the process industry, particularly in the chemical and life science sectors, and ultimately leads to improved profitability and competitiveness.

By looking at the two planning processes of scenario analysis and tactical integrated business planning, the relevant stakeholders were identified and their decision-relevant variables. It was also pointed out that decision-making processes can be described, for example, by the model of Simon (2019). An AI that wants to support this (e.g., with regards to augmented decision support) must take these processes into account. The process flows will be taken into account later in the business model - the business architecture of Re_fish, as they describe the context, the situation in which Re_fish is used in the case considered here.

Stakeholders are further considered in two ways. Firstly, in the chapter on XAI, the various requirements that stakeholders place on an XAI are examined in more detail. The requirements will then be collected later and will be incorporated into the development of the reference architecture as requirements, functional or qualitative (possibly as constraints).

Finding 12: In Chapter 2.3, the corporate planning process of the process companies was presented. In particular, scenario planning, which is to be classified in the strategic planning area, and sales and operations (integrated business planning) planning, which is to be classified in the tactical area. These sub-planning processes have several possibilities to replace or at least support sub-processes with AI solutions. First and foremost forecasting, but also optimisation with regard to constraints - usually linear optimisation models are traditionally used here, but AI methods are already available. The identified stakeholders

and their requirements will be taken into account in the requirements for the reference architecture.

"At the same time, what was becoming clear to me was the extent to which humans, in their wish to escape loneliness, made maneuvers that were very complex and hard to fathom, and I saw it was possible that the consequences of Morgan's Falls had at no stage been within my control." (Ishiguro, Kazuo (2021). Klara and the Sun. Chapter 3)

3 Explainable Artificial Intelligence in Corporate Planning

3.1 Introduction

Artificial Intelligence[23] is a term which was first used as a theme in a funding request to the Rockefeller Foundation for a workshop ("Dartmouth Summer Research Project on Artificial Intelligence"), which then took place at the Dartmouth College in the summer of 1956.[24] This year has come to be seen as the foundation year of Artificial Intelligence. The term itself was invented by John McCarthy as being a topic for a conference; the participants included renowned scientists, such as Marvin Minsky, Nathaniel Rochester, Claude Shannon, Allan Newell, and Herbert Simon. In the following years, textbooks were presented, such as Feigenbaum and Feldman (1963), Nilsson (1971), Newell and Simon (1972), McCorduck (2004), Raphael (1976), Winston (1977), Rich (1983), Charniak and McDermott (1985), Haugeland (1985) and later, the famous textbook by Russel and Norvig (1995). Started in the 1950s, the field of artificial intelligence has developed into an exciting and interesting field of research.

[23] The term "Artificial Intelligence" was coined by McCarthy, as shown above, and is not accepted by all scientists. Instead, some would like to use the term "machine intelligence", e.g., Donald Michie, University of Edinburgh, who founded the "Machine Intelligence Institute" there.

[24] There was another conference 1956, at MIT, the "Symposium on Information Theory", which is now been seen as the foundation year of cognitive sciences.

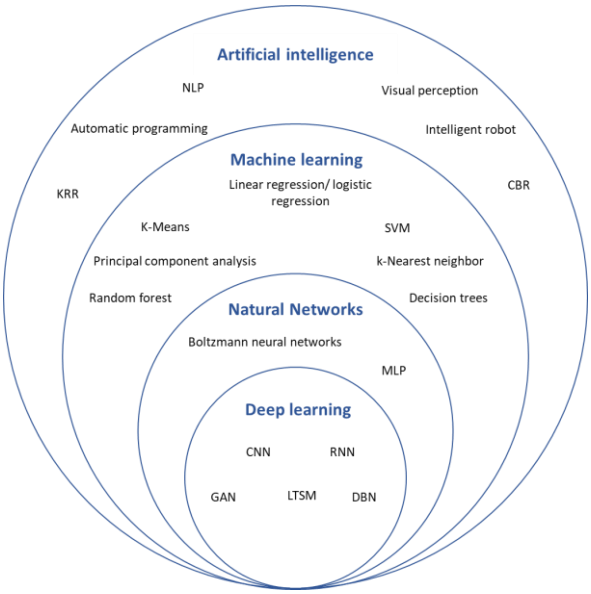

Figure 53: Artificial Intelligence and its “disciplines”

Associated with attempts to realise human-like mental processes and behaviours. Artificial Intelligence (AI) consists of a multitude of different research topics, which are shown in figure 53. Herein, the areas of Knowledge Representation and Reasoning (KRR) are presented within the chapter about knowledge-based systems (Chapter 3.2.2 “Knowledge Based Systems”) and the topic of machine learning as a sub-area of AI, with its associated areas of neural networks and deep learning. This will be done in Chapter 3.2.1 “Machine Learning and Deep Neural Networks” As mentioned in Chapter 1 - when AI systems make decisions and these decisions are not explainable, then the user does not trust those decisions. As a result, the systems are not used (or are not implemented to the extent that they could be). The field of Explainable AI is dedicated to the explainability of AI systems. However, recently it has included a focus on the so-called non-symbolic AI systems. In the past, around the 1970s, the symbolic systems of AI were examined in regards of explainability. In addition, research into explanations of when, why, and how people explain facts and decisions to others has a long history in the social sciences and psychology. Recently, hybrid approaches to explainable systems based on Knowledge Graph databases have emerged as one of the most promising approaches for a generic and therefore comprehensive approach to explanations of AI-based systems or decisions. These topics are presented in Chapter 3.3 “Explainable Artificial Intelligence”, with the presentation of explainable and interpretable AI, as well as the “Knowledge Based Systems of Explainable AI” approach in Chapter 3.3.2. The Chapter 3.2.3 gives a short introduction to Neuro symbolic

AI. In Chapter 3.4 “Ethical AI, Law and Regulatory Requirements of Explainable AI”, the requirements of users with regard to ethics, laws, and regulations are presented as well as how these are accounted for and implemented in the context of the design, development and operation of AI systems. Chapter 3.5 maps the stakeholder of corporate planning (Chapter 2.3.4) with this chapter. Chapter 3 concludes with a summary in Chapter 3.6 “Summary”, in which the findings are collected and presented.

3.2 The Technical Perspective of Artificial Intelligence

There exist many different definitions of Artificial Intelligence among experts – In their renowned book Russell and Norvig (2022) differentiate definitions by describing the objectives of AI in two dimensions, with one dimension focusing more on fidelity towards human performance, and the other towards rationality. The second dimension lies between intelligence as a thought process, and reasoning in the meaning of "thinking" intelligence, and demonstrating intelligent behaviour (which is a behaviouristic, external characterisation, in the meaning of observing intelligent behaviour) -- in the meaning of "acting" intelligent. From these two dimensions, we may derive at least four possible combinations, as follows. *Acting humanly* – this can be best represented by the Turing test approach, which an intelligent agent has to pass. To do so, an agent needs to use natural language processing, knowledge representation, automated reasoning, machine learning, and for the total Turing test, computer vision and robotics, as well. All these disciplines cover almost all AI disciplines. *Think humanly* -- this combination can be seen as the foundation for cognitive sciences, with the idea that one can only learn about human thoughts by way of introspection, psychological experiments, or brain imaging. However, if there is a sufficient and precise theory of mind, then this can be implemented in algorithms within computer programs. Then one would expect that the algorithms can also be working within humans when the results (that is, the output) matches human behaviour. *Thinking rationally* is reflected in the “logicist” tradition within artificial intelligence, the idea of which is to create systems able to build on programs which can solve any logical problem described in logical notation. The problem is that logic needs to have a certain knowledge of the world, which is not in fact possible.[25] Then, the theory of probability is being used to be used for reasoning with uncertain information. This approach is not providing intelligent

[25] S. e.g., one try to explain the whole world logically, s. Wittgenstein (2014). S. bounded rationality.

behaviour. The last combination is the rational agent approach of *acting rationally*. This is about an agent that does the right things, depending on the current information, beside the problem of limited or bounded rationality, when the computational demands are so high and there is not enough time to do all the necessary or desired computations. Acting rationally is also about intelligent decisions. This needed to be done by the agent with regards to its goal; the agent also has to have situational awareness, as it needs to evaluate the context it is in. This definition will be used in this work.

Taking a different look at the objectives of AI and AI research, it may be easier to understand what AI is about through the goals of AI research. Görz et al. (2021) describe the objective of AI research as the construction of "intelligent" systems that make certain human perceptual and intellectual capabilities available to machines (Görz et al., 2021). As a second objective, there is “[c]ognitive modelling, i.e., the simulation of cognitive processes using information processing models.” (Görz et al., 2021).

3.2.1 Machine Learning and Deep Neural Networks

If an agent is observing the world around it and can improve its performance based on those observations, it is learning. If the agent is a computer, this process is therefore called machine learning. Thus, machine learning can be defined as follows: a computer observes data, build a model based on the data, and uses the model as a hypothesis of the world and the software that is able to solve the problem (Russell & Norvig, 2022).

Based on the representation in figure 53, it can be seen that machine learning is a sub-area of artificial intelligence and deep learning is a sub-area of machine learning (Russell & Norvig, 2022).

There are three kinds of machine learning methodologies to be distinguished: supervised, unsupervised, and reinforcement learning. Machine learning needs historical data to be able to learn. If it is possible to map from an input to the output by using labelled historical data, we call it supervised learning. This learning is used for regressi on and for classification, for instance.

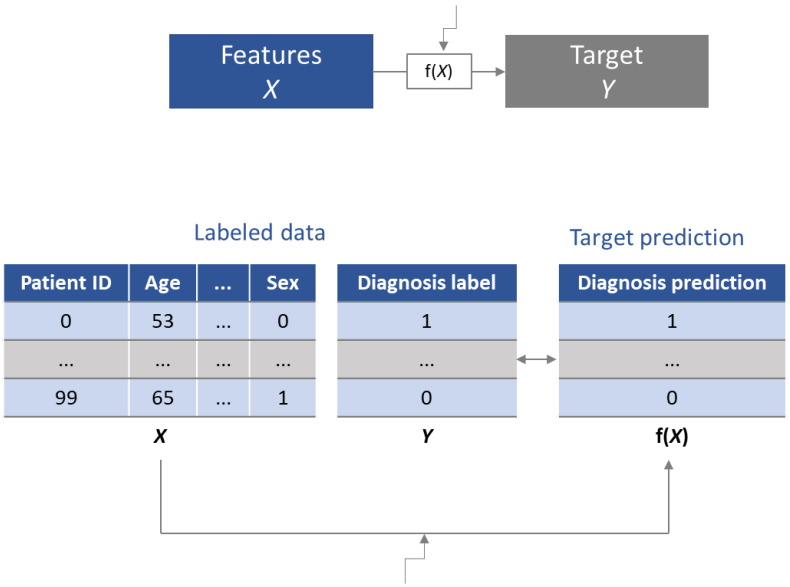

Figure 54: Supervised Learning- based on Thampi (2022)

As can be seen in the figure (s. figure 54) above, the machine learning model is a function that maps the features X on to the target Y. To be able to do this, the machine learning model needs to find labelled historical data; in the figure above, this is the "Diagnosis label" column (values 0 and 1).

In unsupervised learning, no historically - labelled data is available, so the only possible goal for machine learning is to learn a representation (patterns in the data) of the data that best describes it. Unsupervised learning is mainly used for classification.

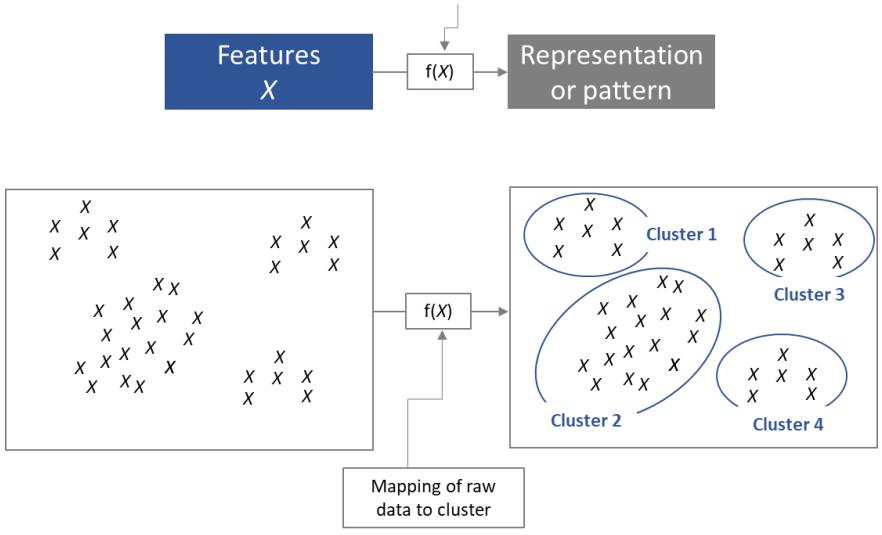

Figure 55: Unsupervised Learning - based on Thampi (2022)

In figure 55 it is shown that unlabelled data is being mapped by the machine learning model into clusters. Therefore, the machine learning model is learning a representation or a pattern of the historical data.

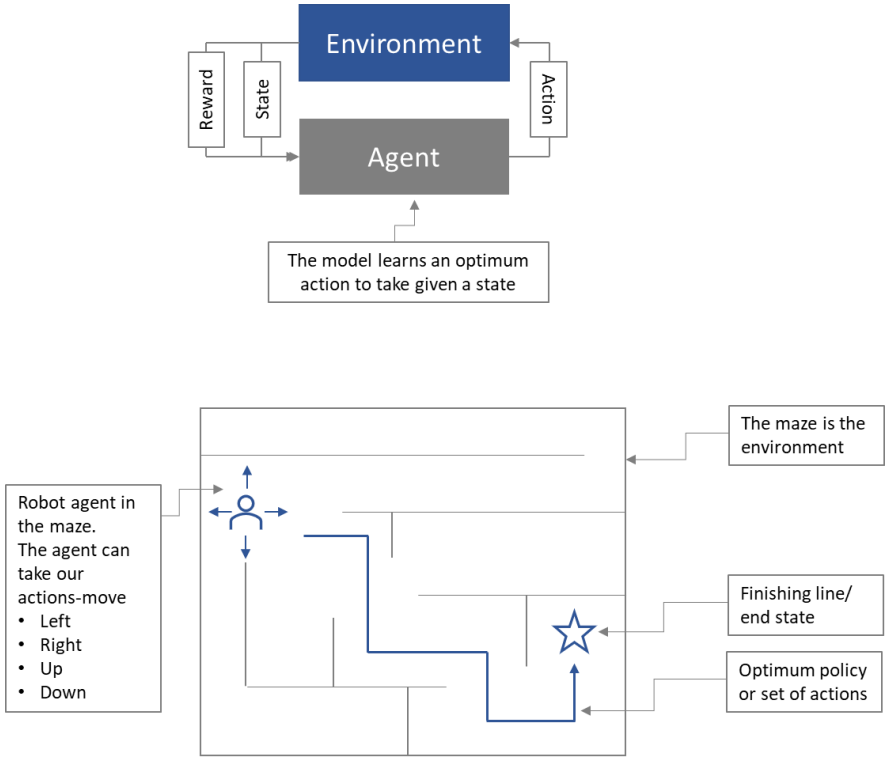

Figure 56: Reinforcement learning model, based on Thampi (2022)

In reinforcement learning (s. figure 56), an agent learns by interacting with an environment (the agent's "world"), and based on its actions, the agent receives a reward or a penalty. The goal of the agent is to maximise the cumulative reward. Therefore, the agents need recurrent feedback in the meaning of an "understanding" of its surroundings, namely the world it is living in (Situation calculus, (McCarthy & Hayes, 1969)).

Neural Networks/Deep Neural Networks

The term Deep Learning refers to a whole series of machine learning technologies. It involves the use of interconnected artificial neurons (or Russell and Norvig (2022) circuits), which are also organised into multiple layers. The number of layers here reflects the depth, and thus the computational paths of the connections from input to output. Deep Learning technologies are successfully used in the fields of visual object recognition, machine translation, speech recognition, speech synthesis, and image synthesis, as well as in the context of reinforcement learning applications.

The term neural network has its origins in the work of McCulloch and Pitts (1943), who attempted to simulate networks of neurons in the brain with computational circuits. As with the other machine learning models, neural networks consist of models that map an input of X to an output of Y:

$$f(x, \theta) \rightarrow y, \qquad \text{(f15)}$$

The inputs and outputs can be scalars, vectors, matrices, and higher-dimensional number packages (which are called “tensors”).

For example, in a digit classification problem, x is a pixel matrix containing the grey values of the image, and y is a probability vector containing probabilities of the possible digits. θ is a vector of free parameters that is modified by an optimisation procedure so that the loss function

$$L(\mathbb{S}train, \theta) \qquad \text{(f16)}$$

for the training data $\mathbb{S}train$ becomes as small as possible. This has the effect, for example, in digit classification, that the probability of the observed outputs (digit classes) is as high as possible. In this way, KNNs "learn" to perform a task without being given rules or instructions on how to perform the task.

Even though the term Deep Learning was coined more recently, many of the basic concepts and algorithms can be traced back to much older work. These include the basic principle that complex functionalities can be generated through the interaction of many uniform elements. In this context, several overview articles describe the development of deep neural networks within three phases. Although these phases were characterised by different objectives, the resulting architectures follow the same basic principles, which have retained their importance to this day.

It is also worth mentioning that DNNs are based on much older concepts, so a distinction is essentially made between three different methods.

Cybernetic approaches were in the foreground in the first development stage of deep neural networks. The goal here was to develop a better understanding of how learning can function in biological systems, via feedback mechanisms, among other things.

Connectionist approaches, which essentially characterised the second phase, aimed to emulate complex cognitive perceptual performances. The starting point was the consideration

of the network-like connections of neurons in the brain, whereby the reproduction of biologically plausible neuronal processing elements tended to take a back seat. In this phase, neuronal networks emerged which were able to achieve their first major successes in the field of image classification, among other things.

While the neocognitron was still adapted to the data by a very simple form of self-organising learning, the next major advance, initiated by Yann LeCun et al. (1989) in 1989, was to train convolutional networks with the help of gradient methods.

Recurrent networks, which form the basis for processing sequences, underwent a similar development. One of the main representatives is the Long Short-Term Memory (LSTM) introduced by Hochreiter and Schmidthuber (1997) is currently a central component of many neuronal architectures.

3.2.2 Knowledge Based Systems

As already mentioned in 3.1, knowledge-based systems are one of the most important fields of Artificial Intelligence. But as seen in figure 53, it also consists of many other fields of AI, e.g., machine learning (s. Chapter 3.2.1) or logic, rules, inferencing, etc.

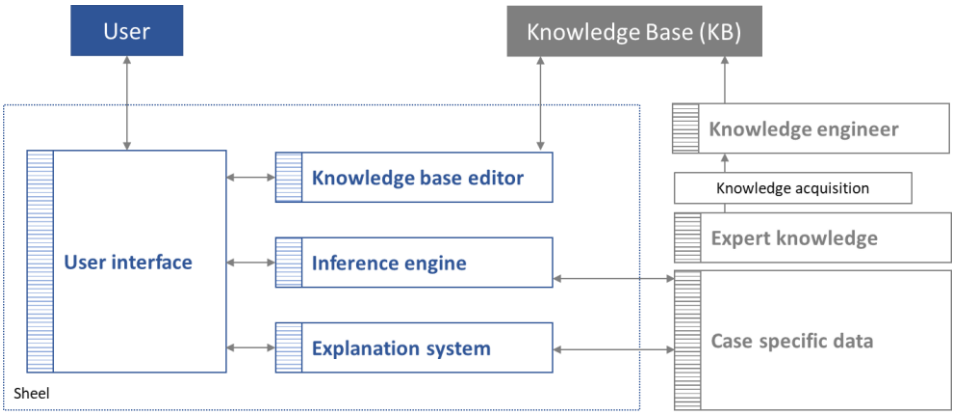

Figure 57: Typical knowledge - based system architecture Samawi et al. (2013)

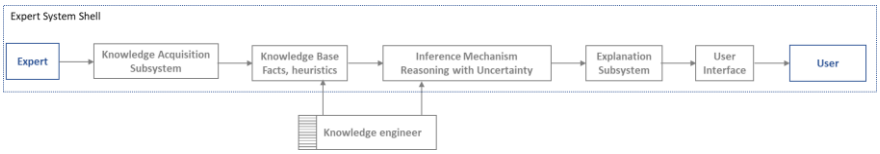

Figure 58: Typical expert system architecture Samawi et al. (2013)

One of the most important knowledge-based systems is the medical expert system MYCIN, which was developed at Stanford University.

The basis of knowledge-based (KB) systems is the logical knowledge representation and inference. Rule based systems are one of the oldest KB systems – they are built upon if-then rules and are easy to understand and handle with classical logic. They are used in well-structured areas, where only 0-1 decisions need to be taken. Machine learning is also one of the most important characteristics of a KB system. Learning can be seen as intelligent behaviour and humans can learn from experience, observation of the world, samples, trial and error, reading, and the like. So, in the following it is of great importance that the system can learn and therefore facilitate the growth of knowledge within the knowledge base.

The expert systems (s. figure 57 and 58) in the past were not very successful, as they had problems with learning. The primary task is to solve the problem of knowledge representation and processing with a new approach. The central idea is to build up a case database in which problems are stored as pairs along with solutions. When being confronted with a problem situation that has occurred the same way before, the systems should be able to provide a decision or at least a recommendation for this near-equivalent problem. The challenge is to find the situation among the present cases and apply the solution stored there. But if the problem situation is new, one tries to adapt the solution of a possibly similar case accordingly. Non-monotonous inference in classic logic it that the set of possible equivalent cases is growing monotone – but it is often not the case.

During such inference processes, it can occur that an inference must be taken back, because of additional knowledge. This is called non-monotonic reasoning (approaches to solving this kind of inference are truth maintenance systems, or default logics).

Planning, like non-monotonic reasoning, is a manifestation with intelligent behaviour. But the goal, as already described in chapter 2, is not to determine if a specific situation exists. Instead, it is to plan a sequence of actions -- the execution of which changes the present state into one where the target description applies. The so-called “situational calculus” provides the logical basis of planning. The knowledge component of an agent is the part of the system where methods and processes for the representation and intelligent processing of information are implemented. Therefore, this component plays a central role for the system. The Expert System MYCIN is the ancestor of all knowledge-based systems, in

which the presentation and processing of uncertain knowledge played a central role. For example, many concluding rules in the field of medicine apply only to a certain extent: there are certain characteristic pain symptoms for appendicitis, for instance. One of the most important characteristics of a knowledge-based system is the separation between the presentation of knowledge about the problem area (the knowledge base) and the processing of this knowledge (knowledge processing). Specific knowledge about the field of application should be found in the knowledge base. The knowledge processing, however, is an application-independent problem-solving component of the system. Due to this, a clear separation between problem description and problem solving can be provided. In contrast to classical programming approaches, for instance, the following aspects, among others, can be realised. Knowledge about the scope of an application can be expressed directly. Expert systems are therefore a kind of special knowledge-based systems in which the knowledge ultimately comes from experts (who can be seen as knowledge engineers, in a sense) (Beierle & Kern-Isberner, 2019). In 5.2.2, there is a detailed description of such an expert system architecture.

One important method for building knowledge is to represent it in graph form, in so-called knowledge graphs (KG). A knowledge graph can be seen as a machine-readable way of representing information about the world, including entities, relationships, attributes, facts, beliefs, and even provenance, including justifications and uncertainty Kejriwal, et al. (2021) provide a classification of different semantic networks by John F. Sowa (2006 and 2010) (can be found under http://www.jfsowa.com/pubs/semnet.htm, s. Sowa, 2006 and Sowa , 2010). The formal framework to describe knowledge graphs is the Resource Description Framework (RDF), which was developed by the WWW (World Wide Web Consortium). Based on this Web Ontology Language (OWL) is an even more powerful framework, using a reduced set of predicative logic developed as an extension of RDF and to provide a solution for its limitations (Dengel et al., 2012 and http://www.w3.org/).

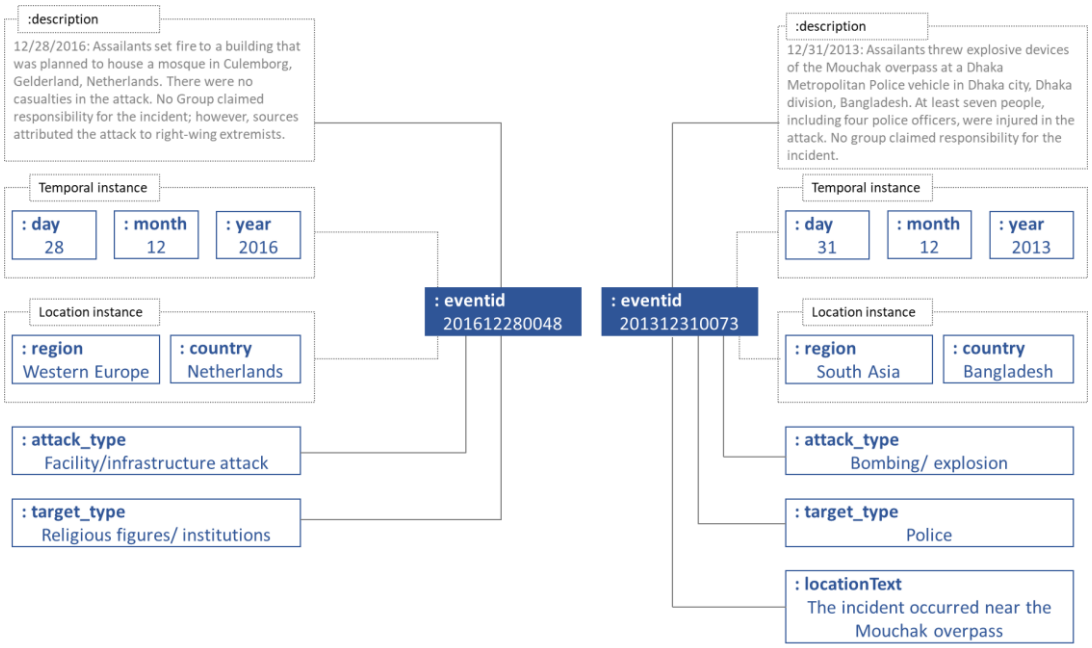

Figure 59: Fragment of an event KG- based on Kejriwal, et al. (2021)

In figure 59, there is a fragment of an event KG to express geopolitical phenomena with, e.g., eventid, attack_type, target_type, description, etc.

As already mentioned in chapter 1 and what is following, when AI models (or agents) carry out decision making or recommend a decision, the user needs explanations. One of the functions of such knowledge graphs as the knowledge representation of an AI system is their explanation capability. With an RDFS encoded ontology, enhanced by SWRL rules, an AI system can provide explanations out of the knowledge base. It is also possible to make inferences on the provided knowledge by using the RDFS framework (Dengel et al., 2012; e.g., using the rdfs:subClassOf relation).

Explanations have a long history, and not only in AI research; as mentioned above, but it is also a research topic in the social sciences, philosophy, etc. Explanations will be a part of chapter 3.3. Ilaria Tiddi (2016) showed in her thesis how the Web of (Linked) Data can be used for pattern explanation. McGuiness et al. (2004) showed an Inference Web approach to generate distributed explanation. The usage of knowledge graphs in explanations is done in Chapter 3.3.2.

3.2.3 Neuro-symbolic AI

At the AAAI fireside chat in 2021, the scientists Kahnemann, LeCun, Hinton and Bengio discussed the future of AI and agreed that the joint development of humans and machines is the way forward. They emphasized the importance of ensuring that AI is designed to work alongside humans rather than replacing them. The scientists also stressed the need for transparency and accountability in AI development, as well as the ethical considerations that must be considered. Overall, they were optimistic about the potential of AI to enhance human capabilities and improve our lives but emphasized the importance of responsible development and deployment, so the goal should be to develop an AI system that produces "semantically sound, explainable and ultimately trustworthy AI[...]", requiring "a sound reasoning layer in combination with deep learning" (AAAI, 2021; Garcez & Lamb, 2023, p.2).

The researchers argue that a neuro-symbolic system that combines the learning capabilities of neural networks with the reasoning capabilities and explainability through symbolic representation for neural models best represents the cognitive model developed by Kahnemann with both systems 1 and 2 (Kahnemann, 2013). The combination, the “hybrid approach”, in the combination of symbolic and subymbolic or non-symbolic connectionist AI is also seen by other scientists (Marcus etc.) as the key to overcoming the above barriers in the use and adoption of the currently so successful subymbolic AI methods, e.g., Deep Learning, as the new, the third wave of AI. The first wave of AI began in the 1980s and was characterised by symbolic logic programming. The second wave began afterwards and is characterised by connectionist neural models. The consensus view is that the third wave will be characterised by the combination of the two methods of symbolic and subymbolic AI. According to Ilkou and Koutraki (2022), there are three differences between the two AI fields. In their research they conclude that symbolic approaches produce logical conclusions, while subymbolic connectionist approaches produce more associative results. It is also worth noting that human intervention is often required when applying symbolic AI methods. In contrast, subymbolic approaches can learn and adapt to the given data independently. These findings are intriguing and could have significant implications for the future development and application of AI technologies. Ultimately, symbolic approaches work best with few but precise data, whereas subymbolic approaches require large datasets. By combining symbolic and subymbolic methods, there are advantages that can be described as follows: they have greater accuracy, efficiency and readability of knowledge,

and they have a higher explanatory capacity without the need for a priori assumptions; they are comprehensive and integrate statistical learning and logical reasoning, and they can work with noisy data, finally, they can combine logical rules with data during learning and fine-tune knowledge based on input data. In principle, therefore, they appear to be suitable for applications that use a large amount of data and require knowledge descriptions (Garcez et al., 2015; Bader & Hitzler, 2005; Garcez, 2019).

Garcez & Lamb (2023) claim that a purely symbolic or neurosymbolic ML system should be able to achieve the three levels of causal reasoning. This is made possible by mapping the neural networks to the symbolic descriptions.

3.3 Explainable Artificial Intelligence

Since there are ever-more AI capabilities involved in systems and, therefore, in business processes, like the management process (where they are especially integrated in planning; s. chapter 1), there is a growing demand for explainability. The decision-makers want to understand the why, what, and how of a decision made, or an action conducted. A difference has also been defined concerning transparency in the meaning of *ad hoc* understandability of an AI model decision process and the *ex-post* explainability that a model explains (an explainer) to the decision-maker or the stakeholder (explainee). We also mentioned that an explanation is context-sensitive, in that it belongs to a specific situation and/or a specific stakeholder impacted by the decision.

Explanations are being used as a communication method to make actions or facts understandable and to build knowledge, so that the communication partners can then make better-informed decisions (Schank, 1986).

In everyday life, explanations can be categorized differently, e.g., Stegmüller (1977) differentiates between causal explanations -- explaining the causality, semantically explanations – explaining the meaning, e.g., of a specific concept, corrective explanations, justifications, descriptions of functionality, and mediations of practical knowledge. Among scientific explanations, the best-known is the Hempel Oppenheim model, which differentiates between deductive nomological, and inductive statistical explanations. These are focusing on answering the “why” questions, which can be answered logically, based in the initial status and by using rules. The fact to be explained is called the “explanandum” and the

facts and rules are the “explanans”. The differentiation between the two types of explanation mentioned above results from the usage of the rules – deterministic or statistical (Hempel & Oppenheim, 1948).

The users build a mental model of how the AI system operates and how it was constructed- and how the data was used to develop and train how it matches the situation. It can also include descriptions of the underlying rationales and reasoning paths the system used to arrive at a conclusion, which in turn can be based on observed statistical regularities, models of underlying mechanisms and causal relationships, and temporal patterns (Chari et al., 2020).

The importance of explainability is more necessary, with collaborative AI systems meant to work in tandem with human users (a human in the loop) in order to augment rather than supplant their skills and capabilities. This can be seen as a “distributed cognition” approach, in which cognition is seen to take place not within the head of any one individual, but rather through the exchange and transformation of representations across multiple actors and artifacts (Elster, 2015). The ability for a system to provide explanations and respond to queries that reference other information relevant to the situation, expands the range of ways in which the system and human actors can interact.” (Chari et al., 2020)

As already mentioned, explainability or explanation has been the research focus for many years. There is a vast amount of research which has already been done in disciplines other than computer science. The explanation has been deeply researched in social sciences, philosophy, psychology, cognitive sciences, etc.

MYCIN is a rule-based systems expert system including a wide range of reasoning components with potentially inductive or abductive reasoning and more traditional deductive reasoning. MYCIN is one of the first expert systems developed at Stanford University and already used an explanation component (Buchanan & Shortliffe, 1984). The idea for the explanation component was that the system keeps explanation templates, which are then enriched by the usage of data from the trace (protocol of the rules of the system, which had been used). The method is simple; however, it assumes that the wordings in the templates are known to the user and the explanation is valid only locally, as it only covers a small part of the problem. The overall strategy of the system cannot be provided, as it is based on complex interactions between the rules. As rules can imply a set of inferences, it might be not understandable. Also, the rules consist only of limited knowledge and therefore not

all relevant explanations can be given, especially justifications as to why a specific action has been done.

NEOMYCIN was built as being a successor of MYCIN, especially to improve the explainability aspect. The main improvement was that NEOMYCIN received implemented meta-rules, which were used to activate rule sets. These sets included all rules necessary to conduct specific tasks. Therefore, the rule could be linked to a context of the task in the system. The overall strategy of the system could be explained by using these meta rules; however, the other issues already mentioned in the context of MYCIN, like justification, could not be addressed, either (Dengel, et al., 2012).

Explanations are highly contextual and depend on users, their roles, prior knowledge, and the situation of the decision being made. Therefore, all relevant context dimensions, e.g., of explainable AI, must be taken into consideration. The more prevalent AI becomes, and the more people are affected by AI, the higher the demand for appropriate explanations. Thus, explainable AI must address the requirements of different groups affected by AI and be respectful of different requirements and presumably different contexts of the stakeholders. Among those, one may include:

Domain experts/users of the model (e.g., employees, physicians, planners) who must trust the model and gain some scientific knowledge -- in the scope of the thesis, management or planning experts are the domain experts.

Users affected by the model decisions (e.g., loan applicant, patient, driver, other planners and plans) who must understand their situation and verify fair decisions.

Regulatory entities/agencies (e.g., auditors) who must certify model compliance with the legislation; in the scope of the thesis, management or the executive board, or the auditors- are the stakeholders here, as well as the public and the government.

Managers and executive board members who must assess regulatory compliance and understand corporate AI applications.

Data Scientists, developers, and product owners who must ensure and improve product efficiency research and new functionalities.

Governments (e.g., on a national or supranational level, like the EU) who must ensure that the model follows specific regulations, e.g., the ethics guidelines for trustworthy AI (Barredo Arrieta et al., 2020; Bejger & Elster, 2020).

In different stages of the lifecycle of an AI system, different stakeholders come into in focus. For the objective to reach the requirements of the guidelines for ethical AI and AI systems, the requirements already have to be considered during design and implementation (Ryan & Stahl, 2021). Methodologies to design and implement an ethical AI could be e.g., done by enriching the CRISP-DM method into CRISP-DM & AI (Bejger & Elster, 2020; Chapman et al., 2000).

Wolf describes that the critical challenge in the deployment of explainable AI systems within complex settings is the understanding of unique requirements by the users/stakeholders. Therefore, he propagates a scenario-based design helping to envision these specific unique requirements. (Wolf, 2019)

3.3.1 Explainable or Interpretable Machine Learning

In the following section, methods are introduced which should serve to make the non-symbolic AI models (and in particular, machine learning models) explainable. However, this first requires some justification. It is more important to understand that machine learning and ML models (the non-symbolic models) are only one part of artificial intelligence. ML can be differentiated from other AI models and methods as non-symbolic AI (s. Chapter 3.1, figure 52)

Much of the literature is concerned with ML explainability rather than AI explainability. ML explainability methods can be distinguished between techniques that are intrinsically or *post-hoc* explanatory. Methods are model-specific or model-agnostic and techniques provide local or global explanations in terms of a specific example or the global model itself.

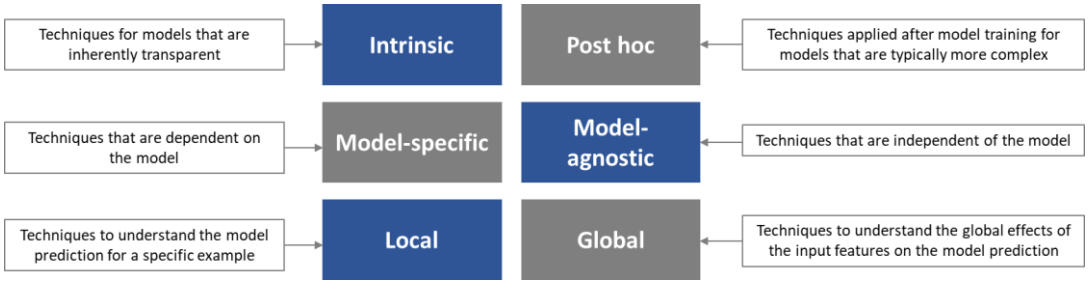

Figure 60: Different categories of techniques to make ML models interpretable- Thami (2020)

It can be seen from the figure 60 that there are different categories of techniques and methods to make ML models explainable. There are techniques for models that are inherently transparent, with so-called intrinsic interpretability, or those with post-hoc interpretability, when the techniques for explainability can only be applied after the model has been trained, and so on.

The following methods, s. table 11(among others), will be briefly explained:

Interpretable (explainable) Machine Learning Methods						
Name	Year	Author/ Creator	Inrinsic/ Post-Hoc	Specific/ Agnostic	Local/ Global	Web
SHAP	2017	Scott M. Lundberg	Post Hoc	Agnostic	Local	https://shap.readthedocs.io/en/latest/index.html
LIME	2016	Marco Tulio Ribeiro	Post Hoc	Agnostic	Local	https://github.com/marcotcr/lime
ELI5	2016	Mikhail Korobov, Konstantin Lopuhin	Post Hoc	Agnostic	Global	https://eli5.readthedocs.io/en/latest/index.html
Skater	2018	Open Source	Post Hoc	Agnostic	Both	https://github.com/oracle/skater
Skope_rules	2017	Gardin et al.	Post Hoc	Agnostic	Both	https://github.com/scikit-learn-contrib/skope-rules

Table 11: Introduced models of interpretable ML/ XAI for ML

SHAP

SHAP stands for SHapley Additive exPlanations and is a Python library-based approach to explaining the outcome of any machine learning model. It was introduced in 2017 by Scott M. Lundberg and Su-In Lee. The approach combines the idea behind LIME (see below) and game theory. Shapley value is a concept from the field of cooperative game theory and quantifies the effect of a coalition of players on the game outcome. In the SHAP approach, the features are the players, and the model itself is the game. The SHAP approach thus makes it possible to calculate the effect of features and "coalitions" of features on the prediction of the ML model, thus rendering it explainable (Molnar, 2019; Lundberg & Lee, 2017; Mangalathu et al., 2020, Mazzanti, 2020, Bhatt et al., 2019)

LIME

LIME stands for Local Interpretable Model-Agnostic Explanations and was introduced in 2016 by Marco Tulio Ribeireo and his team (Ribeiro et al, 2016). The idea is that deep neural networks can use, for example, very complex functions to learn a classification. These functions are difficult to explain "globally". In contrast, DNN decisions can be explained locally; this is achieved by selecting a specific, exemplary result. Then, from a selection of outcomes, a new perturbed dataset is created in which numerical and categorical features are first used means, standard deviations, or frequency distributions. The model is again applied to this new data, and then the distances of the resulting outcome instances are weighted, using an exponential kernel function. This gives a higher value to

the results which are close to the original result instance than to those that are further away. Subsequently, these data are processed with a white box model, e.g., a linear regression. Linear regression enables the representation of feature importance in the results. This makes it possible to explain locally which of the features was relevant to the respective result (Nguyen, 2020; Ribeiro et al., 2016)

ELI5

ELI5[26] ("Explain Like I'm 5") is a library which is based on the programming language PYTHON. The idea behind ELI5 is that it should be used for AI pipelines, in order to visualise and debug various machine learning models by using a unified API. The library provides built-in support for several ML frameworks and a way to explain black box models. The result of using ELI5 could be a table, for instance, where feature importance for a specific model can be seen.

Skater

Skater[27] is an open-source unified framework to enable model interpretation for all forms of models, to help build an interpretable machine learning system -- which is often needed in real-world-use cases. Skater supports models which help to demystify the learned structures of a black box model both globally (inference-based on a complete data set) or locally (inference-based on individual prediction).

Counterfactual Explanations

A post hoc method that can be applied to both text and images. The idea is to change the decision based on the smallest possible change in the input - for example, changing a few pixels in an image or, in the context of a situation, changing a single parameter. Good counterfactual explanations can be defined according to the following criteria:

- The original value and the (artificial) new value are quite similar.

- As few input parameters (features) as possible should be changed.

- A larger number of different explanations seems reasonable

- the (artificial) change of features should be realistic

[26] https://www.benlabs.com/resources/explain-like-im-five-artificial-intelligence/ , accessed 18.06.2023
[27] https://github.com/oracle/skater , accessed 18.06.2023

(s. Pearl, 2019; Wachter et al., 2017; Stepin et al., 2021; Molnar, 2019)

So-called attribution models are used in particular in the use of neural networks, for example in the classification of image content. For example, so-called saliency maps are used in classification (Bejger & Elster, 2019) when it comes to classification - "Border Collie or "cat" - the different components of the neural network are examined to determine which pixel area led to the decision to classify the content of an image as "Border Collie".

Other known applications in the field of attributive methods are, for example:

CAM/ Grad-CAM or Grad-CAM++ (Gradient-weighted Class Activation Mapping).

This method is used post-hoc and can be applied to neural networks such as CNN. A saliency map is generated that is superimposed on the original image and thus helps to identify the decisive area for the decision. (see e.g., Kraus et al., 2022; Zhou et al., 2015; Selvaraju et al, 2019; Chattopadhyay et al., 2017).

LRP- Layer-Wise Relevance Propagation

Is an attributive post hoc approach that considers the influence of individual inputs on the classification result. When classifying content, the aim is to determine which pixels or image parts have influenced the result positively or negatively and to what extent (by assigning a relevance value to each pixel). An output is the sum of the relevance values of the input variables. The calculation of relevance is done iteratively from back to front. (Bejger & Elster, 2019; Bach et al., 2019; Samek et al., 2019; Shiebler, 2017)

The abundance of approaches to making non-symbolic AI models explainable also shows the inherent difficulty: some of the approaches are even more complex to understand and implement than the application of the models, themselves. The issue may be described as follows: if (as claimed above) the application of AI models also depends to a great extent on their being explainable, and on the user feeling trust, then the models of explainability should also be trusted. For why should a complex model that explains another complex model be trusted, per se? If a layperson cannot understand the application of the concepts and models for explanation, why should that person trust them? Moreover, MLs present other major difficulties, which will be further explained below. These are, for example:

Data Leakage:

This issue occurs in the data when features used for training leak information that does not appear in the production environment.

Bias:

Bias means that the data includes a pattern, which leads to an unfair prediction for one specific group over another. Classic biased data examples include the COMPAS AI system used by the US Court to predict future criminals, while the data used to train the model included a severe racial bias (Angwin et al., 2016).

Regulatory noncompliance:

This issue occurs when the data being used is not GDPR-compliant and uses data which violates article 17 of the EU GDPR regulations (s. Bejger & Elster, 2020; and EU regulations).

Concept Drift:

This issue occurs when the properties or distribution of the data changes over time and is not reflected in the training data which was used to train, validate, and tune the model.

In addition to these approaches, there is a large body of work dealing with taxonomies, frameworks, etc. for the field of non-symbolic models. For example, the approach by Lipton (2016) with its “Desiderates” provides a framework and requirements for ML models. Lipton built a framework focusing on machine learning models (supervised learning) and interpretability. He emphasises the goals of XAI, which can be achieved by interpretability, and mentions his "desiderates.", which are:

Trust: Interpretability is a prerequisite for a trustable ML system. Lipton describes trust as knowing “how often a model is right” and “for which examples it is right”, giving an example from the models used by the government to predict crime rates in a neighbourhood. We do not expect any biases that might lead to over-policing some neighbourhoods.

Causality: It is desirable that models pick up more than associations, even if they are optimised for doing so, like in the case of supervised learning. The idea is that causal relationships can also be inferred, in order to generate hypotheses about the natural world. For such interpretability to infer causal relationships from observational data, it would be necessary to solve the causal discovery problem from observational data (Pearl, 2009).

Transferability: This need points towards the resistance of ML models to so-called noisy data and domain shifts. Lipton (2016) describes an example of Caruana et al. (2015), where the deployed model and its usage alter the environment and invalidate the future prediction (domain shift) (cf. Caruana et al., 2015). Lipton also points to examples where stakeholders can "game the system" by intentionally using adversarial manipulation, e.g., in credit ratings (Lakkaraju & Bastani, 2020; Lipton, 2016).

Informativeness: While the learning objective of the ML model is to reduce error in the real world, the idea is to provide helpful information to decision-makers. The idea here is to provide additional (valuable) information to the stakeholder/user.

Fair and Ethical Decision Making: As already stated above, this desire includes the concern of many present politicians, journalists, researchers, and intelligence system designers that the interpretations will provide the possibility to assess or audit the decisions made by an AI system that conforms to ethical standards.

Lipton (ibid.) then provides an overview of techniques and model properties that will lead to the achievement of the desiderata, as mentioned earlier. He differentiates between transparency and *post-hoc* explanations. These methods and techniques are:

Transparency: Due to the limited capacity of cognition, Lipton points out that in his opinion, even linear models or rule-based systems are intrinsically interpretable. This is the opposite of the opacity mentioned above (or the "black box"), which means understanding how a model's mechanism works. Lipton then differentiates this transparency into three-level suitability for the whole model, meaning that a human should be able to take input data and parameters, etc., and in a reasonable amount of time, move through every calculation in order to produce a prediction. The second level is decomposability – the meaning of interpretability in the individual level of components, like the parameters, and at least on the algorithm level (algorithmic transparency). Here, in the case of linear models, the researcher understands the shape of the error surface and can prove that the model will converge into a single unique solution for non-training data. However, this is not given by using deep learning methods, as they lack transparency.

As described above, the non-symbolic models of AI (the machine learning models) are currently extremely successful, though their use is also associated with difficulties, such as those which were briefly presented. By and large, the techniques and methods used to explain them are as complex as the models they explain. This raises doubts as to why a

decision-maker should trust one more than the other. It also seems questionable as to whether the models of explainability can be applied universally in all conceivable contexts. Consider the use of LIME or SHAP in an autonomous vehicle to explain why the controller initiated an evasive manoeuvre (or not, e.g., the Tesla accident with the "white truck" (New York Times, 2021). This raises the suspicion that not only does the use of ML models depend strongly on the context, but the technologies and methods used to explain them do, as well. Miller et al. (2017) aptly formulates in an essay that the explainability provided by mathematical experts and statisticians ensures that their ML models are transparent, and their decisions are also transparent, as if the inmates were running the asylum.
Therefore, approaches that are broader in the sense of AI and use symbolic approaches in addition to non-symbolic approaches appear more promising. These hybrid approaches are briefly presented in the following chapter.

3.3.2 Knowledge Enabled Systems of Explainable AI

The issue of explainability has been an important topic since the beginnings of AI and has interested many researchers. Hybrid systems belong to the category of knowledge enabled systems and have components of both symbolic and non-symbolic AI, meaning that they can contain a wide range of inference components, including potentially inductive or abductive inference, as well as traditional deductive inference.

At the beginning of AI research, AI systems, expert systems, implemented a rule-based approach. These expert systems were conditionally explainable by their construction and design, in which concatenations of rules were used to reach conclusions. This made it possible to generate explanations by providing a detailed or abstracted collection of rule statements, as an explanation for a given conclusion.

The expert systems phase was essentially a matter of explaining the decisions of the systems to the user. It was about explaining the why, what, and how of an outcome of an AI system that produces an outcome.

The “why” explanations were about justifying the conclusions, the “how” explanations were about explaining how the system works, and the “what” explanation was about revealing the variables involved in the decision. The focus in this phase was to explain the functioning of the system. One of the major weaknesses of the expert systems of the time

was that they did not consider the user context - the focus was on providing chains of explanations in response to the questions of how and why.

A well-known example is the MYCIN expert system, which was used for medical diagnosis and was equipped with a rule-based inference engine as an explanatory component (the architecture of expert systems will be discussed in more detail in Chapter 5). In contrast, a deeper understanding and thus explainability must be that a user is able to understand how the underlying AI system works, how it was constructed and how the data used to develop and train (learn) the system fits the situations in which it is used.

However, the currently successful non-symbolic models of AI (e.g., deep learning models) do not provide explanations, but only numbers about the accuracy of the model. One possibility that is suitable to close this gap in terms of representation and inference is semantic technologies, such as the semantic web. These systems use RDF and OWL. The terminology of a semantic web can be used, for example, to explain explanations and the provenance of the information: providing the evidence for the information supports the explanation of what and why by explaining the background of the AI explanations to the user.

External Knowledge Bases

In the case of integration of external knowledge bases, the model provides the decision and also the explanation (based on the database). This method of explanation is model agnostic but can be used preferably for classification problems. An example is access to the PubMed database or, in the case of AISOP (s. Chapter 5.2.3) (van Aken et al., 2021, Holzinger et al. 2017), to the database of electricity providers, weather database, etc. Re_fish also provides for this explanation method - and can also use a company-internal knowledge base in order not to give away any competitive advantage.

Counterfactual Explanations

(s. e.g., Wachter et al, 2017; Stepin et al., 2021; Molnar,2019)

Chari et al. (2020) point out that in the recent past, more and more researchers are calling for AI systems to be equipped with explanation modules. Different systems provide different explanations to the user. For example, as presented above, expert systems provide trace-based explanations (see MYCIN). Chari et al. (2020) therefore call for the next generation of AI systems to be able to go beyond the why-what-how aspects of explanations and provide those that can interpret in terms of the user's setting, i.e., the user's ability to

understand, and the context. A minimum requirement, however, is to provide the provenance of the information so that the user can understand the reasoning or in order to aid comprehension.

The use of interpretable ML, as described in the previous chapter, in explaining the "what" of non-symbolic AI models, can therefore be seen only as an intermediate step. Modern AI systems must be able to provide users with explanations that allow for information attribution and provenance.

Among the motivations for these extensions is to improve the trustworthiness of the information represented in knowledge graphs (KGs), and to provide more context to users.

In certain user contexts, there is a need for personalised reasoning, so that the explanations produced by AI systems are reconsidered from the user's perspective and include components that can "pick up" users in their given context and, for example, train them or orient them to their cognitive model. These AI systems should be able to help users and enable trust in the system, as well as provide information relevant to the user's context.

Chari et al. (2020) define desirable properties of explanations, and these include the following guidelines:

- **Be understandable**

  This requirement demands that the explanation is understandable to a user, by using a terminology the user can understand; if the user cannot, then the system should have the capability to educate the user. Chari et al. (2020) are of the opinion that the system understandability can be significantly raised by providing user feedback and also considering the context the user is in.

- **Include provenance**

  This requirement refers to the fact that the more different content from different sources is processed in the respective AI model and used for a decision, the more necessary it is to provide the origin of the information. This is not only relevant for the collected information, but also for the domain knowledge used in the system and the methods with which the knowledge is obtained - therefore, the systems must also provide the causal information of the conclusion.

- **Appeal to user**

  Appeal to the user is a requirement understood by Chari et al. (2020) to mean that explanations provide facts at the required granularity, so that the user perceives them as resourceful and sufficient. Thus articulated, an explanation is resourceful if it contains content at the appropriate granularity and evidence to appeal to the user's cognition. An explanation is sufficient if the user can use it for their tasks. Chari et al. (2020) acknowledge that these requirements are difficult to verify in real time, and in the real world; here, they refer to the design phase and that these requirements are established during a requirements analysis and enter the design of the system. Moreover, according to the author, this is also a highly context-dependent issue.

- **Adapt to the users' context**

  In addition to the fact that the explanations have to be user-related, they have to be related to the current situation and context the user is in. Therefore, the explanations must use available information about the user (user profile -> stakeholder map) and meet the user's intentions and the right requirements for the user's explanation form; that is, they must connect to the user's mental model and align with the user's intention. Thus, an explanation can be a contrastive hypothesis that relates to the user's intention, or statistical evidence to provide more support to enhance a user's belief. Accordingly, the system must be able to provide different types of explanations, and these are presented below in the context of planning. In addition, the stakeholder table contains detailed information on this (see chapter 2.3.4 and 5.2.3).[28]

In 1991 and 1993, Swartout et al. (1991) and Swartout and Moore (1993) defined the Explainable Expert System Framework and formulated so-called "desiderata", which they consider relevant for the explanation of expert systems. However, these desiderata can also

[28] Another framework for explainability was provided by Lipton (2016); however, the focus of Lipton's framework is on machine learning models/systems and not on an expert system. Lipton's desiderates are for Machine Learning Model Interpretability (s. chapter 3.3.1).

be seen as architectural requirements for expert systems or knowledge-based systems. According to Swartout/Moore, an expert system must be accountable in order to be trusted - it must be able to explain its reasoning and justify its conclusions, much like a human expert. They argue that explanations and concerns must be taken into account if a system is to be desired. Otherwise, the system is unlikely to provide good explanations. Therefore, advances in explanatory capabilities and expert system architectures are closely related.

One main requirement (desiderata) is that an explanation facility imposes some strong requirements on the design of an expert system. It can be difficult or impossible to provide a system with adequate explanations unless those requirements are considered during system design (explanation by design). The five requirements are as follows (Swartout & Moore, 1993):

1. **Fidelity:** The explanation must be an accurate and reasonable representation of what the system does. An inaccurate or misleading explanation is worse than having no explanation at all. The interpreter for expert systems should be as simple as possible and have only a minimal number of special functions (e.g., mechanisms for inference under uncertainty). Special functions built into the interpreter are not part of the system's knowledge base. Therefore, if they are not supported by special built-in routines, they cannot be explained. Any changes or adaptations to the interpreter must also be made for these routines and are therefore a source of potential errors if they are not used.

2. **Understandability:** The explanation the system gives must be understood. Otherwise, the explanation is useless. Comprehensibility is made up of several components, such as the content, the creation of the explanations and the context in which they were produced. – Swartout & Moore (1991) found:
   - **Terminology:** Terms of the explanation must be understandable to the user (stakeholder) of the system, or at least the system must be able to define them in the user's familiar terms (analogy)
   - **User Sensitivity:** An Explanation presented by the system must consider the user's knowledge, objectives and goals, preferences, and concerns.
   - **Abstraction:** The system must be able to present the explanation on different levels of abstraction – depending on the user's needs or preferences. This usually comes with a change in terminology.

- **Summarization:** The explanation the system provides must be on different levels of detail without a change in terminology.
- **Perspectives:** The explanation given by the system must be from different perspectives; Swartout and Moore describe this as “form vs. function in a biological domain” or “safety vs. profitability in a finance domain” (Swartout & Moore, 1991).
- **Linguistic:** The explanations generated by the system should adhere to linguistic principles and constraints, and therefore sound “natural”.
- **Feedback:** If the user does not understand parts of the explanation, he should be able to receive further clarification.

3. **Sufficiency:** Function and terminology should be explainable and detailed enough to justify the decision. The systems should have enough knowledge to explain, depending on the sorts of explanations being offered by the system.
   - **Explanation about the behaviour of the system:** This explains how the system solved a particular problem, how a specific parameter impacted the outcome, and what the effect of a change in the data would be.
   - **Justifications:** This is about the rationale behind the actions and recommendations of the system.
   - **Preferences:** Here the explanation describes why one recommendation or strategy is preferred over the other. This requires knowledge about trade-offs and preferences involved in the selection.
   - **Domain explanations:** Explanation describing the problem domain itself.
   - **Terminology definitions:** These explanations answer questions about the meaning and terms of the system usage.

4. **Low Construction Overhead:** The explanation should not dominate the cost of designing AI. This desideratum is about how the system is designed. A system without an explanation is much easier to design and build than otherwise. The design should be as sophisticated so that there is less impact on the construction of the expert system. This includes all parts of the system lifecycle, as well as maintenance, etc.

5. **Efficiency:** The explanation system should not slow down AI significantly. This desideratum is about how the system is implemented and how it behaves in terms of runtime and costs.

Chari et al. (2020) adopt these requirements, partly in the presentation of their requirements. In addition, they derive requirements for AI systems from the five desiderata mentioned above:

- **Modularity**: An AI system should be modular so that it can adapt to the respective models and to the requirements of the users and scenarios. The AI explanation component should be able to access the other modules and provide the information that the user wants and needs.

- **Interpretability**: This requirement is about transparency in the sense that it enables the user to understand how the knowledge-based system works. If (as Chari et al. (2020) require) the models used in the system are not interpretable, the system must be able to use proxy methods to explain the models (see LIME, SHAP, Chapter 3.3.1).

- **Provenance support**: Chari et al. (2020), echoing the desiderata of Hasan and Gandon (2012), argue that AI systems should store the provenance of the information on which their models are based, using their metadata. They believe that incorporating provenance helps AI systems generate imaginative and sufficient explanations for users. It also allows the user or stakeholder to provide resources for further exploration.

- **Adapt to user's needs:** AI systems should be adaptive and interactive, adjusting their functioning and explanatory capabilities to the particular needs and contexts of the users. The different requirements and explanations result from the analysis of the relevant stakeholders (see Stakeholder Map). Through the modularity of the AI system, the system is able to provide explanations in different forms, which then enable the respective user to better understand and meet their needs.

- **Include explanation facilities:** The design of the explanatory capabilities should be part of the development phase, to ensure that the AI system is able to support their requirements. These explanatory capabilities refer to user interfaces, such as dialogue systems, with which different stakeholders, e.g., an expert or a user, can interact with the system. In addition to the explanation possibilities, the origin of

the information and feedback should be provided. Here, the context is extremely important (think of the provision of explanations in autonomous driving and in contrast, the decision-making described here in the context of corporate planning). It is therefore necessary to link the requirements with those of the requirements for the AI system.

- **Include/Access a knowledge store:** Under this point, knowledge-based systems capable of explanation should store the following types of knowledge:
    - Domain knowledge they use
    - The mental model of the users they address
    - The explanatory components that are generated
    - The incorporation of knowledge should be done through access to a knowledge store, either provided by the system or by another (external) source, which can contribute to and access its knowledge. In this context, the knowledge store is understood to be a knowledge graph (KG) or a semantic representation that can store the above-mentioned types or forms of knowledge.

- **Support compliance and obligation checks:** Explanatory knowledge-based systems should not only host or be able to access the above knowledge store. To ensure that the system complies with the relevant standards, rules and practices, the system should also store the codes of expert knowledge in the relevant domain.

The above-mentioned requirements for AI systems are supplemented by further requirements in Chapters 4 and 5 and thus serve, among other things, as requirements for the RA_Fish reference architecture. In addition to the architectural requirements, it is important to see the types of explanations in the context relevant to the stakeholders and to identify conceivable explanations, e.g., for the area of decision-making in corporate planning.

Wang et al. (2019) developed a framework for a user-centric, explainable AI. The idea of this framework is to bridge the algorithm-generated explanations and human decision-making theories by avoiding common biases. Their research spans the fields of cognitive psychology, philosophy, and decision theories to find patterns in how people think, make

decisions, and then seek explanations, but also to investigate cognitive factors that influence or affect decision-making. The framework is based on three approaches to explanations in explainable AI: the first consists of 'unvalidated guidelines' for approaches that provide little or no rationale for their use, so that the utility of the application to the user is unclear; the second concerns 'empirically derived taxonomies', for approaches that are derived from surveys of the explanatory needs in question (for these approaches in particular, there is still much to be studied and developed, e.g., to what extent something should be explained so that the explanation is not perceived as burdensome); the third category of approaches are the "psychological constructs from formal theories", and focus on the work of Miller (2019), Hoffman and Klein (2017), who examine relevant theories from philosophy, cognitive psychology and the social sciences, but lack the translation of insights into explainable AI systems. [29]

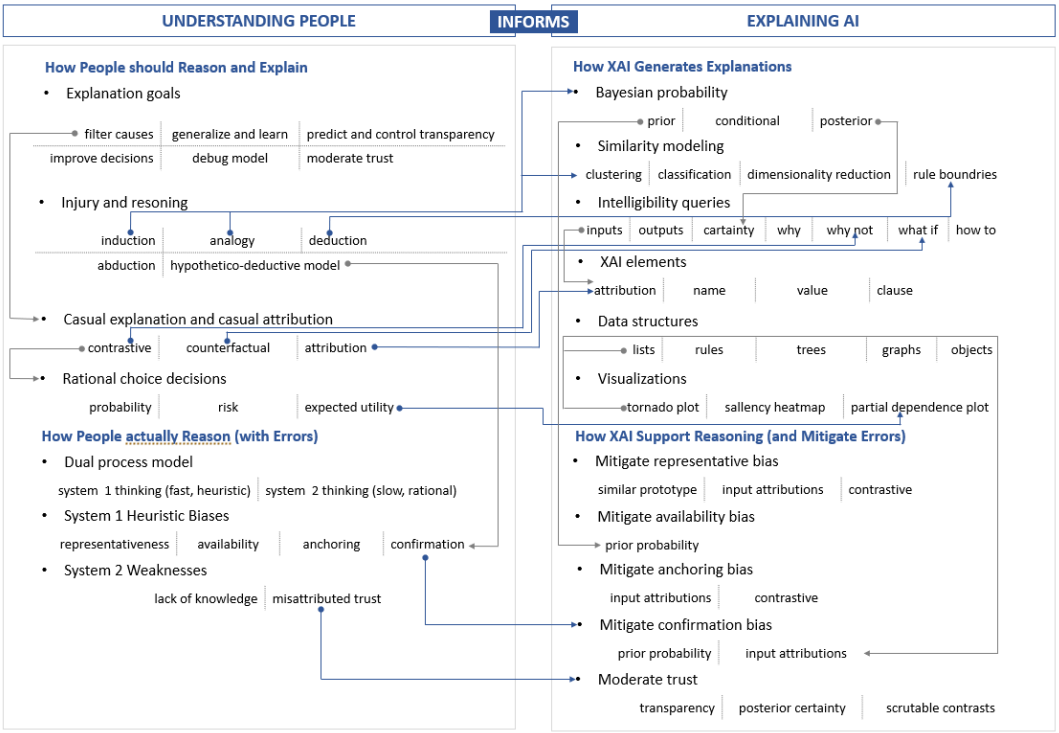

Figure 61: Framework by Wang, D. et al (2019)

The figure 61 shows the framework of Wang, et al. with the left being the “people” side and the right the “explaining AI” side. The connection through the lines shows how human explanations can be described by an explainable AI system.

[29] Wang, D. et al. (2019)

The reasons why users (people) want an explanation are triggered by a (subjectively felt) deviation from an expected behaviour, such as a curious, inconsistent, discrepant, or overall anomalous event with respect to behaviour. Another possible trigger might be that users (people) want to monitor for an expected important or costly event. Miller (2019) found that the main reason people want explanations is to facilitate learning by allowing the user to filter a small set of (reasonable) causes to simplify their observations, after which these observations (inferences) are generalised into a conceptual model, to help them predict and also control perceived future phenomena (Wang, et al., 2019; Miller, 2019).

The reason for the explanation of AI systems can be derived from research by Nunes and Jannach (Nunes & Jannach, 2017). Therefore, explanation by AI systems is to support transparency – to provide users with information about the inner functionality or state of AI systems (models). If AI is used, like in our case for supporting decision making, users would like to get explanations to improve their further decision making. If a system behaves in a way other than expected -- erroneously, users seek to get explanations for proof and to debug and test, to identify the reason for the fault, take over control and adjust. Overall explanations are being used to improve trust and moderate trust among different stakeholders (levels) (Wang, et al., 2019).

The right side of the diagram shows how XAI can generate explanations. For example, there is the following connection between the way people should reason and explain and XAI. People use inductive reasoning based on observations to understand events and test hypotheses (connection between induction and Bayes probability). In this case, XAI uses Bayesian probability theory. Bayes' theorem describes the possibility of an event depending on prior knowledge, prevailing conditions, especially prior and subsequent probabilities, and probability. Understanding the probabilities of outcomes can inform the user about expected utility. The Bayes theorem helps decision makers make decisions, as reason by taking the frequency of events into consideration. For the design and build of the reference architecture, it is important to understand, which explanations are relevant in the context of decision support in corporate planning and how an explainable AI system can provide these explanations. In Chapter 2, the stakeholders for planning were identified and their needs for explanation, as well as the specific explanation type, were presented. The findings are shown in Chapter 2.3.4 and are completed in the following.

Comprehensibility questions (s. figure 61 the connection between “causal explanation and causal attribution” and here in particular contrastive and “why not” and counterfactual and

“what if”). Causal explanation refers to the selection of certain reasons to explain an observation against the background of existing knowledge. Users can ask why not, to understand why the foil did not happen (contrastive). Counterfactual, on the other hand, refers to explaining what needs to change for an alternative to happen. Research by Lim and Dey (2009) found that the Why and Why Not explanations were the most effective in terms of understanding the system and trust.

Other methods of explaining XAI systems are XAI elements, such as attribution, by showing which feature has what relevance to the outcome (see e.g., SHAP, LIME). Furthermore, there are Data Structures, with the simplest way to create explanations, namely through the use of lists, or also the use of rules and decision trees, etc. Visualisations also serve as explanations, such as the use of charts, heatmaps, partial dependence plots, etc.

Chari et al. (2020) derive a catalogue of explanation types from this (s. table 12):

Explanation Type	Definition
Case-based	Answering the question: "To what other situations has this recommendation been applied?" in the sense of a case-based explanation. The user is presented with results from previous cases to confirm the current decision. When using case-based explanations, a system must remember the explanations of previous cases or be able to draw inferences from previous cases by inference. Accordingly, case-based explanations may involve analogical reasoning and rely on similarities between features of the (historical) case and the current situation.
Contextual	Context-based explanations or answering the question: "What broader information about the current situation has led you to make this recommendation now?" The answer to this question or the explanation refers to information that goes beyond the user's current situation. The AI system in question must be context-aware and include information about "the user's tasks, important user attributes, organisational environment, and technical and physical environment".
Conrtrastive	Answering the question "Why should I administer this new drug and not the one I would normally prescribe?" Contrastive explanations explain the outcome as a contrast by relating it to an outcome that did not occur.
Counterfactual	Counterfactual explanations are about answering the question of what results would have been achieved with inputs other than those used. would have been achieved with inputs other than those used. In this context, counterfactual explanations are causal in nature. They are governed by patterns of a particular kind of causal dependence.
Everyday	Everyday explanations are explanations based on common understanding and knowledge of how the world works. They help to understand why certain facts (events, properties, decisions, etc.) have occurred.
Scientific	Scientific explanations use the results of rigorous scientific methods, such as observations and measurements, to explain something.
Simulation-based	Simulation-based explanations are explanations based on the imaginary or implemented imitation of a system or process and the respective results. The use of simulations can be carried out numerous times, e.g. when using Monte Carlo simulations, and the mechanisms can often be directly observed and understood in the simulation.
Statistical	Answer to the question "What percentage of x results from applying y". Statistical explanations use the frequency of occurrence of a certain event under certain conditions to represent the result.
Trace-based	Trace-based explanations by showing the underlying sequence of steps, i.e. a 'chain of reasoning' that the system used to achieve a certain outcome- the functioning of the system/justification of the decision.

Table 12: Explanation types

The most relevant explanations for planning and automated decision-making are the contextual, contrastive, and counterfactual explanations, but certainly with regard to scenario planning, the simulation-based explanation.

Chari et al. propose approaches that can be seen as directions for research. For explainable AI, they see casual methods, neuro-symbolic AI systems, and representation techniques to model the explanation space and enable trustworthy data exchange, or emerging approaches that include distributed ledger technologies, for instance.

One of the most important approaches is causality, which has been researched since about last decade of the last century and pursued independently of semantic technologies. Many researchers consider the use of causality to be crucial in representing explanations to stakeholders. For example, causal inference can be used to explain when a change in performance occurs due to a failure in the real system. The system should therefore encode causal knowledge, which one of the key causality researchers, Judea Pearl, believes is lacking in association-based AI methods. Current ML methods can provide associative explanations, but are unable or weak to use counterfactual explanations, as they must then have causal knowledge.

However, answering questions about intervention knowledge requires that an AI system also understands and encodes knowledge about the world from the data from which it derives a decision. For counterfactual questions, the system would need to know or understand cause-effect relationships.

Chari et al. believe that the semantic representation of causal structures, as indicated in the figure, would lend to the development of causal, neuro-symbolic integrations.

In this Chapter, 3.3, explainable AI was presented generally, as well as in relation to its application in the field of machine learning. This field, however, is only a sub-area of the entire AI (the non-symbolic AI). Therefore, explainability also only covers a small sub-area. More promising are those that use a combination of symbolic and non-symbolic approaches and look back at the extensive research in the area of explainability of expert systems. The use of knowledge graphs is a promising approach for storing knowledge and using it, for example, in the context of inference. The statements made in Chapters 3.3, 3.3.1, 3.3.2 and 3.3.3 regarding requirements for the architecture of AI systems are used in Chapters 4 and 5 to create a reference architecture.

3.3.3 Neuro-symbolic Systems of Explainable AI

According to Garcez & Lamb (2023), the explainability of neurosymbolic models is about the extraction of compact but correct and comprehensive knowledge. Neural networks do not seem to be able to do this, which is why there are, for example, the attributive approaches mentioned above. However, as Garcez & Lamb (2023) argue, a large knowledge graph database is no easier to explain than a neural network. The better explainability of the knowledge graph, however, results in comparison with, for example, a local explainability, because here knowledge graph databases provide a trace in the local explanation - in the sense of a proof of history to show how the result was achieved. In the field of neurosymbolic AI, the primary focus so far has been on investigating the accuarcy of the connection of the extracted knowledge in relation to the neural network. According to Garcez & Lamb (2023), there are essentially 2 possible approaches for explaining XAI in neural-symbolic systems:1. symbols are translated into a neural network, and one seeks to perform reasoning within the network. 2. a more hybrid approach than 1. where the neural network interacts with the symbolic system for reasoning. A third possible approach, in which knowledge is provided by expert knowledge, is disregarded here.

In the first approach, it is necessary to have a symbolic description of the network in order to explain and thus trust or interact with the system. The second approach requires an interface through which the two systems can communicate with each other. This, according to Garcez & Lamb (2023), is currently the 'best solution' to combine reasoning and learning in AI, especially because of the differences that currently exist between the two - the discrete and exact nature of symbolic reasoning and the continuous and approximate nature of statistical learning (Minervini et al., 2020).

3.4 Ethical AI, Law and Regulatory Requirements of Explainable AI[30][31]

As described in chapter one, people only trust the decisions and suggestions of AI if they understand them. Only then can the full potential of AI models be used in the context of

[30] This section is based mainly on the article by Bejger & Elster (2020)

[31] S. law on the regulation of artificial intelligence of 14.06.2023 https://www.europarl.europa.eu/news/de/headlines/society/20230601STO93804/ki-gesetz-erste-regulierung-der-kunstlichen-intelligenz , accessed 18.06.2023

decision support, as well, e.g., using stochastic methods such as artificial neural networks in the context of machine learning (ML), in which human intelligence is imitated and neutral recommendations, decisions, and actions are made in this way, whereby larger amounts of data can be analysed more quickly by appropriate AI models than by a human. As more and more decisions are made by AI, including in certain critical areas such as justice, lending, personnel selection, medicine, transportation, or the military, transparency is a foundation, a "conditio sine qua non", for trust in AI decisions (Holzinger, 2018; DARPA-BAA-16-53, 2016; Laat, de, 2017; Waltl & Vogl, 2018).

The recommendations and decisions by the stakeholders of AI models include not only the developers or users of AI in the company, but also national and supranational society. This is not unique to AI in planning. In principle, this applies to the application of AI models in the economic environment. Understandability and transparency are also the basis for accountability; therefore, many authors demand accountability at different levels, e.g., that of society, state, or enterprise (Sauerwein, 2019).

Aside from the above, there are different phases in the lifecycle of AI models, each involving different stakeholders who influence the model or are impacted. At a technological level, "accountability by design" is a requirement which must be implemented. The designers commit themselves, for example, to Responsible Research and Innovation (RRI), which works like a Hippocratic Oath for developers. Another possible form might be the use of appropriate or adjusted development methods, e.g., through enrichment of the Cross Industry Standard Process for Data Mining (CRISP-DM) process and the requirements demanded by ethical guidelines or regulations. Companies that use AI in their products need to document their social responsibility in the meaning of AI governance, within CSR (Corporate Social Responsibility). Future developments could include that when following given ethical guidelines in conducting business, as well as in the development process, companies might do an audit and get a certification by the government for their product, which uses the ML component as a kind of product feature. In the sense of meta-responsibility, the state has the task of governance by establishing control frameworks, thereby establishing regulations such as those are already in place for the protection of privacy (GDPR (EU) 2016/679). The same applies on a supranational level for the European Union.

The overall question which is discussed in philosophy and computer science, among other fields, is whether transparency and accountability also ensure that decisions do justice to ethical and moral considerations. As computer or AI models and algorithms do not have a per se built-in value system in ethical terms, this is questionable. There is no "built-in" morality in computers or in AI models. Therefore, AI models can instead be seen, as Hannah Arendt puts it, as conscientious instances in the sense of obedient executors; she called them "useful idiots" in connection with the henchmen of the Nazi regime. According to Arendt, morality (and thus ethics) only emerges through an inner "dialogue" with oneself, relating to oneself and the preservation of dignity, as well. An AI model, however, has no "dignity", and does not conduct an inner dialogue. All of that is completely alien to an algorithm. Therefore, only the people who design and use them can be made responsible for the morality of AI algorithms and models (Arendt, 2007; Reichmann, 2019; Bostrom & Yudkowsky, 2014; Mittelstadt et al., 2016).

At the process level, essential questions remain unanswered and AI applications in domains like medicine, justice, personnel selection, etc. cannot be used without the "human entity in the loop" doing the final decision (Dutton, 2008).

At the European Union level, a group of High-Level Experts was formed to define reasonable regulations and informal standards. This is also a task in various other countries and organisations (High-Level Expert Group on AI, 2019).

On the 8th of April 2019, the "High-Level Expert Group on AI" for the European Union presented their so-called "Ethics Guidelines for Trustworthy Artificial Intelligence". These guidelines followed the publication of the first draft guidelines of December 2018. The Group had received more than 500 comments through open consultation and considered them for the 2019 guidelines.

According to these Guidelines, trustworthy AI should be:

(1) lawful - respecting all applicable laws and regulations

(2) ethical - respecting ethical principles and values

(3) robust - from a technical perspective while considering its social environment

The Guidelines put forward a set of seven essential requirements that AI systems should meet to be considered trustworthy. A specific assessment list aims to help verify the application of each of the requirements:

- **Human agency and oversight:** AI systems should empower humans, allowing them to make informed decisions and fostering their fundamental rights. At the same time, proper oversight mechanisms must be ensured, which can be achieved through human-in-the-loop, human-on-the-loop, and human-in-command approaches.

- **Technical Robustness and safety:** AI systems need to be resilient and secure. They must be safe, ensure a fallback plan in case something goes wrong, and be accurate, reliable, and reproducible. That is the only way to ensure that unintentional harm can be minimized and prevented.

- **Privacy and data governance:** Besides ensuring full respect for privacy and data protection, adequate data governance mechanisms must also be upheld, considering the quality and integrity of the data, and ensuring legitimised access to data.

- **Transparency:** Humans must be aware that they are interacting with an AI system and must be informed of its capabilities and limitations. The data, system, and AI business models should be transparent. Traceability mechanisms can help achieve this. Moreover, AI systems and their decisions should be explained in a manner adapted to the stakeholder concerned.

- **Diversity, non-discrimination, and fairness:** Unfair bias must be avoided, as it could have multiple negative implications, from marginalising vulnerable groups to exacerbating prejudice and discrimination. To foster diversity, AI systems should be accessible to all, regardless of disability, and involve relevant stakeholders throughout their entire life circle.

- **Societal and environmental well-being:** AI systems should benefit all human beings, including future generations. It must hence be ensured that they are sustainable and environmentally friendly. Moreover, they should consider the environment, including other living beings, and their social and societal impact should be carefully considered.

- **Accountability:** Mechanisms should be put in place to ensure responsibility and accountability for AI systems and their outcomes. Auditability, which enables the assessment of algorithms, data, and design processes plays a key role therein, especially in critical applications. Moreover, adequate and accessible redress should be ensured.

The requirements for AI due to ethics, laws and regulations are subjects of intensive research. For the work presented here, these requirements are to be considered as architectural constraints, in the sense of basic requirements that the AI systems must fulfil. In this context, particular reference is made to the possibility of auditing AI systems, with regards to the requirements of ethics, laws, and regulations, as mentioned in Chapter 3.3.2, and the consideration of the "explainability by design" requirement, namely considering the explainability of the AI systems during the stage of their design.

3.5 Mapping the Stakeholders and their Requirements

In this section, the aim is to bring together the stakeholders of corporate planning and the stakeholders of XAI in order to obtain the requirements for an AI system in terms of explainability. For this purpose, the terms from 2.3.4 and the XAI stakeholders to be presented in this chapter are mapped together.

If one follows the presentation by Bejger/Elster (2020), one recognises that further stakeholders can be outside the company, such as the owner of the company, the regulator, the auditors, the regional or country-related social society (e.g., society in Germany) and the supranational society (e.g., the European Union).

Requirements/ Constraint	Description	Stakeholder	Solution	References
Acceptance	Improve acceptance of systems	Regulator, Deployer	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI
Accountability	Provide appropriate means to determine who is accountable	Regulator	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
Fairness	Assess and increase a system's (actual) fairness. Affected,	Regulator	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
Informed Consent	Enable humans to give their informed consent concerning a system's decisions	Affected, Regulator	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
Morality/Ethics	Assess and increase a system's compliance with moral and ethical standards	Affected, Regulator	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
Responsibility	Provide appropriate means to let humans remain responsible or to increase perceived responsibility	Regulator	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
Transparency	Have transparent systems	Regulator	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
Trustworthiness	Assess and increase the system's trustworthiness	Regulator	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
Legal Compliance	Assess and increase the legal compliance of a system	Deployer	Audits, quality checks, following the standards	Design and development guidelines, architectural principles, systems tests
Safety	Assess and increase a system's safety	Deployer, User	Audits, quality checks, following the standards	Design and development guidelines, architectural principles, systems tests
Trust	Calibrate appropriate trust in the system	User, Deployer	Audits, quality checks, following the standards	Design and development guidelines, architectural principles, systems tests
Accuracy	Assess and increase a system's predictive accuracy	Developer	Design and development guidelines, architectural principles, systems tests	
Effectiveness	Assess and increase a system's effectiveness; work effectively with a system	Developer, User	Design and development guidelines, architectural principles, systems tests	
Efficiency	Assess and increase a system's efficiency; work efficiently with a system	Developer, User	Design and development guidelines, architectural principles, systems tests	
Robustness	Assess and increase a system's robustness (e.g., against adversarial manipulation)	Developer	Design and development guidelines, architectural principles, systems tests	
Performance	Assess and increase the performance of a system	Developer	Design and development guidelines, architectural principles, systems tests	
Verification	Be able to evaluate whether the system does what it is supposed to do	Developer	Design and development guidelines, architectural principles, systems tests	
Transferability	Make a system's learned model transferable to other contexts	Developer	Design and development guidelines, architectural principles, systems tests	
Debugability	Identify and fix errors and bugs	Developer	Design and development guidelines, architectural principles, systems tests	
Autonomy	Enable humans to retain their autonomy when interacting with a system	User	Design and development guidelines, architectural principles, systems tests	
Confidence	Make humans confident when using a system	User	Design and development guidelines, architectural principles, systems tests	
Controllability	Retain (complete) human control concerning a system	User	Design and development guidelines, architectural principles, systems tests	
Education	Learn how to use a system and system's peculiarities	User	Design and development guidelines, architectural principles, systems tests	
Privacy	Assess and increase a system's privacy practices	User	Design and development guidelines, architectural principles, systems tests	
Satisfaction	Have satisfying systems	User	Design and development guidelines, architectural principles, systems tests	
Science Gain	scientific insights from the system	User	Design and development guidelines, architectural principles, systems tests	
Usability	Have usable systems	User	Design and development guidelines, architectural principles, systems tests	
Usefulness	Have useful systems	User	Design and development guidelines, architectural principles, systems tests	
Security	Assess and increase a system's security	All	Regulations, design and development guidelines, architectural principles, systems tests	

Table 13: Stakeholder Map B- Stakeholders and their requirements/constraints

A stakeholder is "an individual, team, or organisation (or classes thereof), with interests in (or concerns relative to) a system" (Lankhorst, 2017).

The table shows the typical requirements/constraints of the stakeholders of an AI system (s. Chapter 3). The requirement/constraint is named and then described. The stakeholders here are the Regulator (governmental – legal institution), deployer (in our case, the company principal), the user (in our case, business users -- the demand planner, production planner, procurement planner, distribution planner, financial planner, and planning expert as knowledge engineer), the developer (AI developer, administrator) and all those affected, which means an undefined group.

Requirements/ Constraint	Description	Stakeholder	Stakholder Mapping to Corporate Planning
Acceptance	Improve acceptance of systems	Regulator, Deployer	Legal institution, principal (board of directors)
Accountability	Provide appropriate means to determine who is accountable	Regulator	Legal institution, principal (board of directors)
Fairness	Assess and increase a system's (actual) fairness Affected,	Regulator	Legal institution, principal (board of directors)
Informed Consent	Enable humans to give their informed consent concerning a system's decisions	Affected, Regulator	Legal institution, principal (board of directors)
Morality/Ethics	Assess and increase a system's compliance with moral and ethical standards	Affected, Regulator	Legal institution, principal (board of directors)
Responsibility	Provide appropriate means to let humans remain responsible or to increase perceived responsibility	Regulator	Legal institution, principal (board of directors)
Transparency	Have transparent systems	Regulator	Legal institution, principal (board of directors)
Trustworthiness	Assess and increase the system's trustworthiness	Regulator	Legal institution, principal (board of directors)
Legal Compliance	Assess and increase the legal compliance of a system	Deployer	Board of directors
Safety	Assess and increase a system's safety	Deployer, User	Board of directors, business user
Trust	Calibrate appropriate trust in the system	Deployer, User	Board of directors, business user
Accuracy	Assess and increase a system's predictive accuracy	Developer	AI developer, AI system administrator
Effectiveness	Assess and increase a system's effectiveness; work effectively with a system	Developer, User	AI developer, AI system administrator, business user
Efficiency	Assess and increase a system's efficiency; work efficiently with a system	Developer, User	AI developer, AI system administrator, business user
Robustness	Assess and increase a system's robustness (e.g., against adversarial manipulation)	Developer	AI developer, AI system administrator
Performance	Assess and increase the performance of a system	Developer	AI developer, AI system administrator
Verification	Be able to evaluate whether the system does what it is supposed to do	Developer	AI developer, AI system administrator
Transferability	Make a system's learned model transferable to other contexts	Developer	AI developer, AI system administrator
Debugability	Identify and fix errors and bugs	Developer	AI developer, AI system administrator
Autonomy	Enable humans to retain their autonomy when interacting with a system	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Confidence	Make humans confident when using a system	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Controllability	Retain (complete) human control concerning a system	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Education Learn how to use a system and system's peculiarities	Learn how to use a system and system's peculiarities	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Privacy	Assess and increase a system's privacy practices	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Satisfaction	Have satisfying systems	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Science Gain	scientific insights from the system	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Usability	Have usable systems	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Usefulness	Have useful systems	User	Strategic planner, demand planner, procurement planner, production planner, distribution planner, financial planner
Security	Assess and increase a system's security	All	All

Table 14: Mapping the stakeholder to corporate planning

3.6 Summary

Chapter 3 introduced the topic of artificial intelligence and two approaches in particular - non-symbolic AI and symbolic AI. In the key Chapter 3.3 Explainable AI, the aim was to show that research into the explainability of AI models is much older than the currently fashionable approaches of machine learning and non-symbolic AI; it initially dealt with the explainability of expert systems. Moreover, research into explainability is generally much older than AI research and has already occupied researchers from the fields of philosophy, psychology, and sociology. The user and stakeholder requirements for explainability and the requirements of the business planning stakeholders have been mapped together to create the reference architecture in Chapter 5 in such a way that the requirements are met.

Finding 13: In Chapter 3.2, the technical perspective of artificial intelligence was presented, after the economic perspective was presented in Chapter 1.1 and 1.2. In this chapter, the different areas of AI, machine learning, deep learning, knowledge enabled systems and finally the promising approach of neuro symbolic systems, a combination of deep learning and symbolic AI, were presented. Then, in Chapter 3.3, the area of XAI was presented.

Finding 14:
Chapter 3.4 briefly presents the ethical, legal and regulatory requirements for Explainable AI. The risks of AI have been recognised and are already subject to regulation in Europe, for example in the area of the EU GDPR, PE-6-2023-INIT, etc.

"The Friend's Apartment was inside a townhouse. From the window of its Main Lounge I could see similar townhouses on the opposite side of the street. There were six of them in a row, and the front of each had been painted a slightly different color, to prevent a resident climbing the wrong steps and entering a neighbor's house by mistake." (Ishiguro, Kazuo (2021). Klara and the Sun. Chapter 4)

4. Design of a Reference Architecture for Explainable AI

4.1 Introduction

In information systems research, as well as in software development, reference models are used as templates to form instances using generally valid models, which are based, for example, on empirical knowledge and so-called best practices, which then in turn cover special domains or use cases. Reference modelling has been part of information systems research for some time, and there is a wide body of research on the use of such models and on the methodology and forms of their development. For this reason, there are also several different approaches to doing so; the choice of method should be made on the basis of the specific use case, as there is no standard methodology (Pescholl, 2011).

One of the main ideas of this thesis is to develop a domain-specific reference model, which is a reference architecture for explainable AI in the context of business planning and decision making, with a focus on the process industry. This comprises the distinct use case, thus the methods for developing a reference model with respect to a reference architecture must be geared towards it.

Reference models have a normative character, and this is one of their advantages (Schütte, 1998). Reference models can be used as a recommendation (or a kind of blueprint) that is useful for the development of concrete models. Reference models thus represent a solution for an abstract class of problems. According to Pescholl (2011), the process of reference modelling can be divided into two sub-processes: firstly, there is the construction of a reference model which is intended to function as a pattern and thus a generally valid solution for related problems, and secondly, the application of a reference model, in which specific models or problem solutions are developed for a particular application case on the basis of the pattern (derived from the first sub-process) (Schütte, 1998).

According to Schütte (1998), since the validity of a reference model cannot be proven due to its prescriptive and normative character, other quality criteria must be used to evaluate it. These quality criteria are usefulness and applicability, which therefore presuppose the high quality and quantity of use – it follows that the process of creating reference models is of particular importance. After all, a design is the building of structures from components or building blocks through the application of design principles. The making of reference models is a creative activity and a form of design if it is carried out methodically and systematically (Brocke, vom, 2003). Modelling must consider formal and substantive aspects. These are set out, for example, in the Principles of Proper Modelling (Becker et al., 1995).

4.2 Theoretical Basis of Reference Architectures

First and foremost, it is necessary to define what architecture is. There is an oft-quoted, ancient definition from the Roman architect Marcus Vitruvius Pollio; in his ten books on architecture, "De architectura libri decem", he defines the three most important requirements of architecture:

- Firmitas (solidity),
- Utilitas (usefulness), and
- Venustas (beauty).

All three requirements would have to be considered equally. Furthermore, he defines that there are six basic terms for the object of architecture: "ordinatio", "dispositio", "eurythmia", "symmetria", "decor", and "distribution" (Howe, 1999).[32]

In terms of software architecture, these requirements for a software system can be understood to mean that the product must fulfil the functional and non-functional requirements of the stakeholders. The software system must be stable concerning the required quality characteristics, e.g., those related to the number of simultaneous users, longevity, and adaptability to future requirements, so that further developments are possible. The software system should be structured both externally to the user and internally to the developer. On the one hand, this should enable intuitive use and further development (Gharbi et al., 2020).

[32] Vitruvius. Ten Books on Architecture

There are also numerous definitions for the term software architecture (SEI, 2010). In this work, the definition is based on IEEE Standard 1471, which defines architecture as the “fundamental concepts or properties of a system in its environment, embodied in its elements, relationships, and the principles of its design and evolution” (ISO 1471, 2011).

Besides the term software architecture, there is another term, namely enterprise architecture, which can be defined as "a coherent whole of principles, methods, and models that are used in the design and realisation of an enterprise's organisational structure, business processes, information systems, and infrastructure". In this thesis, software architecture is seen as a procedure for the development of a software application; the enterprise architecture can provide the framework or context for the software architecture, in the sense of constraints and guidelines for individual software projects (Lankhorst, 2017).

Another important term is that of the reference model and its distinction from the term “reference architecture”. According to Fettke/Loos (2004), there is no uniform definition of the term "reference model". They distinguish, for example, between the mapping-oriented model as understood according to Hars (1994), with the characteristic that a reference model is helpful in the design of other models. On the other hand, there is Schütte's (1998) definition, which contrasts the purely illusion-oriented definition with the fact that the designer of the reference model exerts a significant influence on the constitution of reality, and therefore understands reference models as a recommendation that serves as a point of reference in the design of information systems. Vom Brocke (2003) emphasises that the degree of generality and the recommendation character are difficult to determine, so that in extreme cases, a modeller declares his model to be a reference model but it is not used, or that a broad group of users recommends the model for reuse and accepts the modelling, but it is not explicitly regarded as a reference model (Fettke & Loos, 2004).

In contrast to a stringent definition, Fettke and Loos (2004) propose a systematisation of reference models, based on various characteristics of quality. For example, they distinguish between reference models as a phenomenon of a field by examining existing reference models and reference models that modellers create as theoretical constructs. The latter can be differentiated according to reference models into a terminological apparatus, a set of singular statements, a set of general statements, a technique, and a set of normative statements. Since the existing models do not clearly show a single quality feature, they are in most cases seen as hybrids between terminological apparatus and singular and general

statements. In the following, after Schütte, a reference model is seen as a theoretical construct of a modeller, which is a recommendation that serves as a reference point in the design of information systems (Schütte, 1998). These theoretical constructions can be understood as a terminological apparatus representing a set of concepts that constitute a collection of terms or a conceptual frame of reference for a subject area (this is congruent with the term ontology in computer science). Another property of the reference model term used here is the set of general statements. The reference model describes a single subject area (one company, several companies) and a whole class. The respective construction method decides whether this description is obtained deductively or inductively (Fettke & Loos, 2004).

Reference models are created by applying reference modelling methods, using reference languages. The context must be considered: the modelling changes are part of a specific real modelling situation. They are therefore subject to psychological, social, organisational, technical, or economic factors, and so forth. Fettke and Loos (2004) mention among examples of organisational factors the power position of the modellers (persons of the modelling agency), and as technical factors -- the choice of modelling tools, entities of companies and general requirements of all stakeholders for the information systems (in this work, especially for trustworthy AI systems).

The terms reference architecture and reference model are not clearly distinguished in the literature and in practice and have the same relationship to one another as architecture and model. They can apply generically across the board, or specifically to an individual company. This work is about developing a generic view, rather than a specific one related to an individual company. As the TOGAF framework describes it, "a generic reference architecture provides the architecture team with a blueprint for an organisation-specific reference architecture that is customised for a particular organisation". For example, a generic reference architecture may indicate that there is a need for data models. An example of a reference architecture is the IT4IT reference architecture, which also defines a common information model for IT management. Another example is the TM Forum eTOM and SID as an organisation-specific reference architecture. Therefore, a reference architecture can be seen as a reference model for a class of architectures (TOGAF, 2022).

The German “Gesellschaft für Informatik” defines reference architectures as follows: “Reference architectures are proven, generic software architectures for specific application domains, such as automotive or e-commerce. They apply across product and company

boundaries. They describe reusable structures, components, interfaces, general design rules and infrastructures for software systems in the respective" (Reussner & Hasselbring, 2008, S. 319)

4.3 Methodology to Develop Reference Architectures

An essential step in the development of reference architectures is the definition and application of a rigorous methodology for their creation. There are different approaches in the literature, perhaps as many as there are definitions of architecture and reference architecture.

In this work, the methods of Galster & Avgeriou (2011), Nakagawa et al. (2014), Bass et al. (2022), and the Architecture Development Method (ADM) of TOGAF (The Open Group Framework) are used. While the first three approaches are intended for the creation of a software architecture, the ADM is intended for the creation of an enterprise architecture. As already described in Chapter 4.2, the creation of a software architecture can be seen in the context of an enterprise architecture, so that these approaches can be combined. This is all the truer, since no statements are made about implementation when creating a reference architecture.

4.3.1 Methods to Develop a Reference Architecture

Galster & Avgeriou (2011) developed and proposed a six-step approach to building a reference architecture (or RA, short for 'reference architecture'; Reidt (2019); Galster & Avgeriou (2011)) Their approach builds on existing reference architectures created and used in practice, as well as on the basis of reference architectures found in the literature. They divide the approach into two main parts: firstly, it consists of "ensuring empirical grounding" (steps 1 - 5), and secondly "ensuring empirical validity" (step 6).

Part I, ensuring empirical grounding:

1. Decision on the type of Reference Architecture

   By using a classification schema, the type of reference architecture is defined (Galster & Avgeriou, 2011; Angelov, 2009). This classification differs in two features. First, it is based on the goal of the reference architecture, whether it serves standardisation or

is to be used as a kind of "blueprint" for the development of (specific) architectures. Secondly, it concerns whether the reference architecture is practice-oriented and describes proven architectures and solutions, or whether the architecture describes future solutions that do not currently exist (Reidt, 2019).

2. Selection of a Design strategy

In this step, it needs to be defined whether the reference architecture is built using best practices and on-project experiences or is to be built completely (or partially) from scratch.

3. Empirical acquisition of data

In this step, it is determined where the data for the creation of the reference architecture should come from. As well, a distinction must be made here as to whether the reference architectures are more practice-oriented or research-oriented. The stakeholders must be differentiated according to whether they provide information for the creation (or instantiation) of the reference architecture, or the stakeholders who apply/implement the designed reference architecture.

4. Construction of the RA

This step is about modelling the reference architecture. Various views are used, and a distinction is made between elements that are used in all instantiations of the reference architecture, i.e., which represent a so-called core component, and those that are present or not, depending on the instantiation.

5. Enabling RA variability

In this step, the variability in the instantiation of the reference architecture is made possible by using specific elements.

Part II, ensuring empirical validity:

6. Evaluation of the Reference Architecture

Evaluation of the reference architecture depending on the type of architecture -- whether its usability or the possibility of adaptation within the instantiation is in the foreground (Galster & Avgeriou, 2011, Reidt, 2019).

The second approach to be presented was developed by Nakagawa et al. (2014) and consists of a four-step approach for the construction, representation, and evaluation of a reference architecture (Nakagawa et al., 2014).

1. Identification of Information Sources.

In this first step, the sources of information are identified. All the elements in this step must be addressed by the reference model of the specific reference architectures. Nakagawa et al. distinguish between four different sources: (I) Investigation of the reference architecture literature. The reference architecture must address business rules, architectural styles (addressing quality styles of the reference architecture), best practices of software developers (architectural decisions, domain constraints, legislation, regulations, and standards), and the software elements that support the development of systems for a specific domain. (II) Knowledge contained in reference architectures. Investigate research on reference architectures within literature (III) Knowledge contained in software architectures. The knowledge contained in software architecture is about five main elements: problem domain, decisions, solution fragments, systems design, and implementation. (IV) Generic models of software systems. Usage and investigating of generic models of software systems. These models have been partially used as an important framework for software systems development. In this first step, all relevant information sources are identified, selected, and investigated. All required information about processes, activities, and tasks, which the information system should later support, must be gathered. Potential information sources are customers, users, and developers – all stakeholders (software systems, publications, domain ontologies, etc.). The knowledge about the domain should be more comprehensive than the one developing a specific architecture. For gathering the information, techniques like interviews, questionnaires etc., can be used.

2. Architectural Analysis -- identification of elements

To gather the definitions of the reference architectures -- extracting and using the elements from the information sources that could become part of reference architectures. The elements will be extracted; all elements contained in the definitions of reference architectures are analysed, summarised, and grouped. Nine elements were found: (i)

business rules (also related to functionalities, business contexts, and the domain problem); (ii) architectural styles (related to foundation, enterprise architecture, abstract framework, and generic architecture); (iii) communication elements (related to data flows and an organisation's Message Bus); (iv) software elements (related to supporting artifacts); (v) domain terminology (related to concepts); (vi) best practices; (vii) architectural decisions; (viii) domain constraints; and (ix) domain request (including domain legislation, regulations, and standards) For gathering the knowledge contained in reference architectures, three elements were identified, namely technical elements, the business model, and customer needs. For the knowledge contained in software architectures, five main elements were taken into respect: problem domain, decisions, solution fragments, systems design, and implementation. Investigating on the generic models of software systems, there is the following:

3. Design of the RA- architectural synthesis

In this step, the architectural description is built as a general framework. This description will be done according to four different groups: the crosscutting group with the general information about the reference architecture, the application group, showing the dynamic behaviour of the systems that will be built based on the reference architecture, the infrastructure group, describing the hardware and software, etc. At last, there is the domain group, which describes the legislation, standards and regulations, etc.

4. Architectural Evaluation

The reference architecture evaluation proposed by Nakagawa refers to a task where the architectural description is checked by the stakeholders in order to detect defects (Nakagawa et al., 2014).

A third very generic approach focusing on "reference models" is proposed by Fettke and Loos (2004). Their approach for building a reference model also consists of a four-step approach:

1. Problem Definition

Define the objective of the modelling process and for which area it is to be developed.

2. Construction of the reference model

Most commonly, according to Fettke and Loos (2004), an inductive process is used using specific existing enterprise models, or a deductive approach based on theoretical assumptions; another approach would be to develop reference models using existing information systems. Based on the previously defined domain and the goal of the modelling process, the reference model is created using a previously selected modelling language. The result is the description of all models, modelling views, and variants, as well as the relationships between them.

3. Evaluation

The evaluation of the model should not be done after the reference model has been completed. It should be done in parallel with the modelling process and thereby consider the “principles of proper modelling”. Typical criteria for the assessment of the reference model are either economic or technical.

4. Maintenance

The creation of a reference model is not a one-time issue but must be maintained on an ongoing basis, e.g., if modelling errors are discovered, appropriate changes are required, or if new requirements arise.

A reference architecture describes a generic software solution (see definition of reference architecture). The methodology for designing a reference architecture as a reference model for other reference architectures is based on requirements engineering. In order to design and implement an information system or software solution that solves a specific problem, one needs to understand the problem and what needs to be solved (Lamsweerde, van, 2009). Therefore, one needs to find out, understand, formulate, analyse, and agree on what problem needs to be solved (the description of services, constraints, and assumptions), why this problem needs to be solved (the objectives) and who (such as stakeholders) needs to be involved. Following van Lamsweerde (2009), requirements engineering is the "coordinated set of activities to explore, evaluate, document, consolidate, revise and adapt the

goals, capabilities, qualities, constraints and assumptions that the future system should fulfil based on the problems posed by the existing system and the opportunities offered by new technologies" (Lamsweerde, van, 2009).

To develop a reference architecture, it is necessary to analyse the needs, perform a system analysis and derive and specify its requirements. There are different categories of requirements. First, there are functional requirements, which define or provide the functionality or features that the reference architecture should fulfil. Thus, they address the "what" of the three questions that the reference architecture should answer. Another category is the non-functional requirements, which can be separated from the constraints or (depending on the author) can be seen in one category. The quality requirements are additional functional effects that the reference architecture should provide, in the form of quality-related properties. Other non-functional requirements can be distinguished into architecture requirements (e.g., distribution or platform requirements) and development requirements that describe or constrain how the reference architecture (or a particular instance of it) should be developed (e.g., in terms of cost, schedules, variability of features, maintainability, reusability, portability, etc.) A key constraint on the re-fish reference architecture is the conformance requirements -- the description of how the reference architecture (or a specific implementation of an information system) will conform to the environment, such as national, international (supernational) laws, international regulations, social norms, ethical norms, cultural and political constraints, and standards (Lamsweerde, van, 2009; Bejger & Elster, 2021).

In this thesis, we will follow the architecture design approach presented by Bass et al (2022) and align it with the ADM TOGAF methodology. Bass et al. describe the Attribute Driven Design (ADD) approach, which aims to make architecture design systematic, repeatable, and cost-effective. Kazman and McGregor (2012) (s. figure 62) argue that in order to use ADD for the design and development of a reference architecture, the ADD approach (and the definition of the architecture) must be adapted.

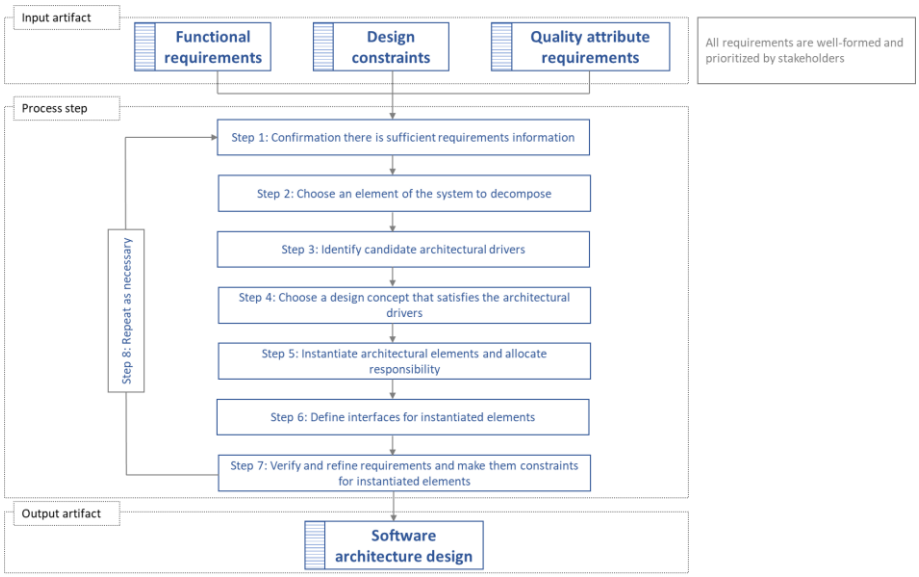

Figure 62: Adjusted ADD approach for RA, Kazman and McGregor (2012)

The architecture design activity consists of design concepts such as reference architectures, externally developed components, tactics, and patterns. This approach turns decisions about architectural drivers into structures. Architectural drivers include and comprise Architecturally Significant Requirements (ASR), functionality, constraints, architectural concerns, and finally, the design purpose. The structures are used to guide the analysis and design process. Before architectural design begins, the scope of the system must be determined.

This is done by defining the context of the architecture or system (Bass, 2021) (s. figure 63).

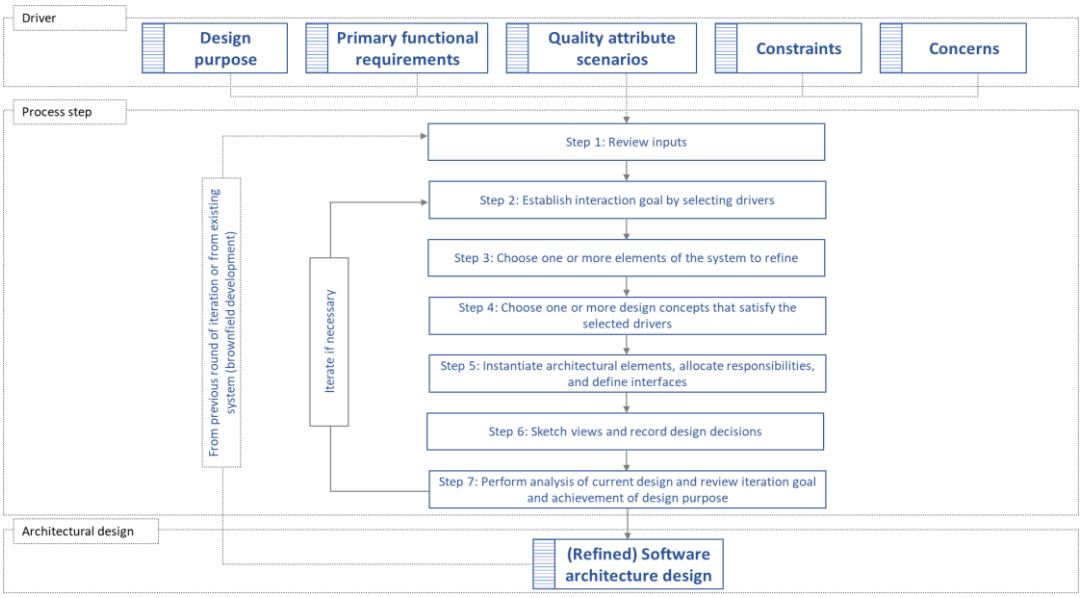

Figure 63: Steps of the Attribute-Driven Design approach of Bass, L. et al. (2021)

- Design Purpose
- Primary functional requirements
- Quality attribute requirements
- Design Constraints
- Concerns

There are 7 steps in the ADD, which need to be iterated, where necessary. Before an architectural design round is started, the architectural drivers need to be available and correct (the iteration, if needed, is sometimes described as being the “step 8”) (Bass, et al., 2021)

Step1: Review inputs and confirm that there is sufficient requirement information.

In this step, the design purpose is reviewed as the goal for the whole process must be clear (especially for one round if it is about incremental design). The primary functionality gathered through the user stories of the main stakeholders must be identified. In this phase, it is necessary to evaluate the requirements of the business conditions that have changed, etc. All the drivers mentioned above will be part of a design backlog. If iterations are done, there must be a design decision which addresses the parts of the architectural drivers comprising parts of the backlog (Bass, et al., 2021).

Step 2: Establish an iteration goal by selecting drivers -- Identify candidate architectural drivers.

Each iteration round of the ADD approach focuses on a specific goal, e.g., fulfils a subset of drivers. That specific goal can be a particular performance goal to be reached, or a specific use case to be fulfilled. In the design of a reference architecture, the goals must be adjusted; accordingly, therefore, a performance goal might be valid for a specific instantiation of the reference architecture, but not for the initial one.

Step 3: Choose One or More Elements of the System to refine or decompose.

To satisfy one or a subset of the architectural drivers, architectural design decisions must be made, which will "manifest themselves in one or more architectural structures". That can be done using an iterative or top-down approach, by refining and with fine-grain elements, or by using a bottom-up approach and gaining a rougher set of elements. Those elements are the ones which are needed to satisfy the specific architectural drivers. Structures are built by modules and components, and refining elements from previous iteration cycles realise them.

Step 4: Choose One or More Design Concepts That Satisfy the Architectural Drivers

In this step, one or more design concepts must be chosen. This is an important step, as there are options to choose from, e.g., tactics, patterns, and reference architectures, for instance, and externally developed components. This is a difficult decision to make, as it must be chosen in relation to the selected iterated goal.

Step 5: Instantiate Architectural Elements, allocate Responsibilities, and Define Interfaces

In this step, the decision has to be made on how the elements can be instantiated according to the design concept, such as if the layer pattern will be one design concept, and the architecture under design will be for an application. Typically, there are three layers to be used: the presentation layer to handle all user interaction with the application, the business layer for the business rules, and the data layer for persistence. The different layers have to be connected; therefore, interfaces are needed to handle the interactivity between the different layers.

Step 6: Sketch Views and Record Design Decisions

The results of the iteration must be preserved as the results of the ADD, e.g., diagrams, as they are essential for the whole process. The different sketches and views are built more formally than the architectural documentation. The views and documentation are the basis for the analysis and review process of the architecture, and how it satisfies the architectural drivers.

Step 7: Perform Analysis of Current Design and Review Iteration Goal and Achievement of Design Purpose

By reaching this step, a partial design should have been created, which is meeting or addressing the goal established for the iteration. This is a proof point that will avoid issues like unhappy stakeholders and later reworking. This can be done best, e.g., by a third person.

The ADD approach is usually designed to build an architecture and not reference architectures in particular; therefore, some modifications might be necessary to develop a reference architecture, due to its more generic and conceptual structure. The modifications necessary in the ADD are that the ADD becomes more “conceptual”, and the elements in ADD will be described as more abstract. Quality attributes (though not precise quality goals) are identified. As well, architecture patterns are described but not concretely instantiated.

At least the ADM (Architecture Development Methodology) within The Open Group Architecture Framework (TOGAF) will be presented and ADD (presented above) and ADM will be aligned. The TOGAF originated as a generic framework and methodology for the development of technical architectures. Since version 8 of the TOGAF framework, it has been oriented towards the creation of enterprise architectures; previously, the main task was to describe technical architectures without a dedicated orientation.

The components of TOGAF are shown in figure 64.

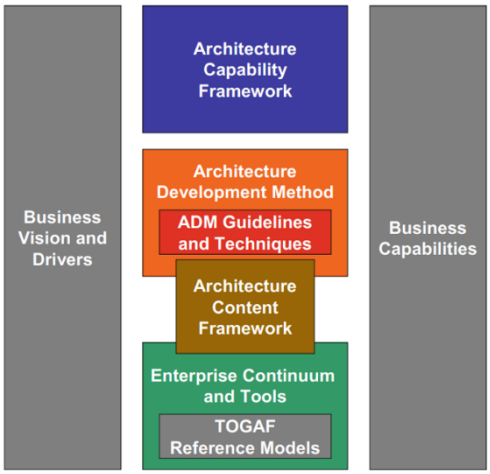

Figure 64: Components of TOGAF

TOGAF consists of an Architecture Capability Framework, which addresses the organisation, processes, skills, roles, and responsibilities required to establish and operate an architecture function within an organisation.

The Architecture Development Method (ADM) is a presentation for architects, a kind of roadmap for the creation of architectures, and a cyclical approach that develops the enterprise architecture step by step. The Architecture Content Framework considers an overall enterprise.

The ADM consists of four closely related architectures: the business architecture, the data architecture, the application architecture, and the technology (IT) architecture.

The Enterprise Continuum shown in the figure consists of various reference models. For example, it includes the Technical Reference Model, The Open Group's Standards Information Base (SIB), and The Building Blocks Information Base (BBIB). The aim is to show how to get from various basic architectures via general system architectures and sector-specific architectures, to a company-specific architecture.

The TOGAF architecture development process is shown in figure 65 with its different phases. It is important to understand that this approach can be run through several iterations and that decisions for every iteration must be made about the

- level of detail
- time horizon and intermediate steps (Lankhorst, 2017)

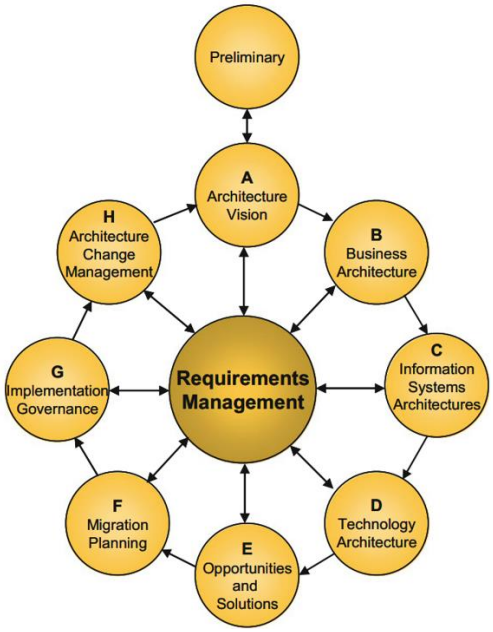

Figure 65: TOGAF ADM

The idea of the following table is to combine the two approaches, ADM and ADD, into one approach to create the reference architecture for Re_fish.

Step	ADD	ADM	Artefact
1	Review inputs and confirm that there is sufficient requirements information	Preliminary, Requirements Management	Motivation View, Use Case View, Stakeholder Map, Requirements, Expert Survey I, Constraints
2	Establish iteration goal by selecting drivers - Identify candidate architectural drivers.	Architecture Vision, Preliminary, Requirements Management	Solution Context, Use Case View, Stakeholder Map, Requirements, Constraints
3	Choose One or More Elements of the System to refine- or decompose	Business Architecture, Information Systems Architecture, Technology Architecture	Solution Context, Use Case View, Stakeholder Map, Requirements, Constraints
4	Choose One or More Design Concepts That Satisfy the Architectural Drivers	Business Architecture, Information Systems Architecture, Technology Architecture	Layered View - Swimline Process View, Solution Architecture View, Component Model 0-2
5	Instantiate Architectural Elements, Allocate Responsibilities, and Define Interfaces	Business Architecture, Information Systems Architecture, Technology Architecture	Layered View - Swimline Process View, Solution Architecture View, Component Model 0-3
6	Sketch Views and Record Design Decisions	Business Architecture, Information Systems Architecture, Technology Architecture	Layered View - Swimline Process View, Solution Architecture View, Component Model 0-4
7	Perform Analysis of Current Design and Review Iteration Goal and Achievement of Design Purpose	Iterations of the ADM will focus on the evolution of the architecture	Expert Survey II
8	Iterate if necessary	Iterations of the ADM will focus on the evolution of the architecture	n/a

Table 15: Mapping ADD and ADM and artefacts

Table 15 shows how the two approaches, ADM and ADD, can be mapped to each other and to the corresponding artefacts. In the following, we will therefore follow the ADM method to develop the reference architecture and its artefacts.

TOGAF and ADD are frameworks. These frameworks must be concretised and developed for the respective use case. This is why, for example, the specific manifestations of the TOGAF framework differ in different companies.

Under TOGAF, there are several reasons why an enterprise architecture needs to be reviewed or developed -- in this case, the development of a reference architecture system for trustworthy AI - Re_fish.

TOGAF calls the key people involved in the creation and use of an architecture or reference architecture stakeholders. These key people want their requirements for the IT (reference) architecture to be implemented and achieved. An architect who undertakes the creation of the architecture must ensure (that is, address) their concerns by:

- “Identifying and refining the requirements of the stakeholders.
- Developing views of the architecture that show how the concerns and requirements are going to be addressed.
- Showing the trade-offs that are going to be made in reconciling the potentially conflicting concerns of different stakeholders”.

Without a structured approach, it is unlikely that all stakeholder requirements will be met.

The preliminary phase of ADM is about the preparation and initial activities to implement the business directives for a new Enterprise Architecture. This includes the definition of the architecture, and the organisation-specific framework and the definition of architectural principles.

The objectives of the Preliminary Phase are mainly related to organisational preparations, when it comes to projects and e.g., the alignment with other frameworks. This has already been implemented in chapter 4.2, with regards to the creation of the reference architecture.

The input for this phase is therefore the TOGAF library and other frameworks.

Non-architectural inputs include the following information: Board strategies and board business plans, the business strategy, the IT strategy, the business principles, the business goals, and the business drivers, if available.

However, the governance and legal frameworks, including architecture governance strategy, appear to be particularly relevant, if they are available.

As described in chapter 3, one of the core requirements for AI models is that they meet the requirements of the "Ethics Guidelines for Trustworthy Artificial Intelligence" and the EU-DSGVO. These are also the framework conditions or constraints in the context of creating the architecture.

In the following chapters, we will work through the architectural design methodology to define which of the tasks and the artefacts are relevant in order to build the RA_fish reference architecture.

For the development of a reference architecture itself, the ADM phases E to H are only partially relevant. They are seen as being relevant to any specific instantiation of the reference architecture.

4.3.2 Phase A: Architecture Vision

This chapter describes the initial phase of the ADM. In this step, the scope of the architecture is defined, the stakeholders are identified, and the architecture vision is created (here and in the following, TOGAF, 2022).

Inputs are reference materials, like reference architectures. They involve a request for architecture work and business principles, business goals, and business drivers.

Architectural inputs are, e.g., constraints on architecture work, the re-use of requirements, and architecture principles, which might also include the business principles.

4.3.2.1 Establish the Architecture Project

In this step, as standard enterprise architecture is also a business capability, the iterations of ADM are a project. In this step, the project management set-up activities are defined and implemented.

- Relevant deliverable for this work: The project definition is more or less given in the description of the architecture vision.

4.3.2.2 Stakeholders, concerns, and business requirements

In this step, the stakeholders are identified, as are their concerns/objectives; key business requirements need to be addressed in the architecture engagement. Stakeholder engagement at this stage is intended to accomplish three objectives:

- To identify candidate vision components and requirements to be tested as the Architecture Vision is developed
- To identify candidate scope boundaries for the engagement to limit the extent of architectural investigation required
- To identify stakeholder concerns, issues, and cultural factors that will shape how the architecture is presented and communicated.

The major deliverable in this step is the stakeholder map for showing the stakeholders how they are involved with the engagement, and the level their involvement. Their concerns will be also included in this map. Importantly, it also gives information about those concerns and provides input for several other important documentations, e.g., the relevant viewpoints for the different stakeholders, which are part of the architecture vision.

Key deliverables of this part are:

- The Stakeholder Map
    - Concerns – Requirements
    - Information about the views/viewpoints

It is also necessary that the scope of the requirements is documented (ideally, in a kind of requirements repository) as it can change over the time and needs to be adapted accordingly in future iterations.

In this work, the stakeholders' concerns are gathered via a literature review.

- Relevant deliverable for this work: Stakeholder Map, with all required concerns resulting functional and qualitative requirements, constraints, and relevant views/ viewpoints.

4.3.2.3 Confirm and Elaborate Business Goals, Drivers, and Constraints

In an architecture project, the business goals and strategic drivers of the organisation must be gathered and documented. As this work is concerned with the development of a reference architecture, there are no concretised business goals for this project. The business-relevant requirements and constraints are gathered and documented in the stakeholder definition.

Evaluate Capabilities and Assess Readiness for Business Transformation are not applicable.

4.3.2.4 Define Scope

In this step, the baseline architecture is described by a scope diagram. This diagram will be the decomposed in the following steps.

- Relevant deliverable for this work: Scope Diagram

4.3.2.5 Confirm and Elaborate Architecture/ Business Principles

In this step of the ADM, the Architecture Principles are reviewed under which the architecture is developed. This is usually part of the preliminary phase (s. above).

- Relevant deliverable for this work: Review or definition of the Architecture Principles.

4.3.2.6 Develop Architecture Vision

In this step, the architecture vision is developed. Therefore, is also provides a high-level view, as information about an overall architecture to be decided upon is given, based on the stakeholder concerns, scope, constraints, and principles. There are different ways to achieve such a high-level overview. One common practice, for instance, is to develop a simple solution concept diagram, which illustrates the major components of the solution

and gives an initial idea on how the solution will result in benefits, if used. Another possibility is the usage of business scenarios (use cases or user stories), which are an appropriate and best practice technique to discover and document business requirements. Along with the scope diagram, it helps to provide an architecture vision that already give a response to the requirements. In phase B, business scenarios will also be used. "This step generates the first, very high-level definitions of the baseline and target environments, from a business, information systems, and technology." (TOGAF, 2022).

- Relevant deliverable for this work: Business scope diagram, user stories/use cases

4.3.2.7 Summary of Phase A

The relevant outputs or deliverables for this work are:

- Architecture principles
- Refined business principles
- Architecture vision
    - Problem description
    - Objective of the work statement
    - Summary views
    - Business scenarios
    - Stakeholder map

4.3.3 Phase B: Business Architecture

The objectives of this phase are to develop the Target Business Architecture. This architecture is used to describe how the corporation needs to operate to achieve its business goals. The Business Architecture is also the answer to the strategic drivers set out in the Architecture Vision. It addresses the stakeholder concerns.

Inputs during this phase are reference material, Architecture Principles, Architecture Vision (with its problem description), the objective of the statement of work, summary of relevant views, business scenarios, refined stakeholder requirements, and the draft version of the Architecture Definition Document. In the following, only those steps which are seen

to be relevant are presented in detail. For a full and detailed description of all steps, the reader may refer to the official TOGAF documentation (TOGAF, 2022).

4.3.3.1 Select Reference Models, Viewpoints, and Tools

This is when there is a selection of the relevant business architecture resources, reference models and/ or patterns. Relevant views/viewpoints area chosen, which are showing that the stakeholder concerns are addressed.

4.3.3.2 Conduct Formal Stakeholder Review

This step is conducted to check the original motivation for the architecture project and also refine the stakeholder requirements. In this work, this step is done via the literature review. In a next iteration or, when the reference architecture will be instantiated, there can be additional tools like surveys, interviews, workshops with the stakeholders etc.

4.3.3.3 Finalise the Business Architecture and update ADD

In this step, the Business Architecture building blocks will be included by re-using as much as possible from the existing architectures being consulted as a reference. The Architecture Definition Document will be updated.

4.3.3.4 Summary of Phase B

For this work, relevant deliverables and therefore the output of phase B are:

- (Target) Baseline Business Architecture – a diagram
- Relevant business functions and business services – in a diagram
- Products the output generated by the business to be offered to customers – Business processes – in a diagram.
- Business roles – reflected in the stakeholders – in a stakeholder map/table
- Business data model – in a diagram
- Views corresponding to the selected viewpoints addressing key stakeholder concerns.

Main deliverable of this phase relevant to this work is the Business Architecture diagram.

ADM also suggests a draft architecture requirements specification, including the business architecture requirements, like a gap analysis. The gap analysis is necessary if there is an existing solution in place and a gap is identified between the current and the target. This is not irrelevant in the context of a reference architecture, but will become relevant in a specific context, by instantiating the RA. Technical requirements and updated business requirements.

4.3.4 Phase C: Information System Architecture

This is the main phase for developing the reference architecture, and it consists of the development of the target Information Systems Architectures building the reference architecture. This description addresses how the architecture will enable the identified requirements of the stakeholders and the enablement of the Business Architecture.

Architectural inputs, which are seen to be relevant for this work, include

- Scope of the organisations impacted
- Constraints on architecture work
- Architecture Definition Document
- Deliverables of previous phases

4.3.4.1 Select Reference Models, Viewpoints, and Tools

In this work, at least four architectures of systems will be analysed, which will be used for building the (first iteration) of the reference architecture, Re_fish, for trustworthy AI systems. The relevant Application Architecture and Data Architecture resources (e.g., from reference models, patterns, etc.) are based on the business drivers, stakeholders, concern, and at least, from Business Architecture. Also, the relevant viewpoints (for example, stakeholders of the data regulatory bodies, users, generators, subjects, auditors, etc.; various time dimensions in real-time, the reporting period, event-driven, etc.; locations; business processes) will be selected. The primary objective is to address the stakeholder requirements.

The recommended process to develop the application architecture, following the ADM, is:

- Understand the list of applications or application components that are required, depending on the requirements.
- Simplify any complicated applications by decomposing them into two or more applications or application components.
- The set of application definitions should be internally consistent, duplicate functionality should be removed as far as possible, and similar applications should be aggregated into one application.
- Identify the logical applications and also the most appropriate physical applications which are required.
- Develop matrices across the architecture by relating applications to business services, business capabilities, data, processes, etc.
- There needs to be a set of Application Architecture views. These views should address how the application will function, become integrated and developed, and which operational concerns or requirements may emerge.

The ADM also recommends building the required matrices showing the association between the related entities. The same goes for the views/viewpoints, which are needed to provide information about how the requirements of the stakeholders are addressed.

Further steps including the development of the baseline application architecture description, the target application architecture description, the gap analysis, the candidate roadmap, the impact analysis, the formal stakeholder review, etc.

4.3.4.2 Summary of Phase C

For this work, relevant deliverables and therefore the output of phase C are:

- Application Architecture – in an application architecture diagram
- Application interoperability requirements – in an application architecture diagram
- Relevant technical requirements – in an application architecture diagram.
- Updated requirements – reflected in the stakeholders – in a stakeholder map/table
- Constraints on the Technology Architecture about to be designed (s. Chapter 4.3.5)

- Views corresponding to the selected viewpoints, addressing key stakeholder concerns.

4.3.5 Phase D: Technology Architecture

In Phase D: Technology Architecture, the objectives are to:

- "Develop the Target Technology Architecture that enables the Architecture Vision, target business, data, and application building blocks to be delivered through technology components and technology services, in a way that addresses the Statement of Architecture Work and stakeholder concerns".
- Define a roadmap between baseline and target architecture, if there is an existing technology architecture.

The input, architectural, and non-architectural are derived from the previous phases.

The steps to develop the Technical Architecture are the same as in the previous two phases, from a methodological point of view. The steps do have a different focus and perspective, which is described in the objectives above.

4.3.5.1 Select Reference Model, Viewpoints, and Tools

In this step of phase D, the set of technology principles have to be reviewed. They are part of the overarching set of architectural principles.

The one for the scope relevant to Technology Architecture resources (e.g., reference models, patterns, etc.) has to be selected based on the business drivers, stakeholders, and their requirements.

The relevant Technology Architecture views and viewpoints must be selected that will enable the architecture to provide how the stakeholder requirements are being addressed by the Technology Architecture.

For the view and the viewpoints of the Technology Architecture, certain steps must be followed:

- Define a taxonomy of technology services and logical technology components (including standards)

- Identify relevant locations where technology is deployed
- Carry out a physical inventory of deployed technology and abstract up to fit into the taxonomy
- Look at applications and business requirements for technology
- Assess whether the technology in place is fit-for-purpose to meet new requirements (i.e., does it meet functional and non-functional requirements)
- Determine the configuration of the selected technology
- Determine the impact of:
    - Sizing and costing
    - Capacity planning
- Installation/governance/migration impacts

For this work, a diagram and matrices will be created. The diagrams present the Technology Architecture information from the different defined and required perspectives (the so-called viewpoints), according to the requirements of the stakeholders.

This activity provides a link between the platform requirements and the hosting requirements, as a single application may need to be physically located in several environments to support local access, development lifecycles, and hosting requirements.

The main illustration will be a stack diagram showing how hardware, an operating system, software infrastructure and packaged applications are combined to run the application architecture. There will be a logical diagram of hardware and software infrastructure, to show the contents of the environment and logical communications between components.

4.3.5.2 Develop Target Technology Architecture Description

The main objective of this step within Phase D is to develop a Technology Architecture description, which enables and supports the Architecture Vision, Target Business Architecture, and Target Information Systems Architecture. The detail of this description is highly dependent on the relevance of the technology elements, which are needed to reach the Target Architecture. A key process in the creation of a broad architectural model of the target system is to use and conceptualise building blocks. These building blocks describe

functionality and how they may be implemented without the detail introduced by configuration or that within the design.

4.3.5.3 Summary of Phase D

The overall deliverables and outputs of Phase D are:

- Reworked and versions of the Architecture Vision phase deliverables, etc.
- Updated or validated technology principles, or new technology principles
- An updated draft Architecture Definition Document (ADD) with baseline Technology Architecture (if appropriate)
- The Technology Architecture
    - Technology Components and their relationships to information systems architecture (application architecture)
- The appropriate technology platforms and their decomposition, showing the combinations of technology, required to implement a particular “stack” of technology
- Environments and location as being a grouping of required technology into computing environments (e.g., development, production) and therefore lifecycle management
- Expected processing load and its distribution across technology components (not applicable to a reference architecture, only to the instantiations)
- Physical (network) communications – as well as hardware and network specifications – not applicable
- Views according to the viewpoints selected, and addressing key stakeholder requirements

For this work, relevant deliverables and therefore the output of phase D are:

- Technology Architecture – in a diagram
- Views corresponding to the selected viewpoints, addressing key stakeholder requirements.

4.3.6 Phases E to H: Implementation of a concrete Reference Architecture

Phases E to H are not covered in detail in this paper, as they are more relevant when it comes to instantiating the reference architecture and implementing it for a specific use case. In phase E of ADM, "opportunities and solutions", the (software) architecture options are evaluated in terms of whether they can fulfil the requirements. These may relate, for example, to which language or languages should be used in the creation of the core components, the modules. These may include Python, Julia, C++, etc., which databases to use for the knowledge base, e.g., Neo4j, which solution to use for workflow management of the ML applications, etc. In this phase, feasibility studies are carried out, e.g., in the context of a PoC, prototyping, with the aim of developing a detailed software architecture specification. In phase F, "migration planning", the transition from the as-is, if it exists, to the target architecture takes place. In addition, the relevant further software development projects must be identified and prioritised in order to organise resources and dependencies. The creation of a roadmap for the implementation and migration of the software architecture concludes this phase. In phase G, "Governance Implementation", governance mechanisms are established to ensure compliance with the standards and guidelines for the software architecture. This involves monitoring and controlling the software development process, checking compliance with the architecture and, iteratively if necessary, making and implementing any necessary adjustments. Phase H "Establishment of architecture change management" completes the ADM cycle. It involves introducing and implementing processes to manage and control changes to the software architecture, as well as conducting impact analyses for proposed changes and evaluating them in terms of their compliance with the software architecture vision, i.e., ultimately evaluating the architecture. Moreover, the integrity and consistency of the software architecture over time should be ensured.

4.3.7 Summary of the Methodology to Develop Reference Architectures

The previous chapters have outlined the methodology for describing the development of an architecture. In addition, a mixture of the ADM and the ADD approach was preferred, which served the purpose of taking into account the more software architecture heavy ADD in the ADM approach. It should be noted that the two approaches do not differ greatly. It is worth noting, however, that the ADD places greater emphasis on iterating individual development phases and thus deepening the scope under consideration and detailing the architecture. - The individual phases A- D of the ADM do not differ significantly, only in

the degree of the viewpoint. A target architecture - here in the sense of the reference architecture - is made up of several components. The artefacts resulting from the implementation of phases A-D are a basis.

4.4 Summary

The task of Chapter 4 was to provide the methodology for the design and development of a reference architecture. for this purpose, the theoretical basis for reference modelling, a reference architecture being a reference model for specific architectures, was laid. Then different approaches to carry out such modelling were examined and a combination of the ADM and the ADD was proposed. It should be noted that the ADM is a framework that can be implemented in different ways for specific companies. The approach was then described in detail so that it can now serve as a basis in Chapter 5, together with the preparatory work from the other chapters, to create the Re_fish reference architecture.

Finding 15: In Chapter 4, the theoretical possibilities for developing a reference architecture were examined and discussed. For Re_fish, the methodology was based on the TOGAF ADM and the ADD methodology. The whole process of designing and developing a reference architecture was described.

"Josie began to lose her strength eleven days after our return from the city. At first this phase seemed no worse than the ones she'd gone through before, but then came new signs, such as strange breathing, and her semi-waking in the morning, eyes open but empty. If during these spells I spoke to her, she wouldn't respond, and the Mother took to coming up to the bedroom early each morning. And if Josie was in her semi-waking condition, the Mother would stand over the bed, repeating under her breath, 'Josie, Josie, Josie,' as though this were part of a song she was memorizing."(Ishiguro, Kazuo (2021). Klara and the Sun. Chapter 5)

5. Development of a Reference Architecture for Explainable AI in Corporate Planning

5.1 Introduction

As the theoretical foundation was laid in Chapters 1-4, the Re_fish reference architecture will be built based on those findings. Therefore, in Chapter 5 of "Design and Development of a Reference Architecture for Explainable AI in Corporate Planning, " the reference architecture will be developed following the ADM methodology aligned with the ADD.

5.2 Development of the Re_fish Reference Architecture

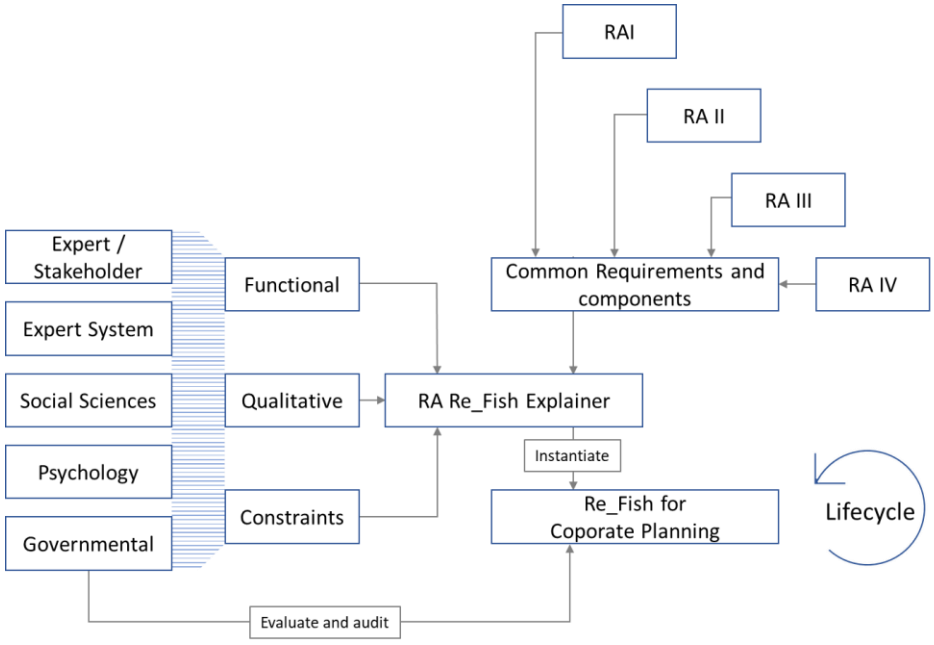

Figure 66: Structure of the design process

Figure 66 shows how the design process of the reference architecture is structured. The left part shows that the functional and qualitative requirements collected in the previous chapters, based on research results from different scientific disciplines, are incorporated into the reference architecture. These requirements are represented in the stakeholder map and are assigned to them. The list of all requirements and constraints is summarised in 5.2.1. The fulfilment of these requirements forms a basis for the evaluation of the architecture. In addition to the requirements and constraints, architectures that have already been created (RA I - RA IV) are also included; the reference architecture created in this way can then serve as a basis for further, specific architectures. An important requirement arises from the possibility of auditing the Explainer/AI system, due to the ever-increasing demands for the explainability of the recommendations and decisions of (automatic) AI.[33]
In particular, the abstraction of existing systems that are already in use, in the sense of an inductive approach, is an essential methodology for developing Re_fish. There are already a number of approaches to knowledge-based systems. For example, Chari et al. describe expert systems that use the EES framework, e.g., MYCIN and NEOMY-CIN. There is also an explainable description logic - CLASSIC - and the development of the EES framework (presented later in this text). Cognitive assistants are largely driven by the DARPA Personal Assistant that Learns (PAL) programme, which was used to build the Cognitive assistant that Learns and Organises (CALO) system. The CALO system uses the Integrated Cognitive Explanation Environment (ICEE) - Intelligent Tutors. The scenario planning system of the DFKI (AISOP). As non-symbolic AI models have recently received special attention due to their enormous power in some areas, many methdodes and approaches have emerged in this domain, as already mentioned in Chapter 3.3.1.

In an architecture development process, the results and deliverables are documented in an artefact, the Architecture Definition Document (Not to be confused with Attributive Driven Design (also ADD for short), hence, the artefact is abbreviated hereafter as ADDC.). The ADDC is the deliverable container for the core architectural artefacts created during the

[33] With the publication of ChatGPT, especially ChatGPT-4, by OpenAI, the discussion about artificial intelligence and the demand for control was accelerated once again. In May, the EDSA ("European Data Protection Committee") set up a task force around ChatGPT - as a reaction to the Italian ban on chatbots (https://edpb.europa.eu/news/news/2023/edpb-resolves-dispute-transfers-meta-and-creates-task-force-chat-gpt_en). Accessed 18.06.2023

whole project. The Architecture Definition Document spans all architecture domains (business, data, application, and technology) and examines all relevant states of the architecture -- baseline, interim state(s), and target.

The architecture definition document is a companion the architecture requirements specification and has a complementary objective: it provides a qualitative view of the solution and is intended to convey the architects' intent. The architecture requirements specification provides a quantitative view of the solution and gives measurable criteria to be met in the implementation of the architecture. In the following, however, only some parts of the architecture design document will play a role. The further development of the artefact takes place in further iterations to refine the reference architecture or in instantiation, in which the reference architecture is used for concrete use cases (TOGAF, 2022).

5.2.1 Preliminary, Purpose and Scope

This step is to be linked to steps 1 and 2 of the ADD[34], and steps 1 and 2 of the ADM method; the scope of the architecture needs to be defined and the stakeholders should be identified, along with their requirements. The architecture vision also needs to be defined.

Inputs:

The inputs for this phase are the architectural drivers, the stakeholders, and the goal[35] (the iteration goal(s)) of the design cycle. Other inputs come from the ADM or are existing reference material and reference architectures. From the enterprise perspective, the inputs are business principles, goals, and business drivers. Since the reference architecture is at an aggregate level, the inputs from the architecture perspective are all constraints on the architecture work. It is essential to consider and check in advance whether there is a possibility to reuse requirements and architectural principles (probably including business principles).

Outputs:

a) Architectural vision

[34] Here “Attributive Driven Design” – in short ADD

[35] An architecture is developed by multiple iterations, therefore for every iteration the goals of this cycle must be defined.

b) Solution context: a high-level architectural diagram

c) Stakeholder Map with requirements

d) Constraints

a) Architecture Vision

Problem Background:

When decisions and actions made by an AI model in corporate planning scenarios and decision-making are not explainable to stakeholders, they are not trusted. As these models need to be more transparent, interpretable, or explainable, they are not used to their full potential (the difference between interpretability/explainability and explanation depends on the situation in which the model is used). This dissertation proposes that most managers and decision-makers in business need more mathematical and statistical knowledge to understand decisions or actions made by subsymbolic black-box machine learning and profound learning models. A sustained lack of stakeholder trust may slow down or even prevent the adoption of AI approaches and models within a corporate planning - business context. Corporate planning is one of the core capabilities of management or leadership, and goal-oriented, forward-looking thinking is not limited to one company. Planning is a core element of business and is central to all business disciplines. It entails the anticipation of future operational events, thus planning transactions by thinking about the future and doing so while having a goal-oriented approach. Such goals must be stated clearly among the different areas and subareas of the company, aside from decision-making. Therefore, planning is a decision problem, which may be examined from different perspectives, e.g., business administration follows a rationality paradigm, with a model of the rational thinking “homo economicus”; the cognitive psychologists prioritise the processes in the mind of the decision maker; game theorists are interested in mathematical decision behaviour; the behavioural economists are interested in the changes in decision-making behaviour in particular contexts, etc. Of note here is that the quality of decision-making is significantly improved through the usage of AI models, as humans tend to bias decision-making with emotions and irrational behaviours. Humans also lack information about the situation the decision must be made within (bounded rationality) (Gigerenzer & Selten, 2002; Russel & Norvig, 2022). Humans tend to base their decision-making on subjective, past experiences

- even when the context of the situation does not fit. Recent studies have found[36] a machine-hybrid approach, which could beat the best chess computers within a game, for instance, and reach better results than AI or a human, alone (Augmented AI, s. e.g., De Cremer & Kasparov, 2022; Frankfurter Allgemeine Zeitung, 2022). Therefore, the proposition is a hybrid approach of human and AI, which leads to better results in planning. The use of AI is particularly helpful in the two sub-disciplines of scenario planning and integrated business planning (sales & operations planning). In scenario planning, for example, there is a large, comprehensive set of alternatives from which only the scenarios that are relevant for the company can be selected. In the area of integrated business planning, the combination of the scope of planning (strategic, tactical and operational), product levels, locations, suppliers and customers, and any external factors that need consideration at different levels can lead to such a high level of complexity that AI models can be used successfully. However, their suggestions and decisions must be explained to the user so that they can be trusted, and its suggestions implemented accordingly. The focus is only on the two parts – scenario planning and forecast (demand, supply, distribution, procurement) and not to automate the whole corporate planning process (therefore, it is more to be to augment the planning process resp. for the planner)

Change Drivers and Opportunities:

The purpose of the reference architecture is defined by the research goal of this work (s. Chapter 1.2 The Research Goal and Research Question):

The main goal of this work is to develop a reference architecture as a reference model which can be used for design development, as well as implementation and runtime of a trustworthy and reliable XAI system. The designed reference architecture is called "Re_fish" (in tribute to Marian Rejewski, the leading Polish scientist solving the Enigma code and the Babelfish – "a fictional universal decoder for any form of language in the universe" (Adams, 2010). The empirical relevance of the reference architecture will be developed with scientific rigour, within a process industry corporate planning context. The reference architecture for trustworthy AI systems should consider the requirements of all relevant stakeholders (see table 7-10, 13 and 14) and ensure explainability by design, as

well as throughout the entire life cycle. The explanation component should be able to account for different models of non-symbolic and symbolic AI.

Business principles and goals derive from the requirements of the process industry mentioned in Chapter 2:

Both process industries, chemical and pharmacy, play significant roles in the global economy and involve complex, interconnected supply chains and processes, in which raw materials are transformed into intermediate and finished products through a series of chemical reactions and physical operations. The XAI system must be able to provide the relevant explanation of its decisions/recommendations, especially causality. Typical (domain) knowledge of the XAI system should include, for example, the following areas:

- Knowledge/Information about different steps - the sequence of the production, e.g., scheduling of the batches, which is relevant within forecasting supply, demand etc.
- Allocation of the right resources in the right volume/quality/time/quality
- Domain knowledge, e.g., clinical trials, research/development activities, regulatory approvals, particularly important for the introduction of new products within the framework of the forecast.
- Knowledge about scenarios, forecasting, forecasting accuracy
- Predictive maintenance - as one of the findings in Chapter 2 was that the process industry is highly asset-intensive industry
- Knowledge about the planning process and the decision variables, the strategic and tactical inputs (s. the stakeholder map)

b) Solution Context- High Level Architecture Diagram

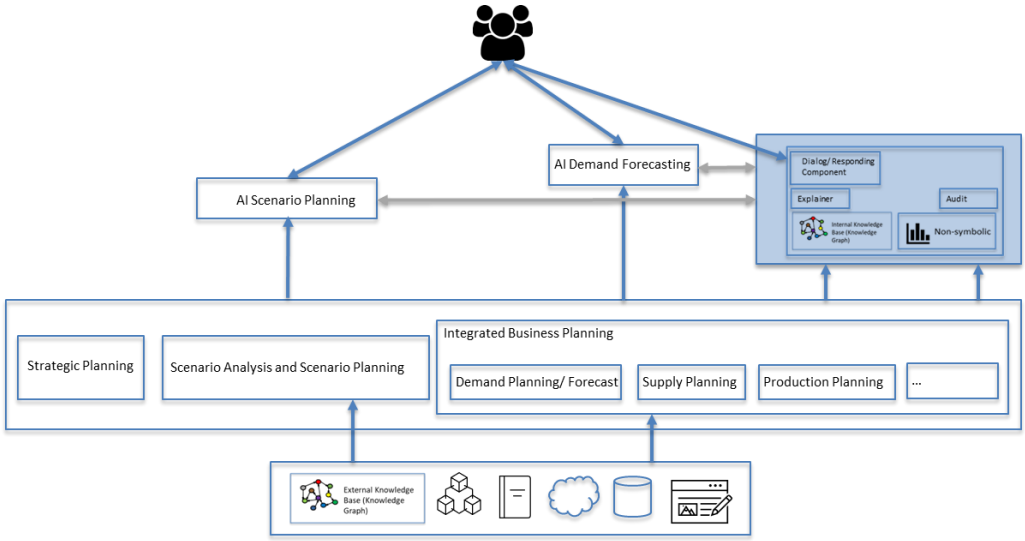

Figure 67: High level conceptual diagram

From the illustration in charts 67 and 68, which is a high-level concept diagram, it can be seen that the Re_fish reference architecture described here (highlighted in blue) is used to explain the AI models used in the context of business planning. This work focuses on the areas of scenario planning and integrated business planning that have been identified as particularly important (s. Chapter 2).

Figure 68: Re_fish high level conceptual diagram

a) Stakeholder Map with requirements

The stakeholder map is divided into two parts. Two mapping tables were created based on the research. The first table shows the entire model of strategic and tactical integrated business planning with its decision variables. The table contains the key input values and the parameters that are to be regarded as external to the model and shows which questions and types of explanations arise in the planning process. In the present case, these are essentially question types 2 and 3 of Pearl's causal hierarchy. Therefore, it must be possible to answer these types of questions within the framework of the explainer.

The tables in Chapter 2.3.4 (tables 1-4, Stakeholder Map A – Model and Decision Variables, parts I to IV) show the parameters relevant within strategic-tactical corporate planning. The parameter ID are SI = Strategic Input, EP = External Parameter, DE = Decision Variable, IP = Input Parameter. Stakeholder Group = mapping in Table Stakeholder B, Stakeholder- Mapping in due Table Stakeholder B, Domain = considered domain, e.g., Procurement Planning, Sales Planning etc. Type = formulation of the ID type. Decision = decision, Input for plan = for which planning the parameter is valid as input, Input from plan = from which planning the parameter is transferred as input. Deliverable = What is the respective deliverable, e.g., demand planning etc.? Industry remarks = remarks if the parameter is particularly relevant to the process industry (chemicals, life sciences). Impacted Stakeholder = Stakeholders who are affected by the decision. Description = general description of the parameter. Decision/specification = specific presentation of the decision, and finally Level of Causal hierarchy = level of the causality hierarchy according to Judea Pearl, Explanation Type according to Judea Pearl (Pearl (2009); Pearl (2018)).

b) Constraints

Table 10 Stakeholder Map Part B lists the constraints per stakeholder and allows them to be checked against the reference architecture. In addition to the constraints identified per stakeholder group, there are other requirements or constraints on an AI model. In the context of lifecycle management, it must be ensured that the model is not subject to bias (see Chapter 5.2.8).

Table 16: Stakeholder Map B – Constraints

ID	Requirements/ Constraint	Description	Architectural Principle/ Constraint	Stakeholder	Stakeholder Mapping	Solution	References
RC1	Acceptance	Improve acceptance of systems	The system must follow the GDPR regulations and ethical guidelines for trustworthy AI or any rules laid down by the regulator.	Regulator, Deployer	Government, Auditor, Management Board	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI
RC2	Accountability	Provide appropriate means to determine who is accountable	The deployer is responsible for the outcome and consequences of the use of the AI if it can be demonstrated that the model is used in accordance with the deployers's specifications. The deployer confirms and is responsible for compliance with the GDPR and ethical principles for trustworthy AI or any rules laid down by the regulator	Regulator, Deployer	Government, Auditor, Management Board	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
RC3	Fairness	Assess and increase a system's (actual) fairness Affected,	The deployer is responsible for the outcome and consequences of the use of the AI if it can be demonstrated that the model is used in accordance with the deployers specifications. The deployer confirms and is responsible for compliance with the GDPR and ethical principles (fairness) for trustworthy AI or any the rules laid down by the regulator.	Regulator, Deployer	Government, Auditor, Management Board	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
RC4	Informed Consent	Enable humans to give their informed consent concerning a system's decisions	The regulator and the deployer are responsible to enable any humans affected by the system's decision to give informed feedback on the decision made.	Affected, Regulator	Government, Society, Customer, Auditor, Management Board	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
RC5	Morality/Ethics	Assess and increase a system's compliance with moral and ethical standards	The regulator and the operating company are responsible for and must assess the AI system for system compliance with GDP regulations and guidelines for trustworthy AI.	Affected, Regulator	Government, Society, Customer, Auditor, Management Board	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
RC6	Responsibility	Provide appropriate means to let humans remain responsible or to increase perceived responsibility	In applications or situations that directly or indirectly endanger life, it must be ensured that the human being remains ultimately responsible or is able to influence the decision (human in the loop).	Regulator	Government, Auditor, Management Board	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
RC7	Transparency	Have transparent systems	If possible, transparent systems and models should be used and preferred to models that are not transparent. The transparency of the systems used and their modes of operation must be guaranteed. For example, through appropriate documentation or tests.	Regulator	Government, Auditor, Management Board	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
RC8	Trustworthiness	Assess and increase the system's trustworthiness	The system must follow the GDPR regulations and ethical guidelines for trustworthy AI or any rules laid down by the regulator.	Regulator	Government, Auditor, Management Board	Regulations, audits, quality check, following the standards	GDPR, ethics guidelines for trustworthy AI, legal frameworks
RC9	Legal Compliance	Assess and increase the legal compliance of a system	The system must follow the GDPR regulations and ethical guidelines for trustworthy AI or any rules laid down by the regulator.	Deployer	Management Board	Audits, quality checks, following the standards	Design and development guidelines, architectural principles, systems tests
RC10	Safety	Assess and increase a system's safety	The deployer is responsible for the security of the AI system in the designated application area. The user is responsible for using the AI system in the intended area of application.	Deployer, User	Management Board, Demand Planner, Supply Planner etc.	Audits, quality checks, following the standards	Design and development guidelines, architectural principles, systems tests
RC11	Trust	Calibrate appropriate trust in the system	The system must follow the GDPR regulations and ethical guidelines for trustworthy AI or any rules laid down by the regulator.	User, Deployer	Management Board, Demand Planner, Supply Planner etc.	Audits, quality checks, following the standards, system tests	Design and development guidelines, architectural principles, systems tests
RC12	Accuracy	Assess and increase a system's predictive accuracy	The accuracy of the system is checked during the development of the system and during operation.	Developer	AI Developer	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC13	Effectiveness	Assess and increase a system's effectiveness; work effectively with a system	The effectiveness of the system is checked during the development of the system and during operation.	Developer, User	AI Developer, Business User, Demand Planner, Supply Planner etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC14	Efficiency	Assess and increase a system's efficiency; work efficiently with a system	The efficiency of the system is checked during the development of the system and during operation.	Developer, User	AI Developer, Business User, Demand Planner, Supply Planner etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC15	Robustness	Assess and increase a system's robustness (e.g., against adversarial manipulation)	The robustness of the system is checked during the development of the system and during operation.	Developer	AI Developer	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC16	Performance	Assess and increase the performance of a system	The performance of the system is checked during the development of the system and during operation.	Developer	AI Developer	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC17	Verification	Be able to evaluate whether the system does what it is supposed to do	The verification of the system is checked during the development of the system and during operation.	Developer	AI Developer	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC18	Transferability	Make a system's learned model transferable to other contexts	The transferability of the system is checked during the development of the system and during operation.	Developer	AI Developer	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC19	Debugability	Identify and fix errors and bugs	The developer is responsible for developing the system in such a way that it is maintainable, and therefore the maintainability of the system must be documented and made transparent.	Developer	AI Developer	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC20	Autonomy	Enable humans to retain their autonomy when interacting with a system	The developer and the deployer, as well as the user of the AI system, are responsible for ensuring that the user retains autonomy when using the system in its intended application	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC21	Confidence	Make humans confident when using a system	The developer and the deployer, as well as the user of the AI system, are responsible for ensuring that the user retains autonomy when using the system in its intended application.	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, Customer, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC22	Controllability	Retain (complete) human control concerning a system	The developer and the deployer, as well as the user of the AI system, are responsible for ensuring that the user retains control if necessary, when using the system in its intended application.	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC23	Education Learn how to use a system and system's peculiarities	Learn how to use a system and system's peculiarities	The developer and the deployer, as well as the user of the AI system, are responsible for ensuring that system is showing it's peculiarities	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC24	Privacy	Assess and increase a system's privacy practices	The developer and the deployer as well as the user of the AI system are responsible for ensuring that the system complies with the system privacy and continuously improves it.	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC25	Satisfaction	Have satisfying systems	The developer and the provider as well as the user of the AI system are responsible for ensuring that the use of the system meets the requirements.	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC26	Science Gain	Scientific insights from the system	The developer and the provider as well as the user of the AI system are responsible for ensuring that the use of the system is satisfying the requirements	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC27	Usability	Have usable systems	Both the developer and the provider of the AI system are responsible for ensuring the use of the system by the user.	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC28	Usefulness	Have useful systems	Both the developer and the provider of the AI system are responsible for ensuring the usefulness of the system when used by the user in the intended area.	User, developer, deployer	AI Developer, Management Board, Business User, Demand Planner, Supply Planner, etc.	Design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests
RC29	Security	Assess and increase a system's security	The developer and the provider as well as the user of the AI system are responsible for ensuring that the use of the system meets the system safety requirements.	Developer, Deployer	AI Developer, Management Board	Regulations, design and development guidelines, architectural principles, systems tests	Design and development guidelines, architectural principles, systems tests

In Table 16, stakeholder map B – constraints, the requirements (or constraints, s. ID, "RC" = requirement or constraint) for the respective stakeholders are shown, regarding Artificial Intelligence. The table is structured in such a way that the first column contains the requirements/constraints found by Langer et al. (2021). The second column contains the description of these requirements, and the third column contains the derived architecture principles or constraints. These are the framework conditions and the unconditional requirements for the creation of a trustworthy AI; some of the requirements/constraints lead to the same principles/constraints. The stakeholders are listed in the stakeholder column. The national and supranational regulators, e.g., the state governments or the EU, correspond to the regulator. The deployer of the model is usually equivalent to the owner of the company that uses the AI system, e.g., in the context of corporate planning, or offers/provides its service. In a solution, the proposal is listed as to how compliance with the requirements and constraints is to be ensured during development, testing, and application in operation. The references column contains references to regulations and specifications or parts of the system or development documentation.

As already described in Chapter 3.5, the demands on AI have increased, especially from society and public institutions. In the recent past, this has been due not least to the so-called foundation models, such as ChatGPT and in particular the ChatGPT4 algorithm. These models are extremely powerful in that they learn a large amount of data and make it available in the context of chat queries, for example. As a result, the call for regulation and restriction of AI has grown strongly. Already on the 8th of April 2019, the High-Level Expert Group on AI of the European Union presented their so-called "Ethics Guidelines for Trustworthy Artificial Intelligence". These guidelines were a follow-up of the publication of the first draft guidelines of December 2018. The Group received more than 500 comments through open consultation and considered them for the 2019 guidelines. (High-Level Expert Group on Artificial Intelligence, 2019) (s. Chapter 3.5)

In order to meet the requirements of the European Union's Expert Group on AI, Bejger and Elster (Bejger & Elster, 2020)) see two essential conditions that can be seen as constraints on a reference architecture for explainable AI. These are, firstly, explainability by design, whose requirement is already listed in the abovementioned requirements and, secondly, the requirement for auditability.

Explainability should be ensured throughout the entire life cycle. To this end, Bejger and Elster (Bejger & Elster, 2020) call for existing life cycle models for AI and machine learning to be adapted so that no bias or the like can occur from the beginning to the end of the use of an AI model. Suresh and Guttag identify seven sources from which a bias can arise for a model, and which must be avoided accordingly.

These sources and how to avoid them are described in Chapter 5.2.8.

They are

1. Historical bias
2. Representation bias
3. Measurement bias
4. Learning bias
5. Evaluation bias
6. Aggregation bias
7. Deployment bias

Although qualitative requirements are important for both AI and the life cycle, functional requirements that stem from the system's usage requirements are equally crucial. In the following chapters, the requirements will be summarised.

5.2.2 Architectures of Knowledge Enabled AI Systems

In this chapter, common components for the Re_fish reference architecture are identified based on four selected system architectures (s. Chapter 5.2 figure 66 - the architectures designated as “RA”). The selection of these architectures was made in the context of a literature review and based on various articles dealing with research and the status of research into hybrid AI systems and their explanatory components.

In the research plan, one step to build the reference architecture Re_fish is to analyse existing systems of explainable AI. The systems to be investigated are listed in table 11.

The selection of these four systems was done on basis of the literature review done in Chapter 2 and 3.

System name	Domain of application	Domain of application
AISOP	Utilities, scenario planning	2022
SPA	Business	2018
CALO	Business	2004
EES	Program Advisor	1991

Table 17: Investigated (X)AI systems and frameworks and their architectures

The methodology to investigate and categorise the systems was laid out in Chapter 3.

Chari, S, et al. (2020), defined in their research the following categories to describe the systems investigated:

- Modularity
- Interpretability
- Support of Provenance
- Adapt to User's need
- Include Explanation Facilities
- Include/Access a knowledge store
- Support compliance and obligation checks
- Domain usage

The criteria mentioned above and in Chapter 3 will be used to categorise the system and for the step of generalising architectures into a reference architecture. Thus, the specific architectures investigated will become one specific instance of the reference architecture.

AISOP

The AISOP (AI-based scenario planning to predict crisis situations) model by Janzen et al. (Janzen et al., 2022) is used for scenario planning predicting energy crisis situations. AI-SOP uses well-defined scenario patterns, in order to capture entities in the crisis situations.

As already shown in Chapter 2.2, the production process in the process industry is highly dependent on electricity. More so than in discrete production, fluctuations or complete failures in the power supply can severely disrupt the production process (here and in the following see Janzen et al. (2022). While in discrete manufacturing, for example, the entire

production line has to be restarted and synchronised in the event of power failures, which also causes enormous costs, in process manufacturing power failures can cause serious damage - think, for example, of glass production and here in particular of melting tanks or the zinc baths in the galvanisation of parts - or also of the melting baths/ crucibles in casting production. If power failures occur here, the entire production can come to a standstill for several days or weeks, with a considerable loss of material and enormous costs for restarting production. Janzen et al. (Janzen et al., 2022) developed AISOP to assess the risk of such scenarios occurring by looking at various events found in current data and comparing them to historical crisis scenarios to improve the resilience of a process industry company's supply chain. In this context, events such as the war in Ukraine, the Covid 19 pandemic, etc. in particular have shown how weak supply chains can be - entire supply chains have collapsed, leading to a halt in production in some companies in the process industry.

There are several recommendations for improving supply chain resilience. One is planning (strategic planning and scenario planning) to monitor ecosystems and anticipate supply chain challenges before they occur. Most companies have supply chain management experts who monitor specific KPIs or assess political and social situations and their impact on the supply chain. However, power outages (as mentioned above) are of great concern in the process industry, especially due to governments controlled by environmental NGOs that are increasingly restructuring energy production towards sustainable energy generation, such as wind power, while at the same time increasing the use of coal-fired power plants (such as in Germany), making even locations such as Germany at risk. Larger companies are therefore already using their own power supply to mitigate the risk of power outages or to stabilise a potentially unstable supply of alternative energy (solar and wind power). The challenge, however, is to anticipate possible events before they can interrupt or affect the power supply and thus negatively impact production. The goal of AISOP is to predict such crises using scenario planning. AISOP does this by mapping data streams to scenario patterns for determining historical crisis scenarios and predicting future crisis scenarios using inductive knowledge and machine learning. The scenario patterns are operationalised in JSON-LD, resulting in a knowledge graph database of crisis scenarios. A unique feature of the model is that it uses semantically enriched scenario patterns to explain predictive analytics to the decision maker. The model has been tested in the process industry. Based on frameworks such as the Resilience Analysis Grid (RAG) or the Functional Resonance Accident Model (FRAM), AISOP works with semantically enriched scenario patterns used to describe the conceptual structure of a crisis by context, actors, resources,

impact, reason, source, action, and history. (Janzen et al. 2022) AISOP also uses data streams mapped to the scenario patterns to derive historical crisis scenarios, which then lead to an intra-organisational crisis scenario knowledge base over time. The model thus has a learning component so that crisis scenarios can be generated from the historical data. These historical crisis scenarios can then be used by an anticipatory component that uses predictive analytics to create a model to predict possible crisis situations. The monitoring component is used on current data to monitor the company's environment and detect a potential crisis. The missing values or slots in the model template are then "filled in" and an appropriate alert is triggered, allowing the user to make decisions to protect the organisation from the threat of the crisis. The response component then provides this rapid and effective response by using the 'knowledge' of the specific scenario pattern and semantic extensions to explain the recommended preventive actions. The architecture of AISOP is shown in figure 69.

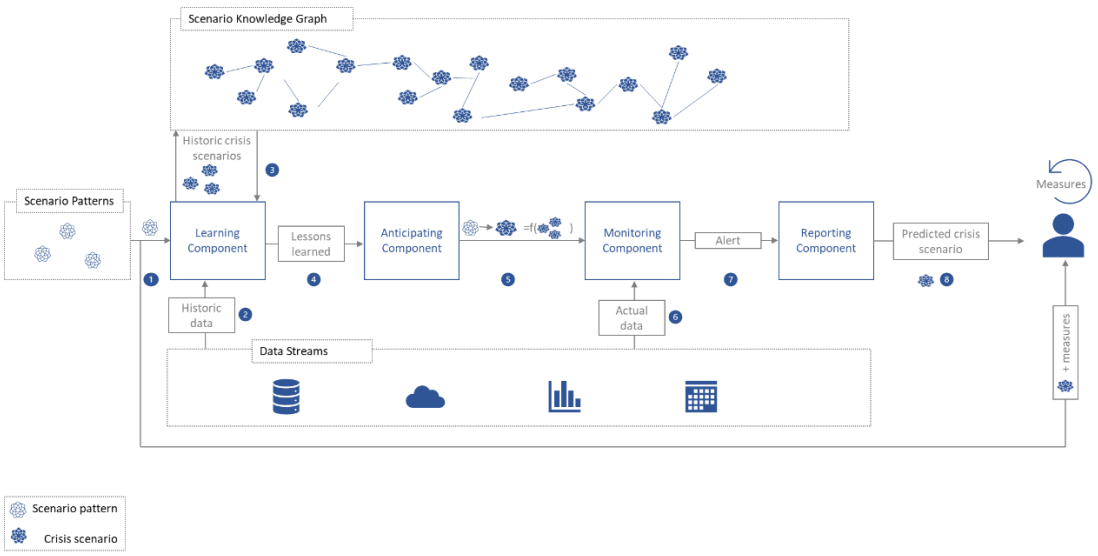

Figure 69: Architecture of AISOP (Janzen et al., 2022)

Entity	Description	Attributes	Example
Identifier	Identifier of scenario	Title, ID, Timestamp	Title: 'Outage' ID: 'Outage_987' Timestamp: '2020-09-13T22:23:05+00:00'
Context	Background information and details	ScenarioDescription, Data, InfluentialFactors	ScenarioDescription:'Outage based on shutdown windturbines' Data: 'wdsp,mxpsd,gust4.0,7.0,26.66' InfluentialFactors: ' Autumn Season'
Source	Origin and reliability of the scenario	Organization	Organization: 'German Federal Network Agency' 'NCEI'
ScenarioLocation	Location of occurrence by the scenario	City, Address, Region, Country,	Country: 'Germany' City: 'Munich' Region: 'Bavaria'
ImpactLocation	Location influenced by the scenario	City, Address, Region, Country	Country: 'Germany' City: '-' Region: 'Bavaria'
Reason	Conditions leading to and explaining the crisis	Precondition Probability	Precondition: 'Wind speed' Probability: '0.78'
Effect	Impact of a scenario and resulting conditions	Postcondition, Complexity	Postcondition: 'Machine downtime' Complexity: 'Low'
Actor	People, groups, departments, taking action	ActorRole, Skillset	ActorRole: 'Worker' Skillset: 'Maintenance work'
Measure	Actions taken to resolve the scenario	Actionstep, Category	Actionstep: 'Plan downtown' Category: 'Precautionary'
Resource	Involved aids and tools	Equipment	Category: 'None'
History	Related historical scenarios	Identifier.ID	Category: 'Outage_913'

Figure 70: AISOP KG entities (Janzen et al., 2022)

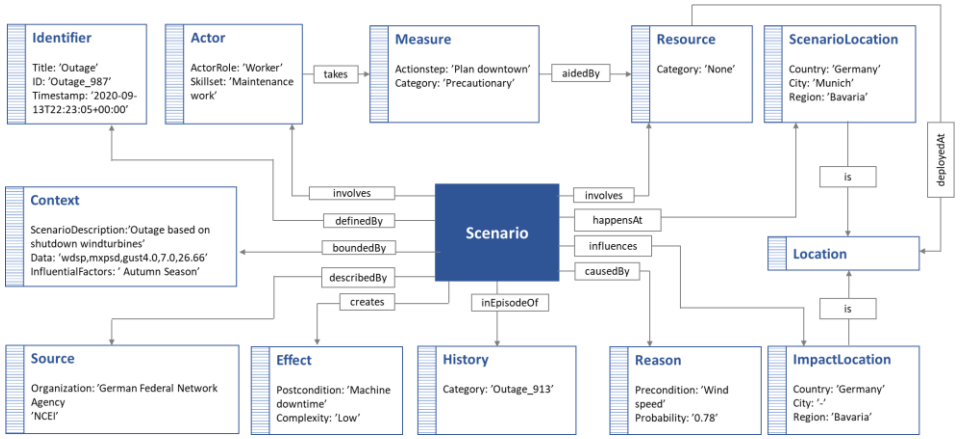

Figure 71: AISOP Knowledge Graph (Janzen et al., 2022)

In AISOP the scenario patterns are building the core of the model and consist of the identifier (see figure 70 and 71): Title, unique ID, "ID", and a "Timestamp". The context entity includes background information, such as “Scenario Description”, the "Data" the scenario is based upon, and "InfluentialFactors". The Source entity provides data "Organization" about the origin of the data. The ScenarioLocation and Location entity describe the location of the scenario using "City", "Address", "Region", "Country". Reason and Effect entities, with "Precondition", "Probability" and "Postcondition", "Complexity" (impact of the effect) provide information about the reason and the effect of the scenario. The Measure and Actor entities provide information about the actor in "ActorRole", the skill-set needed

("Skillset") and with "ActionStep" -- precautionary or sudden actions to mitigate or prevent the crisis. The entity "Resource" includes with "Equipment" the needed equipment for the action. As scenarios evolve over time the entity "History" provides information about historical predecessors by referencing via their unique ID. The core concept of the model is the knowledge base in which the scenario patterns are stored, which can then be used as an explanation in the prediction of crises, in the sense of outages. In a first step, the empty scenario patterns are filled in by the learning component and additionally adjusted and supplemented with further information by experts. Some attributes, such as "Effect", "Reason", and "Location" are filled in using NLP tools and do not need to be adjusted. The scenario patterns thus prepared are transferred to a corresponding instance in the KG (knowledge graph) using JSON-LD. This KG is then used by the anticipating component, which performs a forecast on the current data using ML methods. The monitoring component monitors current data in the respective regions under consideration and these data are made available to the ML forecast, in order to recognise future outages at an early stage. In case of a potential outage, the prediction features are mapped with the context entity. All features, date entries, and outage data are mapped. Identifiers, probability attributes within "Reason", and the "ImpactLocation" are derived from the outage prediction. This inductive learning process (by learning the KG) is then used to explain the results of the forecast to the user. This bridges the gap between symbolic and non-symbolic AI. By combining the results of the non-symbolic forecast model with the attributes of the scenario patterns, further inferences can be made, e.g.,

if Context.Influence = “Autumn Season”.

THEN Reason.Precondition = “Wind Speed”.

and the extension, e.g., by an expert/user, entering the activities Measure.action

Steps = ["planned downtime", "planned maintenance"]; for a so-called crisis scenario, the above rule can be extended to:

If Context.Influence = "Autumn Season"

THEN Reason.Precondition = "Wind Speed".

IF Reason.Precondition = "Wind Speed" THEN Measure.actionSteps

THEN Measure.action Steps = ["planned downtime", "planned maintenance"] (Janzen et al. 2022)

AISOP (s. table 18) relies on a knowledge base, specifically a knowledge graph, to store information. This includes both historical and newly acquired knowledge. If current events are "rediscovered" as patterns in this knowledge, an alarm is issued. The causes of this alarm are communicated to the user. This way of using the knowledge base can, therefore also be used directly as an explanation. In this case, it is not necessary or intended to "explain" the machine learning component and make it transparent for the user. The system has an interface for experts to model new scenarios. It also has a proof of data provenance so that it can be traced where the current data comes from.

System name	Domain of application	Modularity	Machine Learning Explainability	Symbolic Explainability	Support provenance	Adapt to user's need	Support compliance and obligation checks	Learning Component	Knowledge Base	Inference Engine etc.	Data Interface	Dialog Component	Explanation Component	Web Interface	Interface for User	Interface for Auditors	Interface for Experts (Knowledge Engineers)
AISOP	Utilities, scenario planning	yes	no	yes	yes	yes	no	yes	Knowledge Graph	Anticipating component	Monitoring Component	Responding component	N/A	yes	yes	no	yes

Table 18: Result of the analysis of AISOP (Jenzen et al. 2022)

Scenario Planning Adviser (SPA)

SPA is a system that takes input from news and social media and then combines it with expertise to create scenarios and explain the key risk drivers for the different future scenarios (here and in the following Sohrabi et al. (2018)). SPA is a decision support system: it is designed to assist an organisation in creating future scenarios and identifying and managing emerging risks, as well as classifying the key risk drivers. It combines changes in the economy on a global or local level. In doing so, knowledge engineering can ensure that conclusions with a potentially incomplete and biased input are mitigated. The architecture of the SPA is shown in figure 72. The architecture is modular and consists of three parts: the News Aggregator component, the Domain Knowledge component and the Scenario Generation & Presentation component.

The News Aggregator is used to analyse raw data from news channels and social media feeds. Text analysis methods are used to filter, process, and provide the relevant information for the respective area. The relevant information is provided based on a "topic model" and other information. The topic model is provided by the domain expert and contains a list of persons, organisations and keywords that are important for the respective subdomain. The output of the message aggregator is a set of relevant key risk drivers from which the domain expert or the business user can select a subset and use it for scenario generation and presentation.

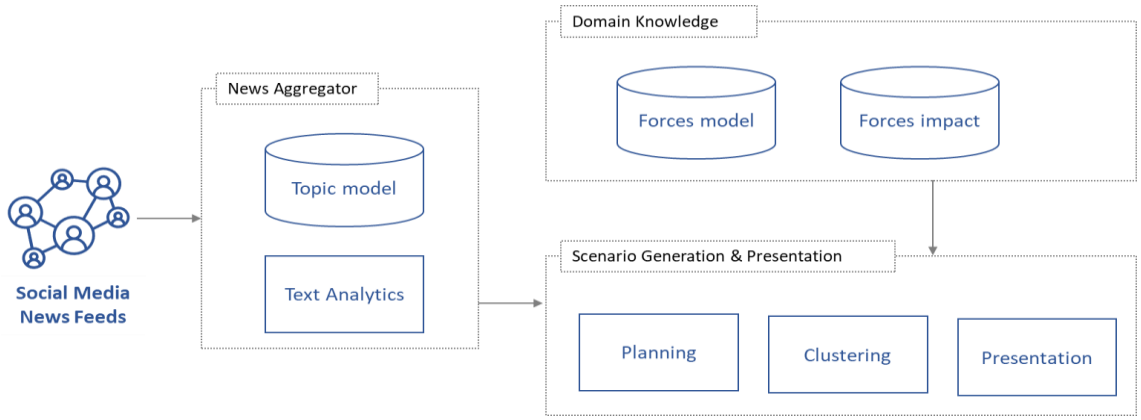

Figure 72: Architecture of SPA (Sohrabi et al. (2018))

The Domain Knowledge component captures the required domain knowledge based on two criteria: Forces Model and Forces Impact. The Forces Model is a description of the causes and consequences of a particular force, for example a social, technical, economic, environmental, and political trend, and is provided by a domain expert with little or no AI planning background. The representation of the force model is done with the help of mind maps. Figure 73 shows such a mind map, which illustrates the connection between the decline of currencies and falling commodity prices.

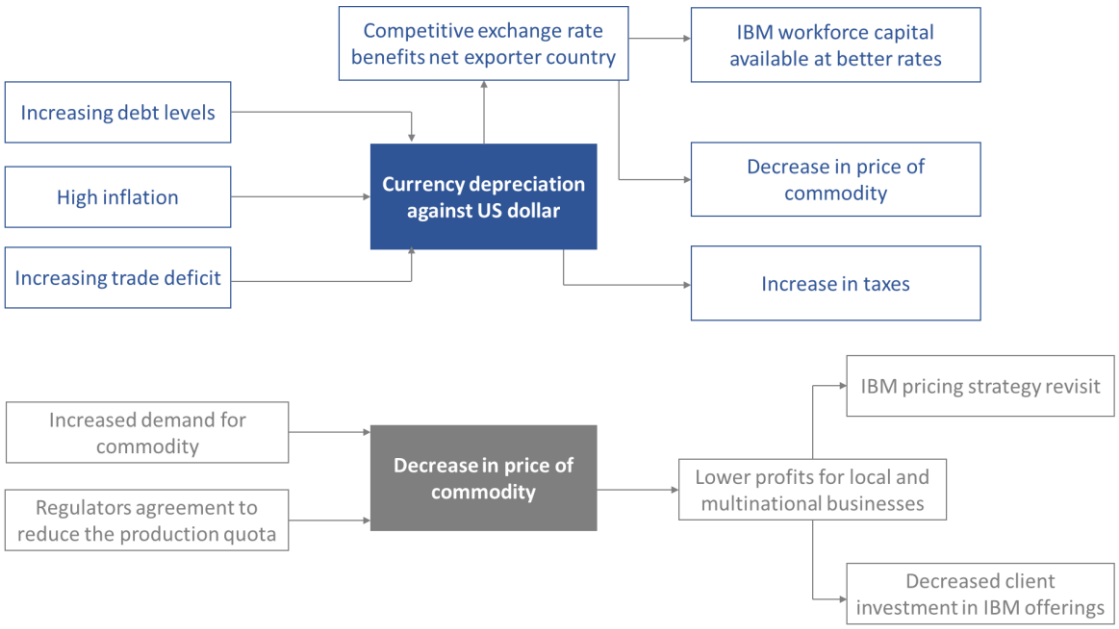

Figure 73: SPA forces model – Sohrabi et al. (2018)

The Forces Impact model is used to represent the probabilities and effects of a cause. The scenario generation component takes the domain knowledge and the main risk drivers and

automatically generates a planning problem from them, the solution of which produces a set of alternative scenarios in the post-processing step. The scenario planning problem (SP problem) is described as a set of tuples SP = {forces-model, forces-impacts, main risk drivers}. Here, the main risk drivers are a subset of the forces describing the current situation. These are proposed by component message aggregation. Each of the forces described in the force model can be used and defined as a main risk driver. The solution to the above SP problem thus consists of a set of alternative scenarios that consider the main risk drivers and describe a range of possible futures. The determination of probability, impact and importance is thereby considered based on the Forces Model and the Forces Impacts. According to Sohrabi et al. (2018), the theoretical background of the Forces model is in AI planning and plan recognition. The theoretical background of the Forces Model is on AI planning and Plan Recognition. The main idea is the planning task,

$$\Pi = \{F, A, I, G, cost\} \qquad \text{(f17)}$$

here being described in the STRIPS[37] formalism. Extended with the operator costs Here F is a set of Boolean flow equations, A is a finite set of actions, "cost" is a non-negative cost function, I is the initial state and G is the goal. The main idea is to minimise the cost, which is cumulative for all actions in the sequence, and thus find an optimal plan s (where s is a subset of the flow form F) for I. The use of mind-maps allows for a finite set of actions, "cost" is a non-negative cost function, I is the initial state, and G is the goal. The experts can process and model the knowledge using mind maps.

SPA uses several components and a knowledge base - a knowledge graph for storing knowledge. In this case, the forces model KG stores the forces influencing a scenario and the forces impact KG stores the effects, etc. If current events are "rediscovered" as patterns in this knowledge, an alarm is issued. Current data are entered into the system via the data service component called News Aggregator. Statements on data provenance are not made in the presentation by Sohrabi et al. (2018). The system has an inference component called Scenario Generation & Presentation. It uses a user interface and one for the experts. No statements are made about compliance with governance rules. There is also no presentation of a machine learning component. The cause-effect relationships are represented graphically using mind maps. The explanation is, thus, provided at the same time in the reporting

[37] S. Fikes & Nilsson (1971)

and does not require any additional explanations, as the data or scenario information (forces model and forces impact) use human-understandable language (s. table 19).

System name	Domain of application	Modularity	Machine Learning Explainability	Symbolic Explainability	Support provenance	Adapt to user's need	Support compliance and obligation checks	Learning Component	Knowledge Base	Inference Engine etc.	Data Interface	Dialog Component	Explanation Component	Web Interface	Interface for User	Interface for Auditors	Interface for Experts (Knowledge Engineers)
SPA	Business	yes	n/a	yes (mindmaps, language)	no	yes	no	yes - domain experts	yes (forces model, forces impact)	Scenario Generation & Presentation	yes "newsaggregator"	yes	yes	yes	yes	no	yes

Table 19: Result summarisation of the analysis of SPA (Sohrabi et al. 2018)

CALO

The Cognitive Assistant that Learns and Organizes (CALO) system was developed within an ambitious and multi-university program, initiated by the Defense Advanced Research Projects Agency (DARPA) program, to build a Personal Assistant that Learns (PAL) (here and in the following, McGuiness et al. 2004 and Chari et al. 2020). The CALO is a cognitive agent whose task is to assist with a variety of everyday office tasks. These tasks can be, for example, sending emails, creating memos, keeping a to-do list, etc. One of the best-known follow-up projects in which CALO or the Calo "technology" was used is the personal assistant Siri from Apple (s. figure 74).

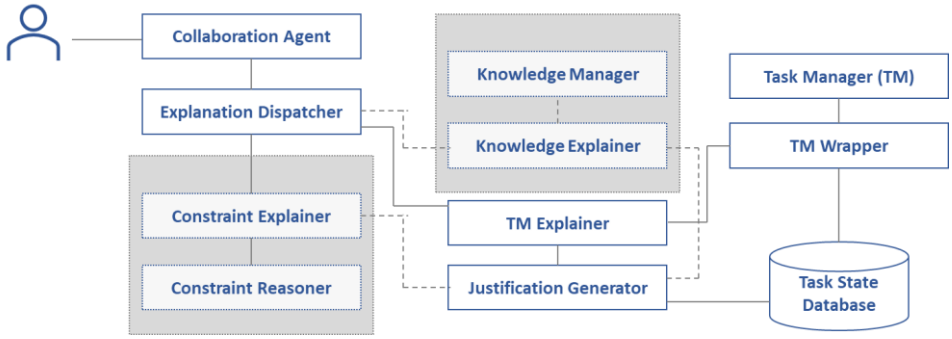

Figure 74: CALO by McGuiness et al. (2004)

CALO based on the Inference web as being one of the early modular explanation frameworks, one of the earlier works of McGuiness et al. (2004) (s. figure 75).

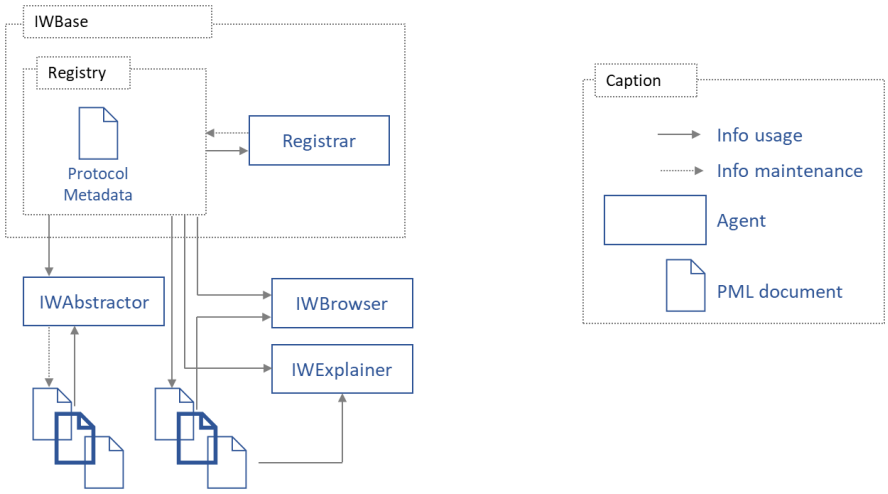

Figure 75: Inference Web (IW) Framework by McGuiness et al. (2004)

In doing so, this web-based system uses explanations created by the semantic web, description logic and expert systems communities. The origin of the information and evidence for inference traces were also provided for the user.

To better support the user's understanding, the system could create summaries for explanations avoiding lengthy proofs that might overwhelm the user. The explanations could be presented in a variety of formats and even had a built-in explanation dialogue that displayed questions and answers and allowed the user to ask follow-up questions.

The framework was based on a modular architecture, used PML and consisted of an IWBase (a data store for the meta-information about the information used by the framework), an IWAbstractor (an abstractor component that converts long Proof Markup Language - PML - proofs into explanations), an IWExplainer (an explanation dialogue component that generates explanations for users) and an IWBrowser (a browser to display the explanations). While the Inference Web Framework did not contain a context-specific component of its own, it did provide some options for context modelling and was thus quite capable of providing a wide range of customised explanation functions.

In terms of task reasoning explanations, ICEE served as an explanatory component in the CALO system. Statistical and deductive methods worked alongside several reasoning techniques, including task processing and numerous learning components. The reasoning techniques used in CALO were able to use multiple sources of knowledge to draw conclusions (s. table 20).

System name	Domain of application	Modularity	Machine Learning Explainability	Symbolic Explainability	Support provenance	Adapt to user's need	Support compliance and obligation checks	Learning Component	Knowledge Base	Inference Engine etc.	Data Interface	Dialog Component	Explanation Component	Web Interface	Interface for User	Interface for Auditors	Interface for Experts (Knowledge Engineers)
CALO	Business	yes	no	yes	yes	yes	yes	yes	yes	Explanation Dispatcher, Constraint Explainer, Constraint Reasoner, Knowledge Manager , Knowledge Explainer, Task Manager Explainer, Task Manager Wrapper, Justification Generator; Task Manager	yes	Collaboration Agent	yes	yes	yes	no	yes

Table 20: Result of the analysis of CALO (McGuiness et al. 2004)

EES Framework

In their research work on second-generation explainable expert systems, Swartout & Moore (1991) defined a list of "desiderata", which explainable expert systems have to follow- the interesting aspect of these "wishes" is that they not only concern the form and content of the explanation – it also concerns the impact of the explanation on the whole system, the design, how it is built, and at last, how it performs (Moore & Paris, 1991; Swartout & Moore, 1993). Here, the word "desideratum" is reformulated into "requirement" for an intelligent system module for the explanation of an XAI system (s. figure 76).

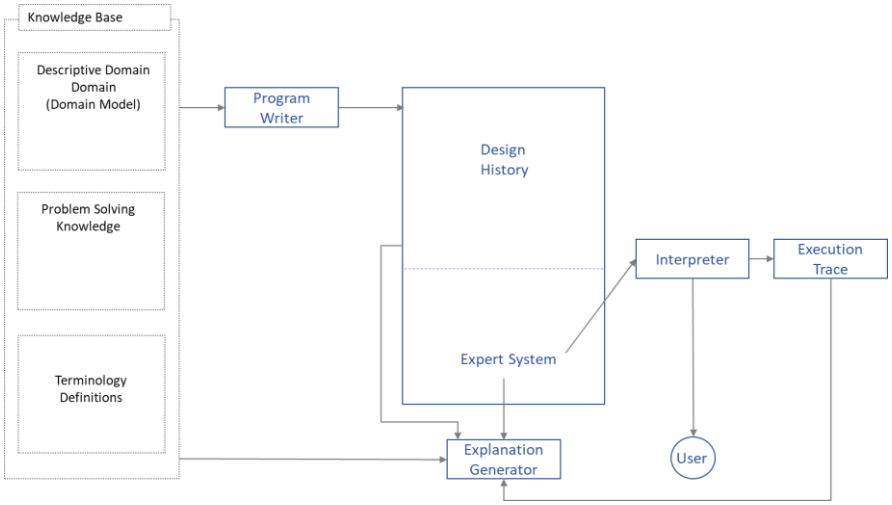

Figure 76: Architecture of the EES Framework (Swartout & Moore, 1993)

The main components of the EES framework consist of the EES knowledge base, which distinguishes between three different types: terminological knowledge, a domain model, and a library of plans for problem solving. Terminological knowledge expresses how terms are defined in the domain and gives these terms explicit semantics. In the original framework, these terms were defined in the Loom programming language. The domain model

contains the facts of the domain, and how the different terms in the domain relate to each other, for example, an electronic component and its circuits served. The domain model describes the domain but not the solution for the problem. The problem knowledge is described in the problem-solving knowledge of the knowledge base. For example, when searching for "diagnose a component":

```
(define-plan diagnose-component
 :capability (DIAGNOSE (obj (c is (inst-of COMPONENT))))
 :method
  (let
    ((actual-symptoms
     (loop for each symptom in (POTENTIAL-SYMPTOM c)
      when
       (DETERMINE-WHETHER-DESCRIBED (obj c) (by symptom))
      collect symptom))
    (FIND-CAUSES (obj actual-symptoms) (of c))))
```

These lines of program code mean:

To diagnose a component, the system finds the potential symptoms of the component. For each symptom, the system determines whether the component exhibits the symptom. These are called the actual symptoms. The system then finds the causes of the actual symptoms and returns them (Swartout & Moore, 1993).

Another important component of the EES is the Automatic Programmer. This works in a kind of refinement driven way. It starts at a high level with a goal that represents what the expert system is meant to do. Then the program writer searches its library of problem-solving knowledge for a plan whose capability matches the description. The goal and its capability description are translated into Loom (programming language => automatic programmer). For Swartout and Moore (1993), the EES framework served to implement the "desiderata" presented herein in Chapter 3.

The EES framework is a concept that relates to expert systems and, therefore, has nothing to say about machine learning. It has a knowledge base with various knowledge such as descriptive domain model, problem solving knowledge and terminology definitions. It has inference components with the program writer, interpreter and explanation generator (s. table 21).

System name	Domain of application	Modularity	Machine Learning Explainability	Symbolic Explainability	Support provenance	Adapt to user's need	Support compliance and obligation checks	Learning Component	Knowledge Base	Inference Engine etc.	Data Interface	Dialog Component	Explanation Component	Web Interface	Interface for User	Interface for Auditors	Interface for Experts (Knowledge Engineers)
EES	Program Advisor	yes	no	yes	no	no	no	yes	yes	yes	yes	yes	yes	no	yes	no	yes

Table 21: Result of the analysis of the EES (Swartout and Moore, 1993)

Summary of RA I- RA IV																	
System name	Domain of application	Modularity	Machine Learning Explainability	Symbolic Explainability	Support provenance	Adapt to user's need	Support compliance and obligation checks	Learning Component	Knowledge Base	Inference Engine etc.	Data Interface	Dialog Component	Explanation Component	Web Interface	Interface for User	Interface for Auditors	Interface for Experts (Knowledge Engineers)
AISOP	Utilities, scenario planning	yes	no	yes	yes	yes	no	yes	Knowledge Graph	Anticipating component	Monitoring Component	Responding component	N/A	yes	yes	no	yes
SPA	Business	yes	n/a	yes (mindmaps, language)	no	yes	no	yes - domain experts	yes (forces model, forces impact)	Scenario Generation & Presentation	yes "newsaggregator"	yes	yes	yes	yes	no	yes
CALO	Business	yes	no	yes	yes	yes	yes	yes	yes	Explanation Dispatcher, Constraint Explainer, Constraint Reasoner, Knowledge Manager , Knowledge Explainer, Task Manager Explainer, Task Manager Wrapper, Justification Generator; Task Manager	yes	Collaboration Agent	yes	yes	yes	no	yes
EES	Program Advisor	yes	no	yes	no	no	no	yes	yes	yes	yes	yes	yes	no	yes	no	yes

Table 22: Summary result overview of the analysis of all systems investigated (RA I – RA IV)

Table 22 shows a summary of all the results. In the following chapter, we will continue to collect the various requirements and create the basis for the creation of the reference architecture.

5.2.3 Gathering and synthesis of the Requirements

The problem was described in Chapter 5.2.1 (see Chapter 5.2.1). Similarly, the corporate planning model was identified and described for the entire scope under consideration and the corresponding decision variables etc. were derived (see chapters 2, 3). Since AI models are currently used in connection with scenario planning and in relation to the support of integrated business planning (see Chapters 2.3, 3 and 5.2.2 AISOP, SPA), strategic and tactical planning with scenario planning and integrated corporate planning, in which forecast models are used, for example, were used as the scope. The stakeholders relevant for the application were identified and placed in the context of the requirements for AI and the decision variables. These requirements essentially related to qualitative requirements and

to constraints. In the following, therefore, further functional requirements will be briefly included based on two use cases. For a summary of the other requirements, please refer to the relevant chapters (s table 23, 24 and 26).

Use Case 1:	**Strategic Planner - Explain Strategic Scenario Analysis and Planning**
Actor(s):	Strategic Planner, Management Board
Summary Description:	(Re_fish) Gives an explanation and reason to a strategic planner and analyst why a specific (set) of scenario(s) was selected by the AI Scenario Planning Application
Priority:	Must Have
Status:	Medium Level of details
Pre-Condition:	• The Strategic Analyst logs in to the explainer component (or the component will be embedded in the strategic planning application) and wants to get an explanation of the selected scenarios. • The "Re_fish" is online properly- status is "green".
Post-Condition(s):	• The strategic analyst got the sufficient explanation of the selected scenarios. • The selected scenarios are approved by the strategic analyst and by the management board and handed over to the next process step. • The (set of) strategic scenarios was rejected by the strategic analyst
• Basic Path:	1. The strategic analyst enters the log in data. 2. The Re_fish verifies the login. 3. The Re_fish provides the Strategic Analysis Explanation frontend. 4. The strategic analyst selects by menu the scenario planning he/she wants to analyse. 5. The requested scenario opens with explanations. 6. The strategic analyst can change the views of the presented scenarios. 7. The strategic analyst can open the dialogue component. 8. The dialogue component opens and greets the strategic analyst. 9. The strategic analyst can start to ask questions in natural language. 10. The dialogue component answers- explains the questions and explains the decisions made by the strategic analysis AI application in natural language. 11. The analyst gets presented with a visualisation of the explanation. 12. The analyst can change the graphical presentation as well as entering questions in natural language. 13. The analyst can request the data provenance of the model and the data the decision was made. 14. Alternative paths

Alternative Path:	14a. The analyst can document the whole analysis and all explanations and “hand over” to the management board to further approve. 14b. The analyst can document the whole analysis and all explanations and reject the (set of) scenarios selected by the AI application with remarks. 15a. The management board can start the analysis by reviewing the analysis made by the strategic analyst. 15b. The strategic analyst can start over to build a new strategic planning- scenario application round.
Functional Requirements:	F1: Web Frontend F2: User/ role-based security. F3: Frontend (role/ user dependent)- interactive with graphical presentation F4: (Embedded) frontend for natural language dialog F5: Accept/ reject function
Business Rules:	B1: Authentication B2: Authorisation B3: User role B4: Selected (set of) scenarios. B5: Reject B6: Accept B7: Handed over for further approval. B8: Reject (set of) scenario(s)
Non-Functional Requirements:	NF1: Logging of all tasks NF2: Security password entry NF3: Explanations regarding stakeholder map fulfilled. NF4: Language support NF5: Compliance regarding requirements – s. stakeholder map fulfilled

Table 23: Use Case sample – use case 1 strategic planner

Use Case 2:	**Tactical Planner – Demand Planner**
Actor(s):	Tactical Planner- Demand Planner
Summary Description:	(Re_fish) Gives an explanation and reason to a tactical planner and analyst why a specific forecast was selected by the AI forecast application
Priority:	Must Have
Status:	Medium Level of details

Pre-Condition:	• The tactical planner logs in to the explainer component (or the component will be embedded in the tactical planning application) and wants to get an explanation of the selected forecast. • The "Re_fish" is online properly- status is "green".
Post-Condition(s):	• The tactical planner got the sufficient explanation of the selected forecast. • The selected forecast(s) are approved by the tactical planner and analyst and by the management board and handed over to the next process step (consensus meeting/ plan). • The forecast was rejected by the tactical planner.
• Basic Path:	1. The tactical planner enters the log in data. 2. The Re_fish verifies the login. 3. The Re_fish provides the tactical planner the explanation frontend. 4. The tactical planner selects by menu the scenario planning he/she wants to analyse. 5. The requested forecast opens with explanations. 6. The tactical planner can change the views of the presented scenarios. 7. The tactical planner can open the dialogue component. 8. The dialogue component opens and greets the tactical planner. 9. The tactical planner can start to ask questions in natural language. 10. The dialogue component answers, explains the questions, and explains the decisions made by the forecast AI application in natural language. 11. The tactical planner gets presented a visualization of the explanation. 12. The tactical planner can change the graphical presentation as well as entering questions in natural language. 13. The tactical planner can request the data provenance of the model and the data the decision was made. 14. Alternative paths
Alternative Path:	14a. The tactical planner can document the whole analysis and all explanations and "hand over" to the management board to further approve. 14b. The tactical planner can document the whole analysis and all explanations and reject the forecast selected by the AI application with remarks. 15a. The management board can start the analysis by reviewing the analysis made by the tactical planner. 15b. The tactical planner can start over to build a new forecast- application round.
Functional Requirements:	F1: Web Frontend F2: User/ role-based security. F3: Frontend (role/ user dependent)- interactive with graphical presentation F4: (Embedded) frontend for natural language dialog F5: Accept/ reject function
Business Rules:	B1: Authentication B2: Authorisation B3: User role B4: Selected forecast. B5: Reject

	B6: Accept B7: Handed over for further approval. B8: Reject forecast
Non-Functional Require-ments:	NF1: Logging of all tasks NF2: Security password entry NF3: Explanations regarding stakeholder map fulfilled. NF4: Language support NF5: Compliance regarding requirements – s. stakeholder map fulfilled

Table 24: Use case sample – use case 1 tactical planner (demand)

Functional Requirement	Solution (proposed)
F1: Web Frontend	Web frontend, configurable, with menu and role based configuration
F2: User/ role-based security.	Integration with a user authentication/ authorisation service (e.g. MSADS)
F3: Frontend (role/ user dependent)- interactive with graphical presentation	Interactive graphical component embedded in frontend
F4: (Embedded) frontend for natural language dialog	Interactive natural language processing ("chat bot") frontend
F5: Accept/ reject function	Web frontend accept/ reject function

Table 25: Sample of functional requirements for use case 1 and use case 2

In a further iteration to Re_fish and/or in an instantiation, the use case descriptions are to be carried out again in the specific case in order to record the requirements and provide them with solutions.

5.2.4 Re_fish Business Architecture

The individual viewpoints of the Re_fish reference architecture are shown below. These are to be further decomposed in a further iteration or else to be developed in the context of an instantiation. Figure 77 shows the Re_fish business architecture and where the Re_fish architecture fits into the scenario under consideration - corporate planning and scenario planning and integrated business planning, with strategic planning shown on the left.[38] The planning process starts with the definition of strategic corporate goals by the management

[38] The description language used here is Archimate Modeling Language (https://www.opengroup.org/archimate-forum/archimate-overview).

board/board of directors. The strategic planning process is divided into the processes develop strategy and strategic analysis. The AI application for scenario analysis and selection can also be found in strategic analysis. This is accompanied by the process explanation AI scenario planning and analysis. The process description for this part of the Re_fish reference architecture is the problem description earlier in this chapter.

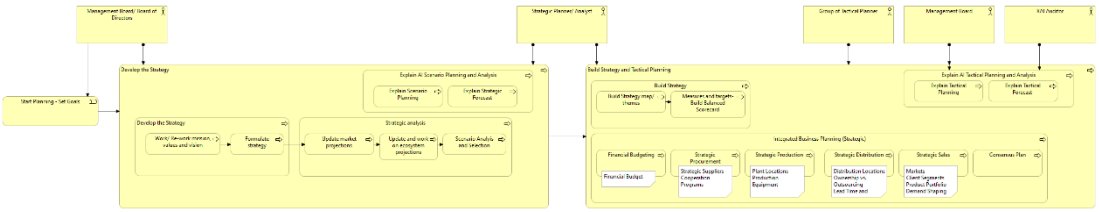

Figure 77: The Re_fish Business Architecture

The right-hand area of the business architecture is tactical planning within the framework of the Re_fish scenario. The transfer from strategic planning to tactical planning takes place after the selection of a selected set of scenarios and corresponding targets, which are also reflected in the balanced scorecard (see Chapter 2.3). Tactical planning in this presentation is integrated business planning and, excluding financial planning, S&OP planning. The actors of this planning are the group of tactical planners (demand, supply, etc.) The planning starts with the further processing of the results (strategic scenario, KPI's Balanced Scorecard, adjustments of the organisational structure, Capex planning, etc., see stakeholder map) from the strategic planning. The processes of tactical planning correspond to the model presented in Chapter 2.3.2. The individual parameters, strategic input parameters, external parameters and decision variables are also mapped. The process "Explain AI Tactical Planning and Analysis" is briefly described as a use case in Chapter 5.2.3. The requirements and constraints for the "Explain" processes, both at the tactical and strategic level, can be found in the stakeholder maps.

5.2.5 Re_fish Application Architecture

The most important architectural representation in the development of the reference architecture for trustworthy AI is certainly the representation of the application architecture (s. Sufi, 2022; Takeuchi et al. 2021). During the development, planning was identified as the

most important component in the management process of companies in the process industry, based on the preliminary analyses. The focus is on scenario planning and, not least because of the high degree of integration of companies in the process industry in highly complex global supply chain networks, on integrated corporate planning (sales and operations planning, S&OP). The Re_fish application architecture is shown in figure 78 as a "Re_fish Explainer" application. - The application is composed of six modules. These modules are Re_fish Data Service, Re_fish Subsymbolic Module, Re_fish Symbolic Module, Re_fish Audit Module, Re_fish Explanation Module and the Re_fish User Dialog Module. The Re_fish Data Services Module is used to transfer data from the AI applications for scenario planning and forecasting (strategic, tactical). In addition to data relating to scenario planning (mapping of scenarios in the KG database), all status parameters of the symbolic and non-symbolic AI models/applications used are also transferred. The Data Services component accesses the same data as the AI Data Services component, thus ensuring that the same data is also used for the explanation. The components in the Re_fish Data Services area have the task of identifying possible biases that are present in the data and are present at the beginning of the life cycle (data transfer) and to correct them if necessary, or at least to point them out. This Data Services component also serves the permanent monitoring of relevant data sources, for example for existing risks and their impact on the company supply chain, as they result from global or local changes in the situation (e.g., through blocking of the Suez Canal, etc.). The AI model or application data, together with the relevant situation data, are evaluated using the subsymbolic module, e.g., with selected XAI machine learning models (see Chapter 3.3.1). In addition, the entire status is available in the task tracker component, so that it can be determined at any time which decision a non-symbolic AI model has made and on the basis of which data. The respective ML XAI models are selected via a library and can be extended or updated. The Re_fish Symbolic Module is used to prepare the data so that they can be persisted in the knowledge base (KG database, e.g., Neo4j). In addition, this module serves to prepare the data within the framework of the Re_fish Explanations Module. This module contains the inference component of Re_fish, which can be implemented based on various existing concepts, for example, as proposed, by implementing a causal inference engine. All relevant metadata about events, status, etc. are collected in parallel in the component called Re_fish Audit Module. This component is separate and thus offers the possibility to ensure compliance, be it the avoidance or just the information about existing biases and the audit of the AI

systems connected to Re_fish. The Re_fish User Dialogue component is used by the various user groups to conduct a graphically guided dialogue to explain the AI decisions, or a dialogue based on natural language to explain the Re_fish system.

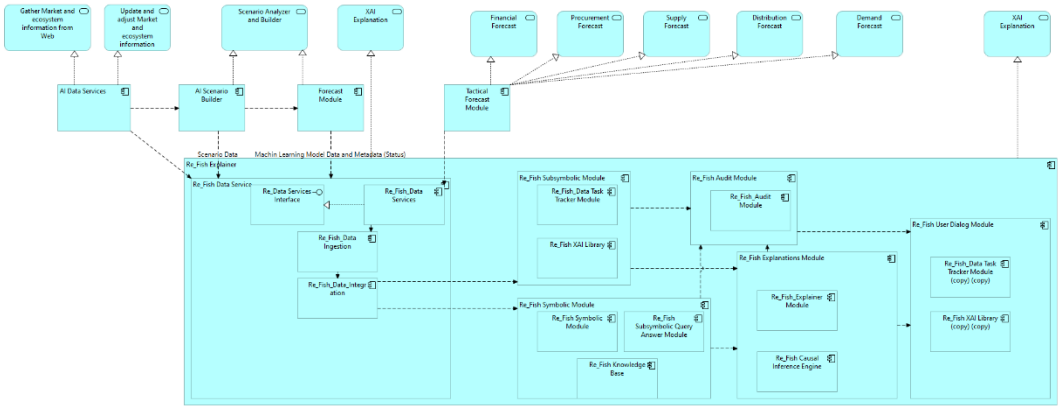

Figure 78: The Re_fish Application Architecture

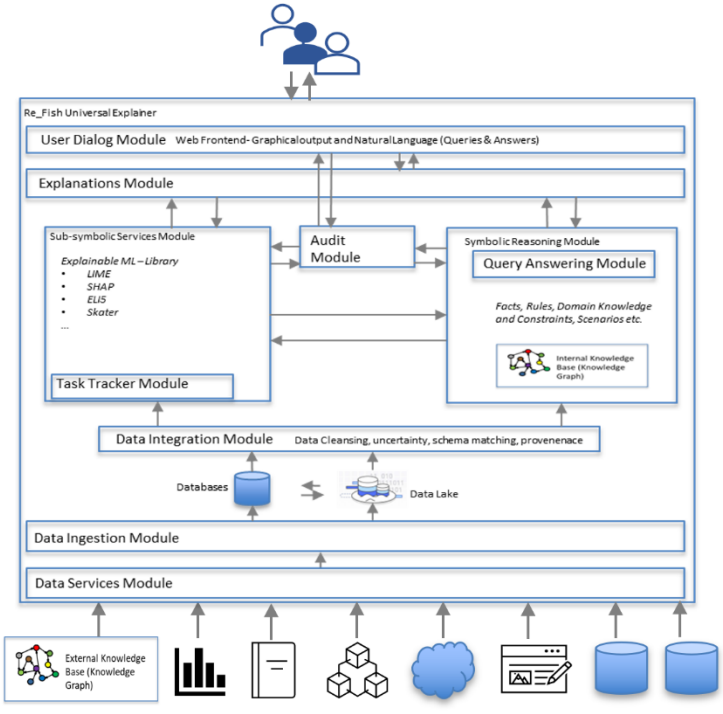

Figure 79: The Re_fish Reference Architecture

Figure 79 shows the conceptual representation of the Re_fish reference architecture. This corresponds to the above-mentioned representation. Here, the different levels in a layered architecture are shown from the bottom, the area of the data sources and the connection of

the data sources. Here, too, a distinction is made between the three levels of Data Services, Data Ingest and Data Integration Module. the two components Subymbolic Services Module and the Query and answering Module are also shown. The Audit module is also shown as a separate module. The dialogue with the user takes place via the User dialogue module. The explanations module is located between the subymbolic and symbolic modules.

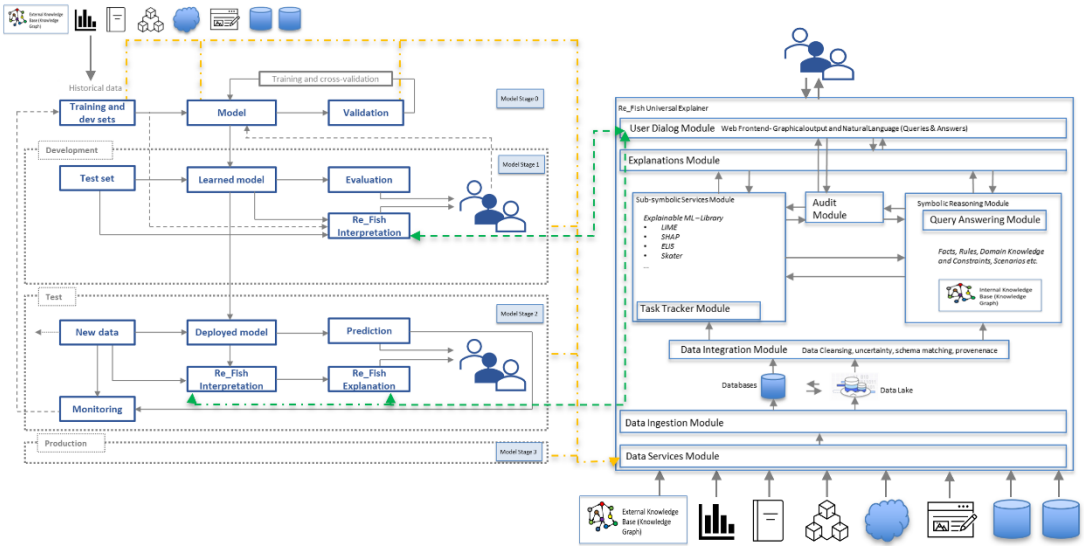

Figure 80: The Re_fish Reference Architecture Sample with ML model (Dev/Test/Prod)

In the illustration s. figure 80, a logical representation of the Re_fish Explainer component in its context is shown. It can be seen that data from different sources, such as the web, are processed via a component and are processed in a knowledge database (a knowledge graph). When AI models use models in different contexts that need to be explained, the "Explainer" component explains the models and the results using the knowledge component. The knowledge available in the knowledge base can be adapted via a frontend.

The explanation of the individual components of the Re_fish conceptual architecture is as follows.

Data Services

The Data Services module gives the Explainer module the ability to provide and process a wide range of data. The idea is that it is possible to collect sentiment data on political trends in the countries where the company operates, or trends in external variables such as commodity prices, etc., that can be used for scenario planning, trends in demand, and so forth. What is important when transferring data is not only the time stamp, but also the data

origin. Up to the possibility of using blockchain technologies, the data can originate from streaming data, semantic databases, social media data, etc. The data service module must have the appropriate interfaces. It is important that the AI models used have the same database as Re_fish.

Data transfer module

In this module, the data is transferred from the data service, and an initial evaluation of the data is carried out. The evaluation can only be done on the basis of rules, which must be carried out in a rule system by a developer, in cooperation with an expert.

Data integration module

The data integration module has the task of further filtering the data and performing data cleansing so that the data can be stored in the Knowledge Graph and also processed further in the Subymbolic Service module. Therefore, data inconsistencies, schema matching, etc. must be performed here.

The Data Integration module also needs an interface towards the AI module/agent to get metadata about the data being used by the AI model and the results. A tracking/login service is logging all steps the AI model is doing and documents these steps.

Subymbolic Services Module

In the Subymbolic Services, the models are provided in a library to support the respective application area. If, for example, the AI model/agent uses a specific ML method, e.g., an ANN, LIME or SHAP can be used here as an interpretation option, the results are passed to the Explanation Module on the one hand, and to the Symbolic Services on the other.

Symbolic Services Module

In the Symbolic Services Module, the explanation of the subsymbolic Interpretation is taken and combined with symbolic knowledge. The rule-based system is able to combine data and low-level knowledge, update and replace high-level knowledge. Query answering is based on well-defined resonating mechanisms, like Judea Pearl's (2019) inference engine. The explanation requests are handled by the query-answering module. This symbolic service module will provide, for example, the evaluation of possible scenarios within the scenario planning component of the AI module. It will also provide explanations for the prediction, by combining the appropriate non-symbolic method with the relevant data from

the ingested and integrated data. Domain knowledge is used in this module to limit inconsistency problems or biased data, or concept drift, by evaluating the results of the subymbolic service module. This service module can be considered as the superego in the AI model, similar to the superego in Sigmund Freud's personality model. There is also a correcting functionality, which must work before answering the question, as a wrong or biased answer could have a severe negative impact.

Explanation Services Module

The Explanation Services Module is providing the overall Explanation, based on the findings regarding the combination of both the symbolic and the non-symbolic services modules.

Query & Answer Service

The results will be provided to the dialogue component; the query and answer service module is providing the HCI (the human-computer interface) to the user, and the Query and Answer Service is providing the Explanation (the Answer of the query in a human-understandable format) that can be in natural language or a graphic output, e.g., a causal graph based on the inference of the explanation.

Admin and Development Service

Component to Design is about developing and maintaining the components, e.g., the domain knowledge, or adding non-symbolic explanation models, and so forth.

Audit Service Module

The Audit Service Module provides a possibility to check the compliance of the AI model as well as the compliance of the explanation service. It is highly secured to prevent manipulation.

The Re_fish Reference Architecture. One of the most important design specifications for knowledge-based systems is the separation between the representation of knowledge and the processing of knowledge, as the way in which knowledge is stored essentially depends on how it is represented and processed. For example, rule-based systems whose knowledge can be represented by if-then rules require a rule-based interpreter. The knowledge base can be subdivided, e.g., as evidence-related or case-related knowledge about the problem

area under consideration. There is rule-based knowledge -- on the one hand, domain-specific knowledge, and on the other, general knowledge. The structure of such a knowledge-based system was shown in chapter 3.3.2; the components are composed of the following:

- The knowledge base, which consists of a temporary working memory and a permanent rule-based knowledge memory.

- A knowledge processing component, which is separate from the knowledge base.

- A knowledge acquisition component that supports the construction of the knowledge base.

- An explanation component that can communicate understandable explanations to the user which can explain how the conclusions were reached.

- The dialogue component, for communication with the expert system. The recommendation here is to distinguish between the component for the experts in charge of building and developing and a user interface for the users.

In Chapter 3.3.2, it was pointed out that causality is one of the most important methods of explanation. Chari et al. believe that in addition to Causal Methods, Neural-Symbolic AI systems, representation techniques (such as Distributed General Ledger (DLG technologies) are also important, as they enable the origin and secure distribution of data.

The following is a brief outline of how a causal inference explainer can be implemented (Explainer Module in Re_fish). To this end, the Causal Inference engine according to Judea Pearl is first introduced.

Causal Inference Engine by Judea Pearl

In the "Book of Why", Judea Pearl (2018) (s. also Pearl, 2019, 2009a, 2009b; Halpern, 2015; Halpern & Pearl, 2005) presented a causal inference machine. Such an inference machine accepts three types of input: Assumptions, Queries and Data. As output, the inference machine also produces three different types. The first output is the answer to the question as to whether the inference machine can answer, assuming both the given causal model and infinite, perfect data. The answer to this question is a yes/no. If the answer is yes, then the inference engine produces an estimate. This is a mathematical formula that can be seen as a recipe for determining an answer from hypothetical data. From the data

entered, the estimand, the recipe, generates an actual estimate with associated statistical estimates for the degree of estimate uncertainty. This uncertainty expresses the data situation and any possible measurement errors or missing data.

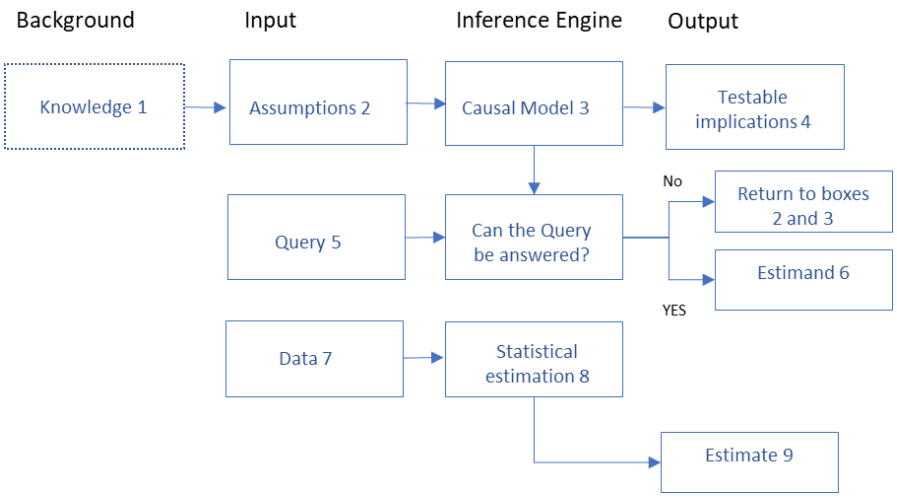

Figure 81: Causal inference engine – based on Pearl (2019)

Pearl describes the inference engine in terms of nine elements (see figure 81). 1. Knowledge: this represents the agent's past experiences and includes past observations, activities, education, and cultural mores, deemed to be interesting for the particular query. This knowledge remains implicit and is not made explicit. 2. Assumptions: only the knowledge that is made explicit by stating assumptions is used, while the other part of the knowledge remains implicit. 3. To show causality, Pearl suggests using a graphical method such as a diagram, in the simplest way, if Y "listens" to X, an arrow from X to Y shows causality. 4. The result of the causal or listening pattern from 3 can be used to test the model. The testable patterns are created by using the data, so another engine is needed that takes the data from the testable implications (4) and data (7), and tests the model for "accuracy". 5. The queries to the inference engine are as follows: What is the probability that X does Y (or P(X| (do Y)))? 6. The Estimand, namely the statistical value of 5., but when the model shows that X and Y depend on some other third variable Z that is not known, the query is unanswerable: P(X|Y,Z) x P(Z). 7. The data enter the estimation model, but the data are mute as to the causal relationship, and the estimator must use the data within the estimation model. 8. Approximation of the estimate. 9. The last step reflects the possibility of the examples studied, so that a statement of the above causality can be made. The

new knowledge thus obtained is then incorporated into the knowledge base. If the models do not provide the anticipated result, then a start-over must take place on box 3.

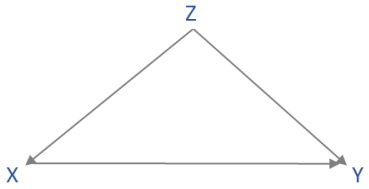

Figure 82: General cause and effect model according to Judea Pearl, Pearl (2019)

The general cause-and-effect model according to Pearl looks as shown in figure 82. It is Q=P(Y|do(X)), where X has an effect on Y and both depend on Z. Pearl formulates the overall problem as a Bayesian equation in that $\exists z = \sum_z P(\mathrm{Y}|\mathrm{X},\mathrm{Z})\mathrm{P}(\mathrm{Z})$ with gender (Z) is a confounder for the effect that an action (X) has on (Y).

Level (Symbol)	Typical Activity	Typical Questions	Examples
1. Association $P(y\|x)$	Seeing	What is? How would seeing X change my belief in$Y$?	What does a symptom tell me about a disease? What does a survey tell us about the election results?
2. Intervention $P(y\|do(x), z)$	Doing	What if? What if I do X?	What if I take aspirin, will my headache be cured? What if we ban cigarettes?
3. Counterfactuals $P(y_x\|x', y')$	Imagining, Retrospection	Why? Was it X that caused Y? What if I had acted differently?	Was it the aspirin that stopped my headache? Would Kennedy be alive had Oswald not shot him? What if I had not been smoking the past 2 years?

Figure 83: Three level causal hierarchy according to Judea Pearl (Pearl, 2019)

The three-level causal hierarchy model is shown in figure 83. The gradation shown here is reflected in the description of the stakeholder map and the questions and explanations mentioned there s. Chou et al. (2021) for an analysis of usage of counterfactuals within XAI).

The symbolic module together with the explainer can thus provide “Why” and “How” questions, using the semantic knowledge base:

Delivery delay of raw material A due to capacity bottlenecks at the port of Kaohsiung causative agent (production plan cannot be executed, switch to alternative supplier (according to supplier list)).

Based on the transitivity $\forall x, y, z(R(x, y) \& R(y, z) \rightarrow R(x, z))$- the non-compliance with the production plan (or the risk) can be represented in this way. By further using the knowledge (domain knowledge) from the ontology (the knowledge base), sentences can be output by means of the explainer and the dialogue component that can be understood by a human: "Production plan cannot be adhered to due to capacity bottlenecks at the port of Kaohsiung for raw material A".

One of the most important questions which has remained unanswered until now is, how can the symbolic and non-symbolic modules of Re_fish be linked? Figure 84 (based on Diwedi et al. 2022) shows how the symbolic module is linked to the non-symbolic module.

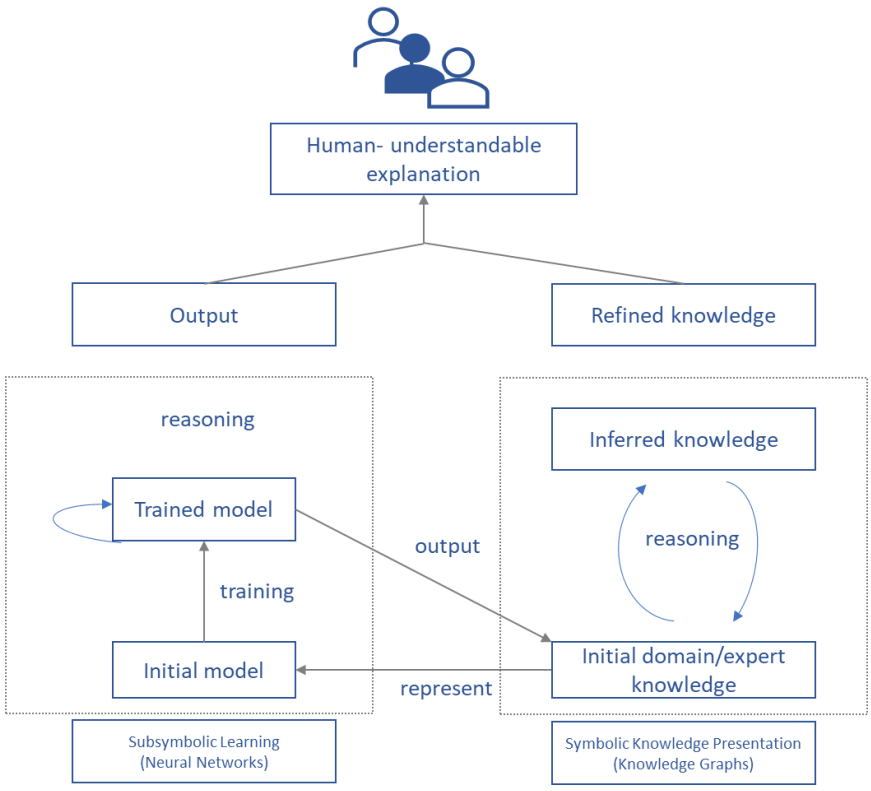

Figure 84: Sample for a neural symbolic system – using Neural Networks and LIME

The neural network is trained to predict disruptions in the supply chain. The symbolic module of Re_fish consists of a knowledge graph database (knowledge base) and contains

typical scenarios and corresponding indicators of disruptions in the company's supply chain. The neural model is now trained with data from all possible and meaningful data sources (see Data Service Module). The knowledge base is also trained with new knowledge about typical parameters and parameter patterns, and the decisions (predictions) of the neural model. The predictions learned from the neural model, such as the analysis of LIME and the predictions learned from the Knowledge graph (as well as the already-existing knowledge), are now used together to output human-understandable statements about the results (Explainer Module - Dialogue Module). The Query & Answer Module (or the dialogue components) can even be used to send queries to the system - What other faults are there? What other faults are there in Taiwan?

Re_fish makes that possible by combining the two approaches, namely the strength and speed of neural approaches with symbolic, human-comprehensible explanations which are based on knowledge.

5.2.6 Re_fish Technology Architecture

The technical architecture of Re_fish is shown in figure 85. A distinction is made between an integration server component, which represents the entire data service component, including the intermediate persistence through storage of the data in other databases and/or a data lake, and the database component with the knowledge database (knowledge graph). The application server component maps the entire level above (in terms of the layered architecture) the data services. This means that the Re_fish subsymbolic module, Re_fish symbolic module, Re_fish audit module, Re_fish explanations module and the Re_fish user dialogue module are technologically mapped here.

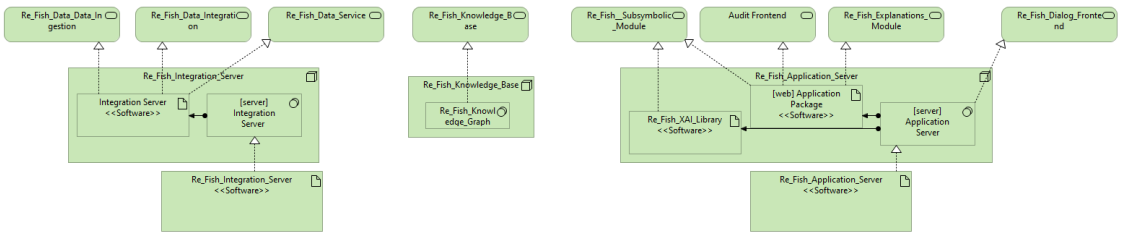

Figure 85: The Re_fish technology architecture

5.2.7 Re_fish Overall Architecture

The overall architecture summarises the individual viewpoints of the Re_fish reference architecture (s. figure 86). These are made up of the business architecture, the application architecture, and the technology architecture. The Re_fish reference architecture, if it is mapped with the components shown and is also instantiated, i.e., implemented, for a concrete use case, fulfils the requirements of the stakeholders - which were recorded in stakeholder maps A and B - as well as the functional requirements in the explicability of AI in strategic scenario planning and in tactical operational planning. Specifically, the requirements - RC1 - RC29, requirements resulting from the decisions DE1 - DE28 and the functional requirements resulting from the concrete application of the Re_fish reference architecture, here F1 - F5, avoidance of bias (see also Chapter 5.2.8) Bias 1- 7, etc. are the most important.

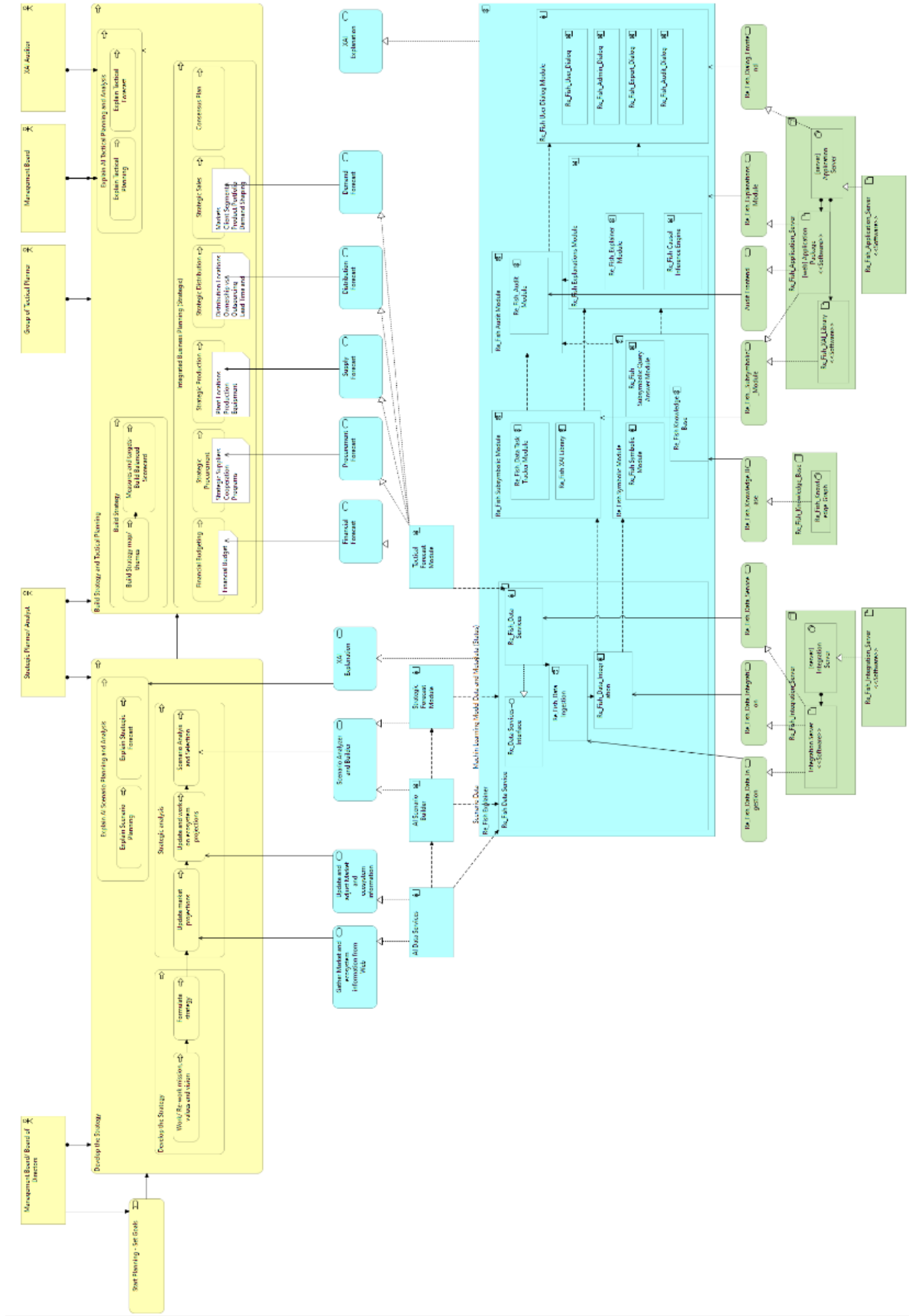

Figure 86: The Re_fish Overall Architecture

5.2.8 Re_fish Lifecycle Management

In the context of lifecycle management for a trustworthy AI system, certain perspectives need to be considered in more detail. Suresh and Guttag (2019) from MIT have identified seven sources that can negatively influence e.g., non-symbolic ML (s. figure 87).

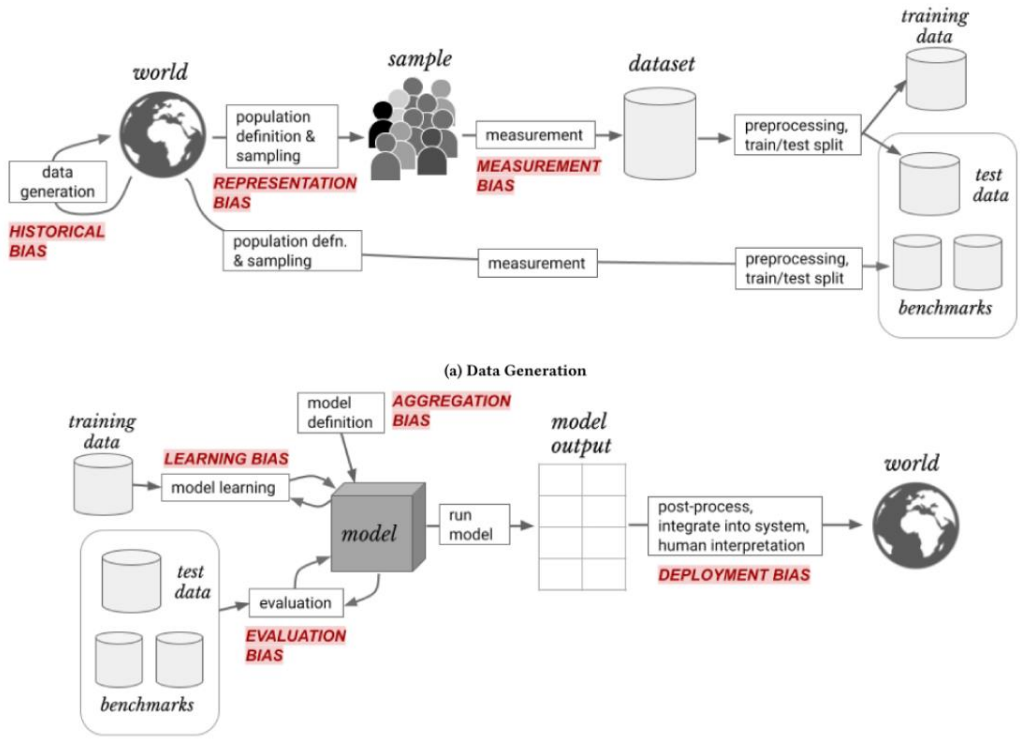

Figure 87: Types of bias – based on Suresh and Guttag

This is commonly referred to as bias. The first possible source of bias is a so-called (1) historical bias. This is characterised by data selection which is made such that a pre-existing (human) bias in the historical data used to learn the model is perpetuated in the model. Representation bias arises from a data constellation when certain groups are underrepresented in the data used to learn the model. This can be achieved, for example, by limiting the trotting data to certain regions or ethnic groups, or more generally, by limiting the data to an unrepresentative target population. Measurement bias can arise from the use of a characteristic that is intended to be a proxy and may be too simple to measure the true target variable, but only part of it. Another possibility is that the measurement of the characteristic is not uniform across all groups in the population. Ultimately, measurement accuracy may be inconsistent across different population groups. Another source of bias can

be aggregation bias, where groups are assembled that do not actually belong with one another. Learning bias occurs when one of the hyperparameters of the ML model is preferred over another parameter, thereby negatively affecting the former parameter. Suresh and Guttag (2019) propose a framework in which a test is carried out from data generation at each stage of the model, using an ideal result and the real result. In the case of deviations, a bias is present, which must then be eliminated from the previous stage.

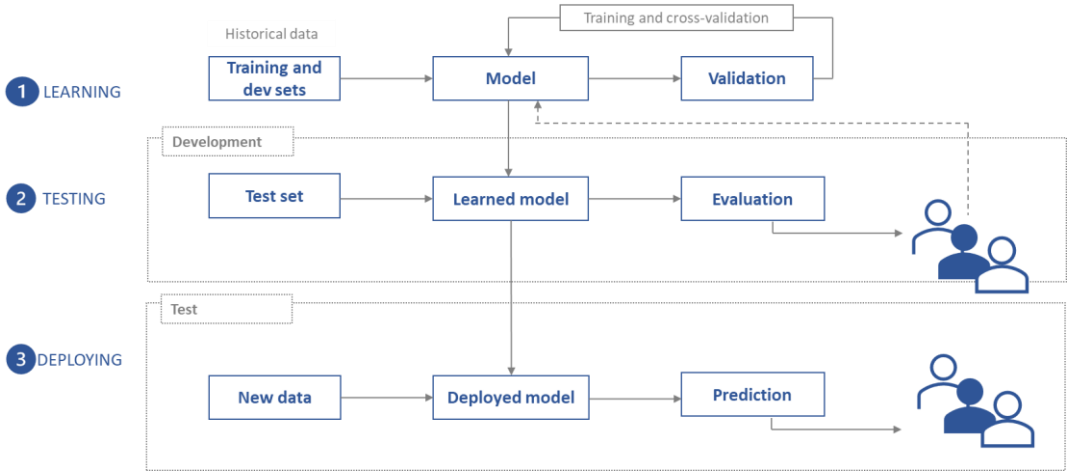

Figure 88: Architetcure of a typical ML system

A typical machine learning system can be split into learning, testing, and deploying (s. figure 88). In the learning phase, historical data is divided into training and development or validation sets. The model is then trained using the training data and authenticated with the validation part of the training data set. The learned model is used on a test data set in the testing phase, and the user evaluates the results. Finally, in the deploying phase, the tested model is deployed, used on new data, and used for prediction.

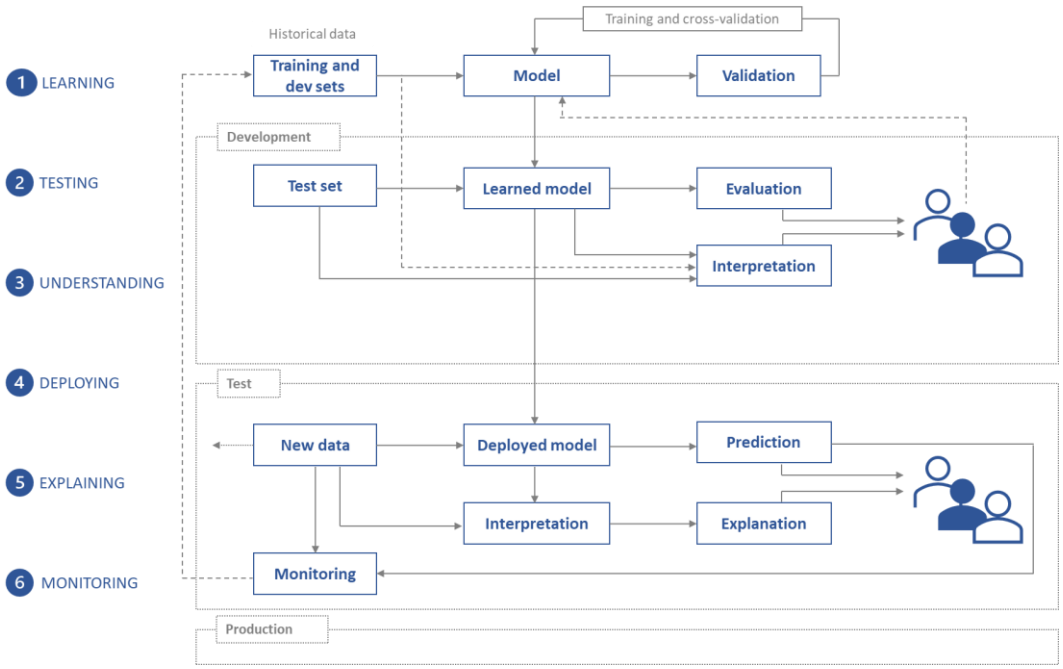

Figure 89: Architecture of a reliable ML system

Figure 89 is an example of how to build a reliable ML system by adding additional layers to the architecture. These layers are shown as understanding, explaining, and monitoring. In the understanding phase, the testers and developers want to recognise why the model chose a specific classification and made a specific prediction. In the explaining phase, the interpretation part is augmented using a human-readable explanation to expose the prediction and explanation to non-expert stakeholders. The monitoring phase is used to permanently monitor the accuracy and explainability of the model in production. The additional two layers are necessary, and not only because of the explanation of the model's prediction. Common challenges are occurring with the use of machine learning models.

The “interpretable” and “explainable” components are processed in the Re_fish model, in the area of the non-symbolic module, by means of the library of XAI methods and corresponding results from the symbolic module on the basis of the results of the model and the context data used, for example, to learn the AI model.

5.2.9 Re_fish Opportunities and Solutions

Since Re_fish is a reference architecture, phases E to H are not dealt with in detail in this paper. They become important in an instantiation of the reference architecture, for example in the development of a prototype. In phase E of ADM, "options and solutions", the options

for the (software) architecture are evaluated to see if they can fulfil the requirements. In phase F, "migration planning", the transition from actual data, if available, to the target architecture takes place. A prioritisation must be created, which also takes into account the current software project and thus any existing side effects. The creation of a roadmap for the implementation and migration of the software architecture concludes this phase. In phase G, "Governance Implementation", governance mechanisms are established to ensure compliance with the standards and guidelines for the software architecture. This also involves the requirements that have already been raised, e.g., compliance with the GDPR, etc., and also the monitoring and control of the software development process, the verification of compliance with the architecture and, if necessary, the iterative making and implementation of necessary adjustments (e.g., the VDE standard can be used for this). Phase H "Establishment of architecture change management" concludes the ADM cycle. Essentially, this phase is about implementing architecture monitoring, using processes to manage and control changes to the software architecture, as well as performing impact analyses for proposed changes and evaluating them in terms of their compliance with the software architecture vision, i.e., ultimately evaluating the architecture. In addition, the integrity and consistency of the software architecture over time should be ensured. Parts of the implementation of these requirements have been presented in chapter 5.2.8 Lifecycle Management. It must be ensured over the entire lifetime of the AI model that no biases arise and that the reliability and integrity and fulfilment of compliance requirements are ensured.

5.3 Evaluation of the Re_fish Architecture- Design Science Evaluation and Expert Survey

The thesis is based on the design science paradigm, which as a problem-solving paradigm, consecutively, is based on engineering and the sciences of the artificial (Simon 1996). The aim of design science is to create such innovations, consisting of ideas, practices, technical skills and new or by combining existing products, analysis, design, implementation, management, and use of information systems can be carried out more effectively and efficiently (Denning, 1997; Tsichritzis, 1998). The design process is a sequence of activities by experts that produce an innovative product, e.g., in the form of an artefact. The evaluation process of the artifact provides feedback information and a better understanding of the problem, which in turn positively influences the design process in improving the artefact (Hevner et al., 2004).

The evaluation of the Re_fish reference architecture was divided into two parts. In the first one, the architecture was evaluated with regard to the criteria of the seven criteria of design science research presented in Chapter 1.5. In the second step, the architecture was evaluated through a survey with experts.

It is important to note that the evaluation was not about assessing the quality of the design of the reference architecture, but about creating a projection of the system quality in terms of the effects that could be achieved by the architecture, when it will be implemented. (Bass et al., 2021; Vasconcelos et al., 2005).

It is therefore necessary to assess whether the qualitative requirements and the constraints identified in the previous chapters (Chapters 2, 3 and 4) have been implemented in the reference architecture (Bass et al., 2021, Vasconcelos, et al., 2005). As described in chapter 4, the reference architecture represents a reference model for a set of (specific, instantiated) architectures. The evaluation can be used to ensure that an instantiated architecture meets the necessary criteria.

The approach to evaluate the reference architecture were therefore as follows:

1. Evaluated the seven guidelines of Hevner et al. 2004
2. Conducted presentation, discussion and survey

Step 1: Part one of the evaluation of the reference architecture Re_fish

In this step, the reference architecture was validated against Hevner's and Design Science research guidelines listed in 1.5 Research theory and design (see Chapter 1.5).

Guideline 1: Design as an Artefact: The reference architecture as a purposeful IT artefact, addressing a fundamental organisational problem: the design, construction, and running of a trustworthy AI system.
The reference architecture Re_fish was designed and created in chapter 5 as an artefact (s. Re_fish business architecture, Re_fish information system architecture and Re_fish technology architecture). The architecture was built by following best practices for design using a combination of ADD and ADM. Therefore, guideline 1 is fulfilled.

Guideline 2: Problem Relevance: The relevance of the business problem is derived from empirical analysis, e.g., that of existing literature and empirical studies. This can be seen

as an unsolved business problem. In this work, insufficient explainability of AI models (or the lack of explainability) in corporate planning comprises such a problem.
The empirical necessity of creating a reference architecture for an XAI system has been sufficiently demonstrated in chapter 1 and chapter 2; it follows that guideline 2 is fulfilled.

Guideline 3: Design Evaluation: The design artefact utility and its efficacy must be assessed using rigorous evaluation (The evaluation can be done in terms of functionality, completeness, consistency, etc. s. (Hevner et al, 2004)). As presented in Chapter 1, the methods proposed by Hevner et al. (2004) are designed to evaluate scientific research rigorously. In this way, the evaluation process provides feedback on the overall design process as well as on the resulting artefact. Because the de-design process is iterative, the quality of the process and the artefact itself is improved. By incorporating feedback into the design process, the next iteration of the artefact will benefit greatly in terms of quality. It is essential to consider all feedback to create the best possible outcome. For this reason, the gaps identified in the evaluation of Re_Fish have been documented and taken into account in the next iteration. The test topics proposed by Hevner et al (2004) are described in chapter 1. The evaluation by the experts can be seen as a so-called informed argument. This evaluation can certainly change in the next iteration and then be, for example, the field evaluation of a prototype. In further iteration steps, the reference architecture must now be further deepened and extended. The result of a further iteration step can be the creation of a prototype. Guideline 3 is thus also fulfilled. (s. below)

Guideline 4: Research Contributions: The research used existing foundations and proven methodologies to provide a verifiable contribution to the design of artefacts, design foundation (e.g., reference architecture) and design methodologies (the evaluation), and the artefact itself. The artefact and its design methodology will be used as a starting point for further iterations. All findings have been documented for further analysis and future research (Hevner et al., 2004, s. Chapter 1.3). Guideline 4 is fulfilled.

Guideline 5: Research Rigour: The work of the thesis was built upon applying rigorous methods in the construction, evaluation, and design of the artefact. In this work, the well-researched area of reference modelling as a foundation for artefact construction has been implemented. The evaluation was done by testing the artefact – gaining expert opinions

and thoroughly gathering valid arguments concerning the utility of the reference architecture. Guideline 5 is fulfilled.

Guideline 6: Design as a Research Process: The artefact utilises available means to reach desired ends and satisfies laws in the problem space (environment). However, design science is an inherently iterative process; therefore, this work can be seen as a starting point to search for the best and optimal solution for a reference architecture in order to build reliable, sustainable explainable AI systems. Therefore, it can be seen as a satisfactory solution – satisficing – without specifying all of the possible solutions (It can be seen as a "starting point" and can help to further investigate and contribute to further research – Simon, (1996)). Guideline 6 is fulfilled.

Guideline 7: Communication of Research: The artefact with respect to the research outcome of this dissertation was effectively presented to both audiences – those who were technology-oriented (with sufficient detail to enable construction and implementation of the artefact) and business-oriented (to enable them to use the artefact in a specific organisational context) (Hevner et al., 2004). Guideline 7 is fulfilled

Step 2: Part two of the evaluation of the reference architecture Re_fish

The presentation was prepared based on the display that was later included in the survey. All selected experts were invited to the presentation meeting. All results and presentations were explained in depth and the questions were answered. It was requested that the discussion points in the subsequent survey be entered in the comments field so that they could be evaluated afterwards and included as requirements in the reference architecture.

The survey was conducted using the method described by Saunders et al. (2023) and Sekaran and Bougie (2019) of the selected experts, 11 participated in the survey. The questionnaire and the survey questions are presented in Appendix B. The survey was conducted online. The names of the participants are known to the author but are not disclosed in this paper for data protection reasons. For the questionnaire, the application Microsoft Forms was used.

When conducting the survey and creating the questionnaire, many requirements were taken into account. First of all, it is important that the respondents have sufficient motivation to

answer the questionnaires. If the question is understood, the person retrieves information from his or her memory. In addition to answering the questions using a five-point Likert scale, the respondents could also justify their decision using a comment field. Basically, according to Hollenberg (2016), the following questions must be answered with "yes" for the aspect's motivation formation, understanding, memory retrieval, judgement formation, consideration, decision, and communication:

- Were the respondents able to assign value to the questionnaire based on the subject matter and its design? - Yes, this was ensured by also having a presentation before the interview and also explaining the questions.
- Were the respondents able to assign a neutral or positive consequence to the answer to the questionnaire? - Yes, the respondents were able to assign neutral/positive consequences to the questionnaire and the information provided, also due to the presentation.
- Was the effort for the respondents acceptable? - Yes, this was checked in the pre-test.
- Was the questionnaire structured in a comprehensible way? - Yes, the questionnaire was also discussed during the presentation, and help was offered while answering the questions.
- Yes, since it was a questionnaire from experts, it was ensured that the participants were professionally in contact with these topics or were working on them.
- Were the respondents able to answer the questions competently? - Yes, see previous question and results.
- Was it easy for the respondents to select and develop clear answers? - Yes, the respondents were able to select the questions using a five-point Likert scale and also justify their results in a free comment field.
- Was it possible to answer the questions freely and unbiased? - Yes, as in the previous question, both the Likert scale and a comment field were available as answer options.
- Was it possible to indicate the decisions made when answering the questions in the questionnaire? - Yes, see above.

The questionnaire was prepared for the evaluation of the architecture by experts. According to Hollenberg (2016), a questionnaire must be able to answer the following questions with "yes" with regard to the quality criteria validity, reliability, objectivity, representativeness, utility, economy and reasonableness:

- Is the questionnaire constructed in such a way that it really measures what it intends to measure? - Yes, the questionnaire was designed according to the evaluation criteria for architecture (see below).
- Did the respondents answer the questions largely independently of the person conducting the survey? - Yes, after the questionnaire was presented in a presentation, the respondents answered the questions independently. Help, if needed, was offered by the interviewer.
- Can the results of the questionnaire be generalised - with regard to the target group? - Yes, a small but experienced sample of experts was formed in order to obtain the highest possible quality feedback in the first iteration of the creation of the target architecture.
- Is there any form of benefit to the target group from the survey? - Yes, the specific benefit is the first high quality feedback for the first iteration of the reference architecture.
- Has the questionnaire been constructed in such a way that it is long enough to cover all relevant aspects, but as short as possible so as not to overburden the respondent? - Yes, this was checked in a pre-test, among other things.
- Was the questionnaire reasonable for all respondents? - Yes, this was also checked in a pre-test.
- The questions were mainly closed questions using the Likert scale (see above). A comment field was available to the respondents. The following criteria should apply to the items (question-answer combinations):
- Could the question be applied meaningfully? - Yes, this was ensured during the pre-test and also during the presentation.
- Was the targeted knowledge area targeted with sufficient precision? - Yes, this is shown not least by the answers.

- Was sufficient information provided? - Yes, within the framework of a presentation with all the required content.
- Are answers suggested or implied to the respondent? - No
- Were the question and the answer options formulated with sufficient precision? Yes, this was confirmed by the results, among others.
- Was the content of the question/answer combination free of contradictions? -Yes, this was also confirmed by the results.
- Could the question be answered exhaustively with the choices? - Yes, this was also made possible by the provision of a comment field.
- Were the questions simple enough not to have a negative influence on the motivation of the respondents? - Yes, this was checked during the pre-test and also confirmed by the results.

A five-point Likert scale was used; an interval scale level is often assumed for the evaluation (Franzen, 2019; Häder, 2019; Schnell et al., 2022). Here, positively, or negatively formulated statements about an issue are given. The respondents can then express their opinion by agreeing or disagreeing with the statements in several predefined gradations. The distance between the answer options is as equal as possible and interpretable (equidistance, Bortz & Döring (2023)). The questions are given a five-point answer scale and combined in a sample. Then the answers are added up to a total value. Only the items with the highest correlation value are then included in the final questionnaire, which is then used in the study sample Schnell et al. (2022) or Bortz & Döring (2023).

During data preparation, missing values were replaced with the median.

The experts' evaluation of the architecture was based on the following criteria (s. Bass et al, 2021):

- Ease of use – the user's ability to use a system effectively. -> Usability
- Performance – the responsiveness of the system - the time it takes to respond to stimuli, or the number of events processed in a given time interval. -Y Performance
- Reliability – the ability of the system to function over time. -> Reliability

- Availability – the proportion of time the system is operational. -> in this iteration N/A
- Security – the ability of the system to resist unauthorised attempts to use it and denial of service, while providing its services to legitimate users. -> Security
- Functionality – the ability of the system to perform the tasks for which it is intended. -> Functionality
- Modifiability – the ability to make changes to a system quickly and inexpensively. - > Modifiability
- Cost – the cost of the system -> in this iteration within Modifiability

A 5-point Likert scale was used to answer the questions:

Weighting scale with

1 = Does not apply at all, strongly disagree

2 = Does not apply, disagree

3 = Neutral (or neither agree nor disagree)

4 = Agree and

5 = Strongly agree

Among the experts of the participants were the following persons (the names will not be disclosed in this thesis):

- Senior Economist, Strategic Planning, PKN Orlen (Polski Koncern Naftowy ORLEN)

- Chief Architect, SAP AG

- Principal Expert, SAP AG, Integrated Business Planning

- Principal Project Manager, SAP AG, Utilities & Process Industry

- HR Manager, NN, selection of personnel, in particular requirements profile regarding AI for employees

- Expert Data Analyst, NN, data collection and data analysis

- Technology Lead & Business Advisor, SAP ICN Potsdam, SAP Decision Simulator (https://www.sap.cn/about/company/innovation/icn.html)

Data Analysis

1. Descriptive

1.1 Items

Descriptive Statistics

Descriptive Statistics

	Item 1	Item 2	Item 3	Item 4	Item 5	Item 6	Item 7	Item 8	Item 9	Item 10	Item 11	Item 12	Item 13	Item 14	Item 15	Item 16	Item 17	Item 18
Valid	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12	12
Missing	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Mean	3.833	4.000	3.917	3.667	4.250	4.083	3.333	3.667	4.333	4.250	4.333	4.167	4.167	4.000	4.167	4.000	4.083	3.750
Std. Deviation	1.267	0.739	1.084	0.888	0.452	0.515	1.723	1.155	0.651	0.754	0.651	0.718	0.718	0.426	0.577	0.603	0.669	0.866
Minimum	0.000	2.000	1.000	2.000	4.000	3.000	0.000	0.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000	3.000	2.000
Maximum	5.000	5.000	5.000	5.000	5.000	5.000	5.000	4.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000	5.000

Table 26: Descriptive statistics of the survey items

The statistical analysis was performed using the JASP[39] program based on R. After cleaning the data, and the Likert scale values were transformed into numerical values following the provided guidelines. The sample size was N =12 values. Table 26 shows the descriptive statistics of the items (items 1 to 18) with their respective mean, standard deviation, and the respective minimum and maximum.

Results

Unidimensional Reliability

Frequentist Scale Reliability Statistics

Estimate	Cronbach's α
Point estimate	0.828
95% CI lower bound	0.599
95% CI upper bound	0.938

Note. The following item correlated negatively with the scale: Item 3. Variables Item 9 and Item 11 correlated perfectly.

Table 27: Reliability testing of the items

[39] https://jasp-stats.org/ , accessed 18.06.2023

Table 27 shows the calculation of the reliability and results in a high value for Cronbach's alpha. This means that there is a very high level of agreement between the items, which thus evaluate the individual aspects of the reference architecture.

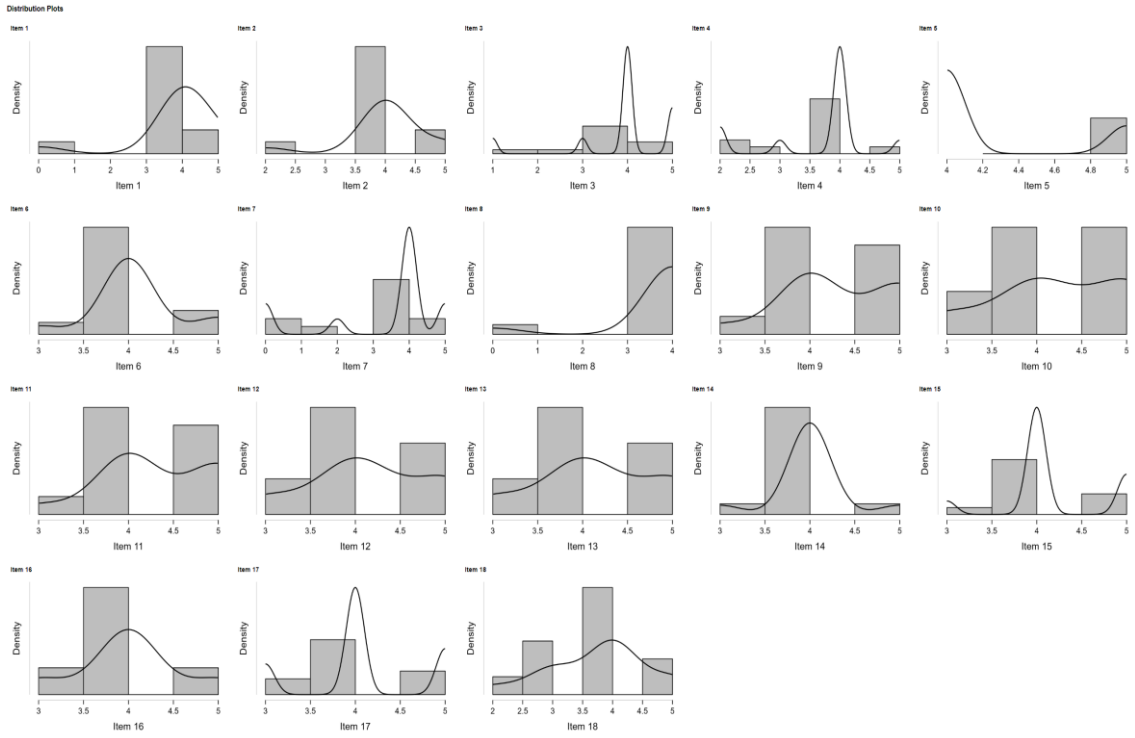

Table 28: Distribution of the transformed Likert values for all items

Table 28 shows that most of the item scores are 4 or above, which also reflects the value of the Cronbach's alpha. This means that most items are rated at least "agree".

1.2 Years of experience

Descriptive Statistics

Descriptive Statistics

	Planning Years	IT Architecture Years	Business Analytics Years	AI years	Project Mgmt. Years
Valid	12	12	12	12	12
Missing	0	0	0	0	0
Mean	9.417	8.750	5.833	4.750	11.417
Std. Deviation	8.218	8.433	6.590	7.098	8.969
Minimum	0.000	0.000	0.000	0.000	1.000
Maximum	23.000	25.000	21.000	25.000	25.000

Table 29: Descriptive statistics of the years of experience of the experts

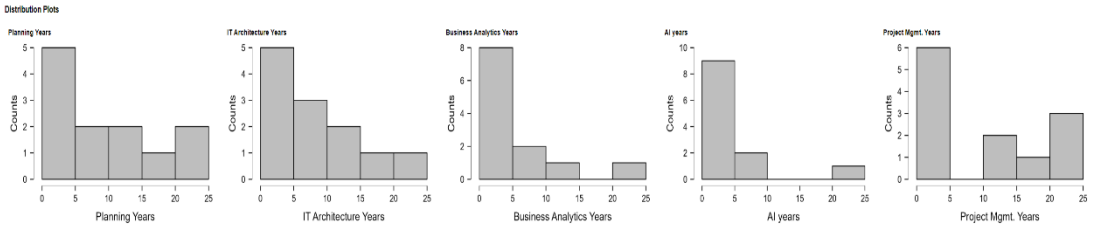

Table 30: Distribution of domain experience

Descriptive Statistics

Descriptive Statistics

	Planning Years	IT Architecture Years	AI years	EVAL
Valid	12	12	12	12
Missing	0	0	0	0
Median	9.000	8.500	2.000	4.083
Mean	9.417	8.750	4.750	4.000
Std. Deviation	8.218	8.433	7.098	0.436
MAD robust	10.378	10.378	2.965	0.412
Shapiro-Wilk	0.895	0.899	0.691	0.931
P-value of Shapiro-Wilk	0.135	0.153	< .001	0.388
Minimum	0.000	0.000	0.000	3.111
Maximum	23.000	25.000	25.000	4.556

Table 31: Descriptive statistics of the years of experience of the experts

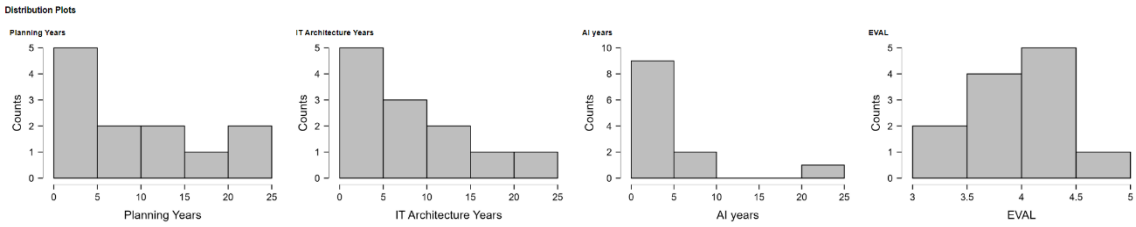

Table 32: Descriptive statistics of the years of experience of the experts and the EVAL variable

EVAL is the arithmetic mean, i.e. the sum of all items divided by the number of respondents.

2. Inferential Statistics

Hypothesis:

2.1 Correlation of AI experience and evaluation of Re_fish

It was pointed out above that the interviewees were experts in their respective fields (project management, AI, business analytics and IT architecture were identified). In addition, the evaluation of the different aspects of the Re_fish reference architecture was presented above as an objective. In the 1980s, Stuart and Hubert Dreyfus (1980) established a multi-level model of what it means to be an expert in a field (critique of the five-level model of Gobet & Chassy (2008 and 2009)). Ericsson et al. (1993) state in their study that it takes about 10,000hrs deliberate training, so approx. 10 years to become an expert, however, depending on the skill to be learned and quality of education etc. This view is, therefore, not accepted here, as the mean is 5 years even for the "youngest" area, AI, all respondents had more than 10 years of professional experience.

H0: The less experience in AI a responder has, the less positive Re_fish will be evaluated

H1: The more experience in AI a responder has, the more positive Re_fish will be evaluated.

Correlation

Correlation Table

Variable		EVAL	AI years
1. EVAL	Pearson's r	—	
	p-value	—	
	Spearman's rho	—	
	p-value	—	
2. AI years	Pearson's r	0.338	—
	p-value	0.283	—
	Spearman's rho	0.434	—
	p-value	0.158	—

Table 33: Spearman’s rho (and Pearson’s r) AI years of Experience and EVAL

Result- Spearman’s rho is 0.434 and with | rho | > .3 shows therefore a moderate positive correlation.

This means that the null hypothesis is to be rejected and the H1 hypothesis is to be accepted.

2.2 Correlation of IT architecture (ITA) experience and positive evaluation of Re_fish

H0: The less experience in ITA a responder has, the less positive Re_fish will be evaluated

H1: The more experience in ITA a responder has, the more positive Re_fish will be evaluated.

Correlation

Correlation Table

Variable		EVAL	IT Architecture Years
1. EVAL	Pearson's r	—	
	p-value	—	
	Spearman's rho	—	
	p-value	—	
2. IT Architecture Years	Pearson's r	0.441	—
	p-value	0.152	—
	Spearman's rho	0.389	—
	p-value	0.211	—

Table 34: Spearman's rho (and Pearson's r) ITA years of Experience and EVAL

Result- Spearman's rho is 0.389 and with | rho | > 3 shows, therefore, also a moderate positive correlation.

This means that the null hypothesis is to be rejected and the H1 hypothesis is to be accepted.

Both results express that the experts value the reference architecture as being "useful" regards to the evaluated aspects.

In the following, some evaluations are presented on the basis of MS Excel.

Question		Strongly Agree	Agree	Neither Agree nor Disagree	Disagree	Strongly Disagree	Neutral	Sum	Result	
									Agree	Disagree
1	Absolute	2	8	1				11		
	Percentage	18%	73%	9%	0%	0%	0%	100%	91%	0%
2	Absolute	2	8		1			11		
	Percentage	18%	73%	0%	9%	0%	0%	100%	91%	9%
3	Absolute	1	7	1	1	1		11		
	Percentage	9%	64%	9%	9%	9%	0%	100%	73%	18%
4	Absolute	1	7		2		1	11		
	Percentage	9%	64%	0%	18%	0%	9%	100%	73%	18%
5	Absolute	3	8					11		
	Percentage	27%	73%	0%	0%	0%	0%	100%	100%	0%
6	Absolute	2	8				1	11		
	Percentage	18%	73%	0%	0%	0%	9%	100%	91%	0%
7	Absolute	2	6		1		2	11		
	Percentage	18%	55%	0%	9%	0%	18%	100%	73%	9%
8	Absolute	10					1	11		
	Percentage	91%	0%	0%	0%	0%	9%	100%	91%	0%
9	Absolute	4	6				1	11		
	Percentage	36%	55%	0%	0%	0%	9%	100%	91%	0%
10	Absolute	4	5				2	11		
	Percentage	36%	45%	0%	0%	0%	18%	100%	82%	0%
11	Absolute	4	6				1	11		
	Percentage	36%	55%	0%	0%	0%	9%	100%	91%	0%
12	Absolute	3	6				2	11		
	Percentage	27%	55%	0%	0%	0%	18%	100%	82%	0%
13	Absolute	4	5				2	11		
	Percentage	36%	45%	0%	0%	0%	18%	100%	82%	0%
14	Absolute	1	9				1	11		
	Percentage	9%	82%	0%	0%	0%	9%	100%	91%	0%
15	Absolute	3	7				1	11		
	Percentage	27%	64%	0%	0%	0%	9%	100%	91%	0%
16	Absolute	2	7				2	11		
	Percentage	18%	64%	0%	0%	0%	18%	100%	82%	0%
17	Absolute	2	7				2	11		
	Percentage	18%	64%	0%	0%	0%	18%	100%	82%	0%
18	Absolute	2	5		1		3	11		
	Percentage	18%	45%	0%	9%	0%	27%	100%	64%	9%

Table 35: Evaluation of the Re_fish reference architecture in percentages

Question	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Strongly Agree	18%	18%	9%	9%	27%	18%	18%	91%	36%	36%	36%	27%	36%	9%	27%	18%	18%	18%
Agree	73%	73%	64%	64%	73%	73%	55%		55%	45%	55%	55%	45%	82%	64%	64%	64%	45%
Neither Agree nor Disagree	9%		9%															
Disagree		9%	9%	18%			9%											9%
Strongly Disagree			9%															
Neutral				9%		9%	18%	9%	9%	18%	9%	18%	18%	9%	9%	18%	18%	27%
Summe	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%

Table 36: Evaluation of the Re_fish reference architecture in percentages

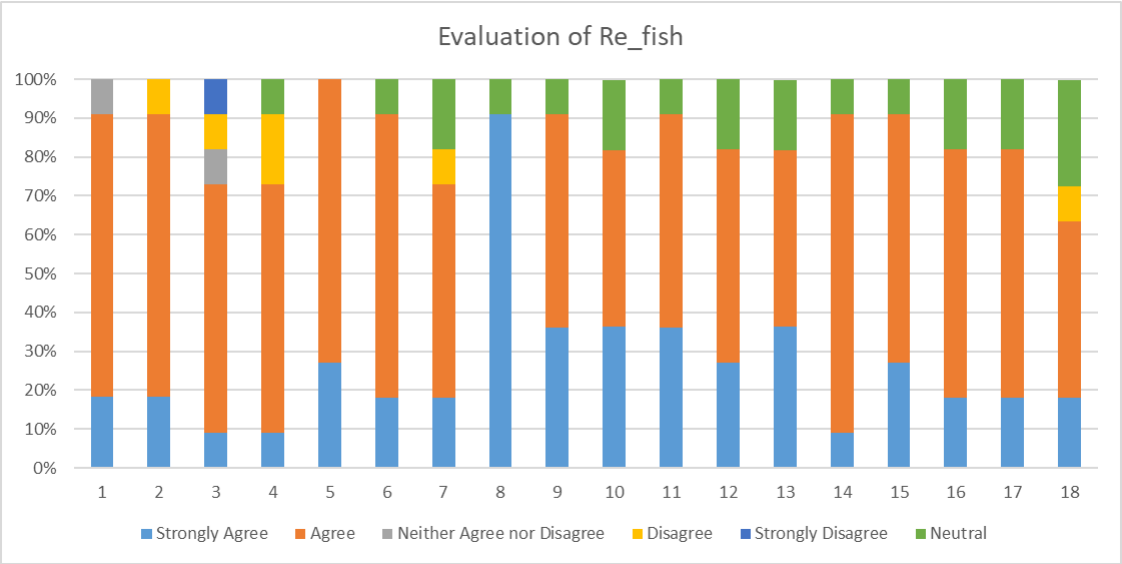

Table 37: Evaluation of the Re_fish reference architecture in percentages categories per question

A total of 15 people took part in the survey by the deadline - N = 15

After correcting incorrect and/or duplicate entries, a total of 12 usable results remained. These results are shown in tables 38 to 40.

Overall, the assessment of the Re_fish reference architecture was positive. If the ratings for "Strongly Agree" and "Agree" per question were cumulated and compared with the opposite ratings "Disagree" and "Strongly Disagree", the Re_fish reference architecture was rated positively for all 18 items asked.

Design Evaluation Methods	
1. Observational	Case Study: Study Artifact in depth in business environment
	Field Study: Monitor use of artifact in multiple projects
2. Analytical	Static Analysis: Examine structure of artifact for static qualities (e.g., complexity)
	Architecture Analysis: Study fit of artifact into technical IS architecture
	Optimization: Demonstrate inherent optimal properties of artifacts or provide optimality bounds on artifact behavior
	Dynamic Analysis: Study artifact in use for daynamic quallities (e.g., performance)
3. Experimental	Controlled Experiment: Study artifact in controlled environment for qualities (e.g., usability)
	Simulation - Execute artifact with artifical data
4. Testing	Functional (Black Box) Testing: Execute artifact interfaces to discover failures and identify defects
	Structural (White Box) Testing: Perform coverage testing of some metric (e.g., execution paths) in the artifact implementation
5. Descriptive	Informed Argument: Use information from the knowledge base (e.g., relevant research) to build a convincing argument for the artifact's utility
	Scenarios: Construct detailed scenarios around the artifact to demonstrate ist utility

Table 38: Design Evaluation Methods by Hevner et al. (2004)

As mentioned above within the description of guideline 4, Hevner et al. (2004) provide twelve design evaluation methods. These can in turn be divided into five categories. These

include the methods shown in table 42. In the author's opinion, this dissertation meets the guidelines of Hevner et al. (2004), as shown above. The presentation, the discussion/interview and the subsequent survey were conducted in accordance with the evaluation method of an "informed argument". In an informed argument, a convincing argument has been built up on the basis of the knowledge base (e.g., through relevant research) to prove the utility of the artifact. In a subsequent iteration, in which the gaps identified by the experts (s. chapter 5.4) will be incorporated into the iterated design, a prototype will be built, which will then undergo further evaluation in the context of use in a concrete context in the field of corporate planning. In conclusion, however, it can already be stated that the Re_fish reference architecture represents a solution to the problem presented in the hypothesis.

5.4 Adjustment of the Reference Architecture Re-fish

In the following, the significant results and the comments and recommendations of the participants in the discussions and the survey are presented. The additional requirements are included in a catalogue and numbered consecutively. They are given the ID AR = Adjusting Requirement. An overview of the survey questionnaire and the questions can be found in Appendix B- "Survey Questions for Architecture Evaluation". The names of the questions in the survey are based on the four areas to be evaluated, see Chapter 5.3 - Usability = U, R = Reliability, F = Functionality, M = Modifiability, Performance = P, Security = S, Quality = Q.

Introduction:

Reference architecture quality attribute 1: Usability

Artificial intelligence needs transparency and security. Especially in corporate planning processes, the explainability of AI is an essential factor in creating trust in AI applications. The reproducibility of the results of multi-layer machine learning processes is a prerequisite for this.

I like the explanations about Re_fish from page 7 and 8 the most. Interesting conception. The potential users would surely need a support while implementing it in their businesses

U1:

In my opinion, it is not "easy" to understand the Reference Model, but this is not due to the presentation of the model but rather to the complexity of the problem and its solution.

Finding and recommendation:

In the next iteration, the views of the architecture should be simplified. This also results from the fact that in the next iteration the architecture components will have to be further decomposed. -> AR 1

U5:

I like the possibility to use a chat-bot to deepen certain questions and to point out - possibly overlooked by the user - connections in the dashboard, for very helpful and useful.

I like slides 17 and 18 because they explain the process and users' profits in an easy way. The possibility to ask further questions is very important.

Finding and recommendation:

N/A -> AR N/A

U6:

I believe it is useful and important to provide (as much as possible) causal explanations (always assuming that there is also a clear cause-effect relationship between two variables). The chosen presentation in the form of the "causality ladder" seems to be suitable for this.

The implied function that the presentation dynamically adapts to further questions in the course of the analysis is very useful. Whether this type of interface is sufficient for all types of users would have to be determined by corresponding analyses in the field and practical use.

In my opinion, it would also be good if it is made transparent for the user with which probabilities or with which certainty one or the other statement is made, or causality is established.

Finding and recommendation:

The experts consider the causality of the decisions to be a very important function of the reference architecture. It should therefore be developed in detail in the next iteration step. -> AR 2

U7:

The explanations of the machine learning models and the presentation of the results of the explanatory model used in each case is useful and sufficient for certain user groups. Users with less mathematical understanding (often decision makers, managers) might be put off by this rather "technical" way of presentation. Here, one could think about how to present the information provided by the explanatory models in a suitable way and in regard of specific actors. In any case, it should be considered to offer a kind of "consolidated view" across all explanatory models in order to avoid uncertainties regarding the results and deviations between the respective models.

Finding and recommendation:

It was already pointed out during the development of Re_fish that different users have different skills (models of mind) and the HCI must adapt accordingly. For this reason, user-specific dashboards were initially included in the design. However, in a next iteration of the Re_fish design, it must be developed more intensively so that the different users, their skills, their model of mind and also their requirements in general are better supported. ->AR 3

P1:

In any case, this is an interesting and promising approach.

P3:

The idea of the work to combine the interaction of symbolic and non-symbolic AI with a representation of data provenance is very promising. Whether this is sufficient to establish the desired confidence in the results must be proven by tests in practice

Finding and recommendation:

In this case, one expert agrees with the approach in the thesis and supports the continuation of the development of the Re_fish reference architecture by developing a prototype and thus shares the opinion expressed here in this thesis.

S1:

In principle, the solution architecture allows different user roles to be distinguished. However, these user groups are - in my opinion - not distinguished clearly enough in the illustrations.

Finding and recommendation:

In the next iteration of the development of the Re_fish Reference architecture, the assignment of users (stakeholders) to the respective groups and roles must be more strongly and better represented. -> AR 4

S2:

Especially for planning processes and decisions made on the basis of corresponding analyses, it is necessary that the results of the analyses are documented and reproducible. This means that the conditions under which the results were obtained must be historicized, etc. The solution architecture provides appropriate functions for this purpose.

Q2:

I am convinced that the necessary technologies and concepts for a suitable and functioning implementation of the reference architecture are available today.

Q3:

Naturally, I have a hard time with this answer - because of my many years of project experience! But what speaks for the calculability and predictability of project costs is that many functional building blocks of the reference architecture are already available today and "only" need to be adapted.

Further findings (s. table 42)

Adjusting Requirement Catalogue		
#	ID	Description
1	AR	In the next iteration, the views of the architecture should be simplified. This also results from the fact that in the next iteration the architecture components will have to be further decomposed.
2	AR	The experts consider the causality of the decisions to be a very important function of the reference architecture. It should therefore be developed in detail in the next iteration step
3	AR	It was already pointed out during the development of Re_fish that different users have different skills (models of mind) and the HCI must adapt accordingly. For this reason, user-specific dashboards were initially included in the design. However, in a next iteration of the Re_fish design, it must be developed more intensively so that the different users, their skills, their model of mind and also their requirements in general are better supported.
4	AR	In the next iteration of the development of the Re_fish Reference architecture, the assignment of users (stakeholders) to the respective groups and roles must be more strongly and better represented. -> AR 4

Table 39: List of further findings

5.5 Summary

The task of Chapter 5 was to combine the results and theoretical considerations of the previous chapters and use them to create an artefact. For this purpose, the results of the stakeholder analyses and, in particular, their requirements and XAI were collected and summarised via literature review. The method of creating architectures, and principally reference architectures, was also considered. All results were used in Chapters 5.1.2 to 5.2.8 to create the Re_fish reference architecture. Chapter 5.3 described the evaluation of the architecture in the form of a two-step approach, in which the first part relates to the evaluation of the implementation of the design science method, and the second one to the evaluation (first iteration) of the Re_fish reference architecture.

Finding 16: In Chapter 5, the reference architecture of Re_fish was created - the results of the previous chapters were used and taken into account. The result is a reference architecture that serves as a reference model to be used as a starting point in a concrete use case. In Chapter 5.2.8, statements were made about the lifecycle, and it was presented how Re_fish can be integrated into a machine learning pipeline using various modules.

Finding 17: The Re_fish reference architecture was evaluated by experts from the business and technology sectors. The result of the evaluation was positive and thus supported a continuation and next iteration.

The Sun's special nourishment proved as effective for Josie as it had for Beggar Man, and after the dark sky morning, she grew not only stronger, but from a child into an adult. As the seasons – and the years – went by, Mr McBain's vehicles cut down the tall grass in all three fields, leaving them a pale brown color. The barn now looked taller and more sharply outlined, but Mr McBain still didn't build additional walls for it, and on cloudless evenings, as the Sun went towards his resting place, I was still able to see him sinking to the far side of the barn before fading into the ground." (Ishiguro, Kazuo (2021). Klara and the Sun. Chapter 6)

6. Summary and Outlook

The present work is both theoretical and empirical. The hypothesis and the main goal of the dissertation, G1 (s. chapter 1.4), was to develop a reference architecture for trustworthy artificial intelligence in the context of corporate planning in the process industry. This artefact, together with further iterative refinements and additions, will serve as a basis for concrete implementation projects in the future. The created reference architecture is named "Re_fish" (which a composition in honour of Marian Rejewski, the leading Polish scientist who solved the Enigma code and the Babelfish, "a fictitious universal decoder for every form of language in the universe"). The development of this reference architecture followed the research approach of design science research. The empirical relevance of the reference architecture has been developed in this work with scientific rigour in the context of corporate planning in the process industry.

This hypothesis was based on the observed phenomena in literature and practice that decisions and actions taken by an AI model in the context of corporate planning scenarios and decision-making are not always explainable to stakeholders and, therefore, will not be trusted. Since most AI models, especially subsymbolic, are not transparent, interpretable, or explainable, users do not trust their outcomes. As a result, their potential is not fully realised (the difference between interpretability/explainability and explanation depends on the situation in which the model is used).

In this dissertation, it has been assumed that stakeholders – users need a user presentation of the results of the AI methods to understand the decisions or actions taken by e.g., subsymbolic "black box" machine learning and deep learning models. This is especially relevant for managers and decision makers.

In the design science research approach, an empirical observation, or a resulting problem out of such an observation leads to changing this situation by providing a solution. The development of such a solution is done through the scientifically rigorous use of the existing knowledge base, i.e., the foundations, such as theories, frameworks, tools, etc., and the methodologies, such as data analysis techniques, formalisms, etc.

In addition to the above-mentioned main goal, the dissertation has covered the following secondary objectives:

G1.1: The dissertation has provided an overview of the actual status and research of the impact of Artificial Intelligence on the economy. This goal has been mapped to Chapter 1.1 and 1.2

G1.2.: The dissertation has provided an overview on the specifics of the process industry, challenges the process industry is facing and how AI can support business in the process industry. This goal has been mapped to Chapter 2

G1.3: The dissertation has provided an overview of the actual status and research in AI and XAI. This goal has been mapped to Chapter 3

G1.4: The dissertation has provided an approach on how to develop a reference architecture for a trustworthy AI (XAI) system. This goal has been mapped to Chapter 4

G1.5: The dissertation has provided a system reference architecture – Re_fish, which can be used by instantiating to build a trustworthy AI- XAI system. This was a direct subgoal of the main goal (repeating it) and mapped to Chapter 5

Further, Chapter 1.1 and Chapter 1.2 introduced the economic background of AI and the research approach and adapted it to the facts at hand. Subsequently, the special features of the process industry were discussed in Chapter 2. The following results have been obtained:

- AI has a significant impact on the economy as it has the possibility to be disruptive transformative. AI changes the role of the traditional production factors, labour and capital.

- Companies in the process industry are characterised by asset-intensive production and thus have a high fixed cost block. Therefore, the demand for utilisation is an essential element in planning and target setting to ensure sufficient ROI. High plant utilisation creates further challenges in terms of maintenance, etc.

- Companies operate in highly complex supply chain networks that are very vulnerable to disruptions, such as pandemics or wars, etc.

- EU27 companies are important to the economies in which they are located. They require a consistent supply of energy due to their production process, which can lead to a production stoppage of days or months in the event of an interruption.

- Chemical and life sciences companies operate on a highly regulated market.

The above objective was covered in Chapters 1.1 and 1.2, in findings 1 to 11. Generally, these results reflect the economic relevance of AI. Therefore, G1.1 and G 1.2 are considered fulfilled.

Explainable Artificial Intelligence was presented in Chapter 3. First, the field of AI was introduced and then the individual areas - subymbolic, symbolic, and neuro-symbolic AI were explained. This showed that current XAI approaches are mainly used to explain machine learning models. Hybrid (combination of symbolic and subymbolic AI) or knowledge-based approaches were also demonstrated, as well as neuro- symbolic AI approaches, discussed as an extended, tested approaches which also enable people to directly understand explanations related to AI.

With these outcomes, G1.3 is considered fulfilled.

In Chapter 4, the methodological foundations for the creation of a reference architecture have been developed. All results and elaborations from Chapters 2, 3 and 4 were then used in Chapter 5 to create a reference architecture for trustworthy AI. The artefact to be created was evaluated through an evaluation of experts. Any gaps have been documented and thus incorporated into the reference architecture for a further iteration. The evaluation result

was positive for this first iteration and the reference architecture Re_fish is seen as a solution for the observed problem and the hypothesis was fulfilled.

With this result, G1 and G1.5 are considered fulfilled.

The research questions mentioned in Chapter 1.4 have been fulfilled. To proof this, see the summary of the findings of this thesis below:

Finding 1: Impact of AI on economy
The impact of AI on economy can be described as the importance of former production factors, labour and capital, become less important, or grow together into a single factor.

Finding 2: Potential growth opportunities through AI:

1. Intelligent automation. With the help of AI, intricate and strenuous physical tasks can now be replaced. – Replacement case.
2. Additionally, virtual work can also be carried out through software agents, which can replace non-physical tasks such as matching outgoing invoices with payments, within the framework of robotic process automation (RPA[40]). – Replacement case.
3. There is also a potential for advancement by building upon existing work, as outlined in this dissertation, which could ultimately exceed human capabilities. - Augmentation case.
4. Another opportunity for growth arises when innovations spread from one area to another, resulting in increased efficiency through the use of AI and leveraging synergies – Raising synergies through diffusion.

Finding 3: Impact on Labour:
The impact of AI on the labour market is not viewed uniformly. There are different opinions about the strength and direction of the impact. However, the impact can be differentiated according to the growth drivers outlined in Finding 2. For example, some work will fall under the so-called replacement case, others under the augmentation case, and new work will be created, for example, through the diffusion of innovation into other areas.

[40] Robotic Process Automation, s. e.g., https://www.sap.com/germany/products/technology-platform/process-automation/what-is-rpa.html (accessed on 18.06.2023)

Finding 4: Front runners participate most. This could lead to “supercompanies”.

Finding 5: XAI can help to overcome barriers against AI- XAI can be also a needed requirement for specific industries to use AI (s. regulations in process industry- s. chapter 2)

Finding 6: The impact of AI on the situation of work at the company level, as a competition for the greatest talents and the best skills, is closely linked to the "front runner" benefit most. Because the "front runner" companies will also gain the best talents and skills. As a result, according to studies, companies have the task of training their employees extensively in order to ensure the best possible use of AI in the company.

Finding 7: Process companies have some special economic features. These result from the production process. The industry is very heterogeneous, but in general this production process is not easy to stop and restart, for example. Production is extremely equipment-intensive and requires large investments. The impact on the environment is also relevant in terms of sustainability and climate protection. Production itself is less labour-dependent than discrete manufacturing. Companies in the research-based life sciences have a complex, extensive and extremely expensive research process that is subject to many regulations - AI could bring significant improvements here, on the one hand in economic terms, but also in terms of curing generally still incurable diseases.

Finding 8: Competition in the process industry sector is very high and has led to continued concentration over the last 30-40 years. Globally, there are currently only three countries (or groups of countries) that achieve significant sales volumes - these are the USA, the EU and, far ahead of the two aforementioned, China.

Finding 9: Key trends in the process industry are digitalisation, sustainability, including in complex and networked supply chains, and further process optimisation. This industry is highly automated due to its production process, but experts suspect that the available data is not yet being used extensively for process optimisation.

Finding 10: Challenges in the process industry ergeben sich, wie bereits oben beschrieben, aus der hoc The challenges in the process industry arise from various aspects. On the one hand, there is the high level of regulation, the fierce competition, which is also reflected in

the increased concentration that has taken place since the 1970s. The search for qualified workers severely restricts the search for locations. There are also challenges posed by the enormous energy requirements and extremely high plant costs, which also have to be maintained over the long term. On the other hand, there are the short time intervals in which, for example in the pharmaceutical industry, sales can be made that cover the development costs.

Finding 11: The use of XAI in companies in the process industry naturally depends on the use of AI in the companies. Potential applications have been identified in the areas of scenario planning, sales and operation planning, e.g., forecasting, process control, etc., which, when considering the use of AI in the area of research and development as well as in automated process control, have a significant - positive economic impact in the sense of the economic growth drivers presented in Chapter 1.1. and Chapter 1.2 respectively.

Finding 12: In Chapter 2.3, the corporate planning process of the process companies was presented. In particular, scenario planning, which is to be classified in the strategic planning area, and sales and operations (integrated business planning) planning, which is to be classified in the tactical area. These sub-planning processes have several possibilities to replace or at least support sub-processes with AI solutions. First and foremost forecasting, but also optimisation with regard to constraints - usually linear optimisation models are traditionally used here, but AI methods are already available. The identified stakeholders and their requirements will be taken into account in the requirements for the reference architecture.

Finding 13: Chapter 3.4 briefly presents the ethical, legal and regulatory requirements for Explainable AI. The risks of AI have been recognised and are already subject to regulation in Europe, for example in the area of the EU GDPR, PE-6-2023-INIT, etc.

Finding 14: In Chapter 3.2, the technical perspective of artificial intelligence was presented, after the economic perspective was presented in Chapter 1.1 and 1.2. In this chapter, the different areas of AI, machine learning, deep learning, knowledge enabled systems and finally the promising approach of neuro symbolic systems, a combination of deep

learning and symbolic AI, were presented. Then, in Chapter 3.3, the area of XAI was presented.

Finding 15: In Chapter 4, the theoretical possibilities for developing a reference architecture were examined and discussed. For Re_fish, the methodology was based on the TOGAF ADM and the ADD methodology. The whole process of designing and developing a reference architecture was described.

Finding 16: In Chapter 5, the reference architecture of Re_fish was created - the results of the previous chapters were used and taken into account. The result is a reference architecture that serves as a reference model to be used as a starting point in a concrete use case. In Chapter 5.2.8, statements were made about the lifecycle, and it was presented how Re_fish can be integrated into a machine learning pipeline using various modules.

Finding 17: The Re_fish reference architecture was evaluated by experts from the business and technology sectors. The result of the evaluation was positive and thus supported a continuation and next iteration.

RQ1: What are the specifics of the process industry?

RQ 1.1: What are the main and differentiating characteristics of the process industry?
The RQ 1.1 was addressed and answered by finding 1,2,3,4,5, 6, 7, 9, 10 and 11

RQ 1.2: What are the specific market conditions of the process industry?
The RQ 1.2 was addressed and answered by finding 4,8,9 and 10

RQ 1.3: What does the planning process look like within corporate planning?
The RQ 1.3 was addressed and answered by finding 12

RQ 1.4: What special planning sub-processes in corporate planning are of particular importance for the process industry?
The RQ 1.4 was addressed and answered by finding 12

RQ 1.5: What decisions are made in these sub-processes that AI systems can/ will take over?
The RQ 1.5 was addressed and answered by finding 12

RQ 1.6: What are the requirements for explaining decisions made in the sub-processes?
The RQ 1.6 was addressed and answered by finding 12

With this result and the fact that the findings have answered all subsequent research questions, the research question RQ1 is considered as being answered.

RQ 2: What is Explainable AI and how can it support decision making in the corporate planning process?

RQ 2.1: What is AI
The RQ 2.1 was addressed and answered by finding 11,13 and 14

RQ 2.2: What is Machine Learning?
The RQ 2.2 was addressed and answered by finding 11,13 and 14

RQ 2.3: What are knowledge-based systems?
The RQ 2.3 was addressed and answered by finding 11,13 and 14

RQ: 2.4 What is explainable Artificial Intelligence?
The RQ 2.4 was addressed and answered by finding 11,13 and 14

RQ: 2.5 What are the Stakeholders of XAI and how do they relate to the stakeholders in corporate planning?
The RQ 2.5 was addressed and answered by finding 11,13 and 14

With this result and the fact that the findings have answered all subsequent research questions, the research question RQ2 is considered as being answered.

RQ 3: How is a Reference Architecture for an explainable AI system being designed and developed?

RQ 3.1: What are the various theoretical approaches for constructing a reference architecture?
The RQ 3.1 was addressed and answered by finding 15

RQ 3.2: What methodology for designing and developing a reference architecture can be provided?
The RQ 3.2 was addressed and answered by finding 15

With this result and the fact that the findings have answered all subsequent research questions, the research question RQ3 is considered as being answered.

RQ 4: How to provide guidance on creating a reference architecture for explainable artificial intelligence in the operational planning context?

RQ 4.1: To create a reference architecture, what preparations and basic assumptions need to be taken into account? Moreover, what factors should be considered throughout the lifecycle to guarantee explainability?
The RQ 4.1 was addressed and answered by finding 16 and Chapter 5.2.1

RQ 4.2: What are some existing architectures that could be used as a foundation?
The RQ 4.2 was addressed and answered by finding 16 and Chapter 5.2.2

RQ 4.3: How can the requirements be summarised?
The RQ 4.3 was addressed and answered by finding 16 and Chapter 5.2.3

RQ 4.4: What is the Business Layer of Re_fish?
The RQ 4.4 was addressed and answered by finding 16 and Chapter 5.2.4

RQ 4.5: What is the Application Layer of Re_fish?
The RQ 4.5 was addressed and answered by finding 16 and Chapter 5.2.5

RQ 4.6: What is the Technology Layer of Re_fish?

The RQ 4.6 was addressed and answered by finding 16 and Chapter 5.2.6

RQ 4.7: What is the process for managing the lifecycle of an explainable AI system?
The RQ 4.7 was addressed and answered by finding 16 and Chapter 5.2.8

RQ 4.8: How can a reference architecture be evaluated?
The RQ 4.8 was addressed and answered by finding 17 and Chapter 5.3

RQ 4.9 What is the gap between the generic framework and expert opinion?
The RQ 4.9 was addressed and answered by finding 17 and Chapter 5.4

The next goal is to carry out another iteration in which the contents of the evaluation are incorporated and used to improve the reference architecture by a prototype to be created in the next (iteration) step as an instantiation. This prototype will also be about an implementation of a neuro-symbolic method.

Future research can and should follow several paths. One is to better understand what decisions need to be made in strategic planning and business decisions in general. In the area of explaining how people explain people and how machines should do the same, there is still a need for deeper analysis - especially with regard to causality. This especially also in the context of neurosymbolic systems and the connection of deep learning and Symbolic AI

REFERENCES

AAAI Fall Symposium; Association for the Advancement of Artificial Intelligence; Association for the Advancement of Artificial Intelligence Fall Symposium. (2017). *The 2017 AAAI Fall Symposium series: (collected in one volume)*. *Technical reports / Association for the Advancement of Artificial Intelligence FS*. AAAI Press.

Abbott, R. (2019). *The artificial inventor project*. https://www.cipco.uzh.ch/dam/jcr:1a3a7015-02c8-4b38-954b-961ef12308d0/pr%c3%a4sentation%20abbott_cipco%20online%20work-shop%2011.06.2021.pdf

Acemoglu, D., & Autor, D. (2011). Skills, Tasks and Technologies: Implications for Employment and Earnings. In *Handbooks in economics: Vol. 5. Handbook of labor economics* (Vol. 4, pp. 1043–1171). North-Holland. https://doi.org/10.1016/S0169-7218(11)02410-5

Acemoglu, D., & Restrepo, P. (2018). The Race between Man and Machine: Implications of Technology for Growth, Factor Shares, and Employment. *American Economic Review*, *108*(6), 1488–1542. https://doi.org/10.1257/aer.20160696

Adams, D. (2010). *The Ultimate Hitchhiker's Guide to the Galaxy: Five Novels in One Outrageous Volume*. *Hitchhiker's Guide to the Galaxy Ser*. Random House Publishing Group. https://en.wikipedia.org/wiki/The_Hitchhiker%27s_Guide_to_the_Galaxy

Adams, R. J., Smart,P., & Huff, A. S. (2017). Shades of Grey: Guidelines for Working with the Grey Literature in Systematic Reviews for Management and Organizational Studies. International Journal of Management Reviews, 19(4), 432–454. https://doi.org/10.1111/ijmr.12102

Adrien Bibal, & Benoît Frénay (2016). Interpretability of Machine Learning Models and Representations: an Introduction. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*. https://www.researchgate.net/publication/326839249_Interpretability_of_Machine_Learning_Models_and_Representations_an_Introduction

Agrawal, A., Gans, J., & Goldfarb, A. (Eds.). (2019). *National Bureau of Economic Research conference report. The economics of artificial intelligence: An agenda*. The University of Chicago Press.

Aguinis, H., Ramani, R. S., & Alabduljader, N. (2018). What You See Is What You Get? Enhancing Methodological Transparency in Management Research. *Academy of Management Annals*, *12*(1), 83–110. https://doi.org/10.5465/annals.2016.0011

Ahmad, N., Ribarsky, J., & Reinsdorf, M. (2017). *OECD Statistics Working Papers.* https://doi.org/10.1787/18152031

Ahmed, M., & Sundaram, D. (2008). A Framework for a Scenario Driven Decision Support Systems Generator. *International Journal of Information Technology and Web Engineering*, *3*(2), 45–62. https://doi.org/10.4018/jitwe.2008040104

Ahmed, D. M., Sundaram, D., & Piramuthu, S. (2010). Knowledge-based scenario management — Process and support. *Decision Support Systems*, *49*(4), 507–520. https://doi.org/10.1016/j.dss.2010.06.004

Alicioglu, G., & Sun, B. (2021). A survey of visual analytics for Explainable Artificial Intelligence methods. *Computers & Graphics.* Advance online publication. https://doi.org/10.1016/j.cag.2021.09.002

Alpaydin, E. (2019). *Maschinelles Lernen* (2. Auflage). *De Gruyter Studium.*

Andrews, D., Criscuolo, C., & Gal, Peter, N. (2016). *"The Best versus the Rest: The Global Productivity Slowdown, Divergence across Firms and the Role of Public Policy"*. OECD Publishing; Centre for the Study of Living Standards. https://doi.org/10.1787/9789264279179-en

Andrews, D., Criscuolo, C., & Gal, P., N. (2017). *OECD Productivity Working Papers.* https://doi.org/10.1787/24139424

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica*. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

Appio, F. P., La Torre, D., Lazzeri, F., Masri, H., & Schiavone, F. (2023). *Impact of Artificial Intelligence in Business and Society: Opportunities and Challenges*. *Routledge Studies in Innovation, Organizations and Technology Series*. Taylor & Francis Group.

Arnold, T., Kasenberg, D., & Scheutz, M. (2021). Explaining in Time. *ACM Transactions on Human-Robot Interaction*, *10*(3), 1–23. https://doi.org/10.1145/3457183

Augier, M., & Teece, D. J. (Eds.). (2019). *Springer eBook Collection*. *The Palgrave Encyclopedia of Strategic Management*. Palgrave Macmillan. https://doi.org/10.1057/978-1-349-94848-2

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & van Reenen, J. (2017). Concentrating on the Fall of the Labor Share. *American Economic Review*, *107*(5), 180–185. https://doi.org/10.1257/aer.p20171102

Autor, D., Dorn, D., Katz, L. F., Patterson, C., & van Reenen, J. (2020). The Fall of the Labor Share and the Rise of Superstar Firms *The Quarterly Journal of Economics*, *135*(2), 645–709. https://doi.org/10.1093/qje/qjaa004

Autor, D. H., Levy, F., & Murnane, R. J. (2003). The Skill Content of Recent Technological Change: An Empirical Exploration. *The Quarterly Journal of Economics*, *118*(4), 1279–1333. https://doi.org/10.1162/003355303322552801

Bader, S., & Hitzler, P [Pascal]. (2005, November 10). *Dimensions of Neural-symbolic Integration - A Structured Survey*. https://arxiv.org/pdf/cs/0511042

Bansal, N., Agarwal, C., & Nguyen, A. (2020). SAM: The Sensitivity of Attribution Methods to Hyperparameters. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. https://openaccess.thecvf.com/content_CVPR_2020/html/Bansal_SAM_The_Sensitivity_of_Attribution_Methods_to_Hyperparameters_CVPR_2020_paper.html

Barbot, N., Miclet, L., & Prade, H. (2019). Analogy between concepts. *Artificial Intelligence*, *275*, 487–539. https://doi.org/10.1016/j.artint.2019.06.008

Barredo 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

Bass, L. (2013). *Software architecture in practice* (3rd ed.). *SEI series in software engineering*. Addison-Wesley. https://learning.oreilly.com/library/view/-/9780132942799/?ar

Bass, L., Clements, P., & Kazman, R. (2022). *Software architecture in practice* (Fouth edition). *Always learning*. Addison-Wesley.

Baur, N., & Blasius, J. (Eds.). (2019). *Springer eBook Collection*. *Handbuch Methoden der empirischen Sozialforschung* (2., vollständig überarbeitete und erweiterte Auflage). Springer VS. https://doi.org/10.1007/978-3-658-21308-4

Been Kim, Khanna, R., & Oluwasanmi O. Koyejo (2016). Examples are not enough, learn to criticize! Criticism for Interpretability. *NIPS'16: Proceedings of the 30th International Conference on Neural Information Processing Systems*, 2288–2296.

Been, K., & Khanna, Rajiv, Koyejo, Oluwasanami. Examples are not enough, learn to criticize! criticism for interpretability.

Bejger, S., & Elster, S. (2019). Das Blackbox Problem: Künstlicher Intelligenz vertrauen. *AI Spektrum*, *1*, 24–27.

Bejger, S., & Elster, S. (2020). Artificial Intelligence in economic decision making: How to assure a trust? *Ekonomia I Prawo. Economics and Law*, *19*(3), 411. https://doi.org/10.12775/EiP.2020.028

Benjamin, J. J., & Müller-Birn, C. (2019). Materializing Interpretability. In S. Harrison, S. Bardzell, C. Neustaedter, & D. Tatar (Eds.), *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion* (pp. 123–127). ACM. https://doi.org/10.1145/3301019.3323900

Berens, W., Delfmann, W., & Schmitting, W. (2004). *Quantitative Planung: Grundlagen, Fallstudien, Lösungen* (4., überarb. und erw. Aufl.). Schäffer-Poeschel.

Bergeaud, A., Cette, G., & Lecat, R. (2018). The role of production factor quality and technology diffusion in twentieth-century productivity growth. *Cliometrica*, *12*(1), 61–97. https://doi.org/10.1007/s11698-016-0149-2

Bernard, D., & Arnold, A. (2019). Cognitive interaction with virtual assistants: From philosophical foundations to illustrative examples in aeronautics. *Computers in Industry*, *107*, 33–49. https://doi.org/10.1016/j.compind.2019.01.010

Bernhaupt, R., Mueller, F. '., Verweij, D., Andres, J., McGrenere, J., Cockburn, A., Avellino, I., Goguey, A., Bjørn, P., Zhao, S., Samson, B. P., & Kocielnik, R. (Eds.) (04212020). *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. ACM.

Bhatt, U., Ravikumar, P., & Moura, J. M. F. (2019). Building Human-Machine Trust via Interpretability. *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 9919–9920. https://doi.org/10.1609/aaai.v33i01.33019919

Blackman, R. (2022). *Ethical machines: Your concise guide to totally unbiased, transparent, and respectful AI* (First ebook edition). Harvard Business Review Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=3077296

Bogaerts, B., Gamba, E., & Guns, T. (2021). A framework for step-wise explaining how to solve constraint satisfaction problems. *Artificial Intelligence*, *300*, 103550. https://doi.org/10.1016/j.artint.2021.103550

Bolander, T. (2019). What do we loose when machines take the decisions? *Journal of Management and Governance*, *23*(4), 849–867. https://doi.org/10.1007/s10997-019-09493-x

Bonin, H., Gregory, T., & Zierahn, U. (2013). *Übertragung der Studie von Frey*. https://scholar.google.com/citations?user=53vwm932ulec&hl=en&oi=sra

Bostrom, N. (2017). *Superintelligence: Paths, dangers, strategies* (Reprinted with corrections 2017). Oxford University Press.

Bottou, L [Leon]. (2011). From Machine Learning to Machine Reasoning. *Tr*-. https://arxiv.org/pdf/1102.1808

Bougie, R., & Sekaran, U. (2020). *Research methods for business: A skill-building approach* (Eight edition). Wiley.

Bower, p. (2012). Integrated Business Planning: Is It a Hoax or Here to Stay? *Journal of Business Forecasting*.

Breiman, L. (2001). Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author). *Statistical Science*, *16*(3), 199–231. https://doi.org/10.1214/ss/1009213726

Brennan, D. J. (2020). *Process industry economics: Principles, concepts and applications* (Second edition). Elsevier.

Brewster, S., Fitzpatrick, G., Cox, A., & Kostakos, V. (Eds.) (2019). *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Brewster, S., Fitzpatrick, G., Cox, A., & Kostakos, V. (Eds.) (2019). *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. ACM.

Briner, R. B., Denyer, D., & Rousseau, D. M. (2009). Evidence-Based Management: Concept Cleanup Time? *Academy of Management Perspectives*, *23*(4), 19–32. https://doi.org/10.5465/amp.23.4.19

Brynjolfsson, E., & Collis, A. (2019). How should we measure the digital economy? Focus on the value created, not just the prices paid. *Harvard Business Review*, *97*(6), 140-.

Brynjolfsson, E., & McAfee, A. (2016). The second machine age: Work, progress, and prosperity in a time of brilliant technologies (First published as a Norton paperback). W. W. Norton & Company.

Buchanan, B., G., Shortliffe, E. H. (1984). *Rule- Based Expert Systems : The MYCIN Experiments of the Stanford Heuristic Programming Project*. http://papers.cumincad.org/cgi-bin/works/show&_id=caadria2010_044/paper/ec87

Buchanan, D. A., & Bryman, A. (2011). *The SAGE handbook of organizational research methods* (Paperback ed.). SAGE.

Bunn, J. (2020). Working in contexts for which transparency is important. *Records Management Journal*, *30*(2), 143–153. https://doi.org/10.1108/RMJ-08-2019-0038

Cao, L. (Ed.) (2015). *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *ACM Digital Library*. ACM.

Cardoso Ermel, A. P., Lacerda, D. P., Morandi, M. I. W. M., & Gauss, L. (2021). *Literature Reviews*. Springer International Publishing. https://doi.org/10.1007/978-3-030-75722-9

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible Models for HealthCare. In L. Cao (Ed.), *ACM Digital Library, Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1721–1730). ACM. https://doi.org/10.1145/2783258.2788613

Caruana, R., Lundberg, S., Ribeiro, M. T., Nori, H., & Jenkins, S. (2020). Intelligible and Explainable Machine Learning. In R. Gupta, Y. Liu, J. Tang, & B. A. Prakash (Eds.), Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (pp. 3511–3512). ACM. https://doi.org/10.1145/3394486.3406707

CEFIC. (2023). *2023 Facts And Figures Of The European Chemical Industry*. https://cefic.org/a-pillar-of-the-european-economy/facts-and-figures-of-the-european-chemical-industry/

Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2019). Balancing Explicability and Explanations in Human-Aware Planning. In S. Kraus (Ed.), *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)* (pp. 1335–1343). International Joint Conferences on Artificial Intelligence. https://doi.org/10.24963/ijcai.2019/185

Chakraborti, T., Sreedharan, S., & Kambhampati, S. (2020). The Emerging Landscape of Explainable AI Planning and Decision Making Chan, F. T. S. (2005). Application of a hybrid case-based reasoning approach in electroplating industry. Expert Systems with Applications, 29(1), 121–130. https://doi.org/10.1016/j.eswa.2005.01.010

Chapman, P., Clinton, J., Kerber, R., Khabaza, T.,. Reinartz, T. P., Shearer, C., & Wirth, R. (2000). CRISP-DM 1.0: Step-by-step data mining guide. In

Chan, F. (2005). Application of a hybrid case-based reasoning approach in electroplating industry. *Expert Systems with Applications*, *29*(1), 121–130. https://doi.org/10.1016/j.eswa.2005.01.010

Chari, S., Gruen, D.M., Seneviratne, O., & McGuiness, D.L. (2020). Foundations of Explainable Knowledge-Enabled Systems. In Tiddi, I., Lécué, F., & Hitzler, P. (Eds.). (2020). Studies on the semantic web: volume 047. Knowledge graphs for explainable artificial intelligence: Foundations, applications and challenges. IOS Press; AKA. Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. Journal of Business Research, 122, 502–517. https://doi.org/10.1016/j.jbusres.2020.09.009

Chari, S., Gruen, D.M., Seneviratne, O., & McGuiness, D.L. (2020). Directions for Explainable Knowledge-Enabled Systems. In Tiddi, I., Lécué, F., & Hitzler, P. (Eds.). (2020). Studies on the semantic web: volume 047. Knowledge graphs for explainable artificial intelligence: Foundations, applications and challenges. IOS Press; AKA. Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. Journal of Business Research, 122, 502–517. https://doi.org/10.1016/j.jbusres.2020.09.009

Charniak, E., & McDermott, D. V. (1987). *Introduction to artificial intelligence* (Reprinted with corrections). *Addison-Wesley Series in computer science*. Addison-Wesley.

Chattopadhay, A., Sarkar, A., Howlader, P., & Balasubramanian, V. N. (2017). *Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks*. https://arxiv.org/pdf/1710.11063 https://doi.org/10.1109/WACV.2018.00097

Chen, P.H. C., Gadepalli, K., MacDonald, R., Liu, Y [Yun], Kadowaki, S., Nagpal, K., Kohlberger, T., Dean, J., Corrado, G. S., Hipp, J. D., Mermel, C. H., & Stumpe, M. C. (2019). An augmented reality microscope with real-time artificial intelligence integration for cancer diagnosis. *Nature Medicine*, *25*(9), 1453–1457. https://doi.org/10.1038/s41591-019-0539-7

Chiang, L. H., & Braatz, R. D. (2003). Process monitoring using causal map and multivariate statistics: fault detection and identification. *Chemometrics and Intelligent Laboratory Systems*, *65*(2), 159–178. https://doi.org/10.1016/S0169-7439(02)00140-5

Chiusi, F., Fischer, S., Kayser-Brill, N., & Spielkamp, M. *Automating Society Report 2020.*

Chou, Y.L., Moreira, C., Bruza, P., Ouyang, C., & Jorge, J. (2021, March 7). *Counterfactuals and Causability in Explainable Artificial Intelligence: Theory, Algorithms, and Applications*. https://arxiv.org/pdf/2103.04244

Menzel, C., & Winkler, C. (2019). *Zur Diskussion der Effekte Künstlicher Intelligenz in der wirtschaftswissen-schaftlichen Literatur*. https://www.researchgate.net/profile/christoph-menzel-6/publication/331231649_zur_diskussion_der_effekte_kunstlicher_intelligenz_in_der_wirtschaftswissenschaftlichen_literatur

Chu, E., Gillani, N., & Priscilla Makini, S. (2020). Games for Fairness and Interpretability. In A. E. F. Seghrouchni, G. Sukthankar, T.-Y. Liu, & M. van Steen (Eds.), *Companion Proceedings of the Web Conference 2020* (pp. 520–524). ACM. https://doi.org/10.1145/3366424.3384374

Clancey, W. J. (1983). The epistemology of a rule-based expert system —a framework for explanation. *Artificial Intelligence*, *20*(3), 215–251. https://doi.org/10.1016/0004-3702(83)90008-5

Cleve, J., & Lämmel, U. (2016). *Data Mining* (2nd ed.). *De Gruyter Studium*. De Gruyter.

Coats, P. K. (1988). Why Expert Systems Fail. *Financial Management*, *17*(3), 77. https://doi.org/10.2307/3666074

Coldrick, A., Ling, D., & Turner, C. (2003). *Evolution of Sales & Operations Planning-From Production Planning to Integrated Decision Making*. https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=00cef55eb75bb6379fe451ea0dff67a6c20a9c63

Feigenbaum, E.A., & Feldman, J. (1963). Computers and thought. McGraw-Hill. https://ojs.aaai.org/index.php/aimagazine/article/download/1618/1517/0

Conitzer, V., Hadfield, G., & Vallor, S. (Eds.) (2019). *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Costa, J., Vasconcelos, A., & Fragoso, B. (2020). An Enterprise Architecture Approach for Assessing the Alignment Between Projects and Goals. *International Journal of Information Technology Project Management*, *11*(3), 55–76. https://doi.org/10.4018/IJITPM.2020070104

Could advanced AI drive explosive economic growth. (2021). https://www.openphilanthropy.org/wp-content/uploads/could-advanced-ai-drive-explosive-economic-growth_.pdf

Cremer, D. de, & Kasparov, G. (2022). The ethics of technology innovation: A double-edged sword? *AI and Ethics*, *2*(3), 533–537. https://doi.org/10.1007/s43681-021-00103-x

Crnkovic, I. (Ed.) (2011). *Proceedings of the joint ACM SIGSOFT conference -- QoSA and ACM SIGSOFT symposium -- ISARCS on Quality of software architectures -- QoSA and architecting critical systems -- ISARCS*. *ACM Conferences*. ACM.

Cropanzano, R. (2009). Writing Nonempirical Articles for Journal of Management: General Thoughts and Suggestions. *Journal of Management*, *35*(6), 1304–1311. https://doi.org/10.1177/0149206309344118

Da Xu, L., Xu, E. L., & Li, L. (2018). Industry 4.0: State of the art and future trends. *International Journal of Production Research*, *56*(8), 2941–2962. https://doi.org/10.1080/00207543.2018.1444806

Daun, A. (2013). *Referenzmodell für den Einsatz von Bildungsmethoden für E-Learning, Wissens- und Kompetenzmanagement* [Duisburg, Essen, Universität Duisburg-Essen, Diss., 2013, Universitätsbibliothek Duisburg-Essen, Duisburg, Essen]. Deutsche Nationalbibliothek.

Davenport, T. H., & Miller, S. M. (2022). *Working with AI: Real stories of human-machine collaboration*. *Management on the cutting edge*. The MIT Press. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=7069285

David, J.-M., Krivine, J.-P., & Simmons, R. (Eds.). (1993). *Second Generation Expert Systems*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-77927-5

Dengel, A. (2012). *Semantische Technologien: Grundlagen - Konzepte - Anwendungen*. Spektrum Akademischer Verlag. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=3067327

Denning, P. J. (1997). A new social contract for research. *Communications of the ACM*, *40*(2), 132–134. https://doi.org/10.1145/253671.253755

Denyer Denis. (2009). *The Sage Handbook of Organizational Research Methods.*

Dev, S., Sameki, M., Dhamala, J., & Hsieh, C.-J. (2021). Measures and Best Practices for Responsible AI. In F. Zhu, B. Chin Ooi, & C. Miao (Eds.), *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (p. 4118). ACM. https://doi.org/10.1145/3447548.3469458

Doering, N. (2023). *Forschungsmethoden und Evaluation in den Sozial- und Humanwissenschaften* (6., vollst. uberarb., akt. u. erw. Aufl. 2023). Springer-Verlag Berlin and Heidelberg GmbH & Co. KG.

Dolgui, A., Bernard, A., Lemoine, D., Cieminski, G. von, & Romero, D. (Eds.). (2021). *Springer eBook Collection: Vol. 632*. *Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems: Ifip WG 5.7 International Conference, APMS 2021, Nantes, France, September 5–9, 2021, Proceedings, Part III* (1st ed. 2021). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-85906-0

Doshi-Velez, F., & Kim, B. (2017, February 28). *Towards A Rigorous Science of Interpretable Machine Learning*. http://arxiv.org/pdf/1702.08608v2

Doshi-Velez, F., Kortz, M., Budish, R., Bavitz, C., Gershman, S., O'Brien, D., Scott, K., Schieber, S., Waldo, J., Weinberger, D., Weller, A., & Wood, A. (2017, November 3). *Accountability of AI Under the Law: The Role of Explanation*. http://arxiv.org/pdf/1711.01134v3

Dubiel, M. (2018). Towards Human-Like Conversational Search Systems. In C. Shah, N. J. Belkin, K. Byström, J. Huang, & F. Scholer (Eds.), *Chiir'18: Proceedings of the 2018 Conference on Human Information Interaction and Retrieval: March 11-15, 2018, New Brunswick, NJ, USA* (pp. 348–350). ACM Association for Computing Machinery. https://doi.org/10.1145/3176349.3176360

Durand, R., Grant, R. M., & Madsen, T. L. (2017). The expanding domain of strategic management research and the quest for integration. *Strategic Management Journal*, *38*(1), 4–16. https://doi.org/10.1002/smj.2607

Durkin, J. (1994). *Expert systems: Design and development*. Prentice-Hall.

Dwivedi, R [Rudresh], Dave, D., Naik, H., Singhal, S., Omer, R., Patel, P., Qian, B., Wen, Z., Shah, T., Morgan, G., & Ranjan, R. (2023). Explainable AI (XAI): Core Ideas, Techniques, and Solutions. *ACM Computing Surveys*, *55*(9), 1–33. https://doi.org/10.1145/3561048

Dwivedi, Y. K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R [Rohita], Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P. V., Janssen, M., Jones, P., Kar, A. K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., . . . Williams, M. D. (2019). Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *International Journal of Information Management*, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

E.J. Russo, & P.J.H. Schoemaker (1992). Managing Overconfidence. *Sloan Management Review*, *33*(2), 7–17. https://www.researchgate.net/profile/paul-schoemaker-2/publication/306940378_managing_overconfidence

Ehsan, U., Harrison, B., Chan, L., & Riedl, M. O. (2018). Rationalization. In J. Furman, G. Marchant, H. Price, & F. Rossi (Eds.), *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 81–87). ACM. https://doi.org/10.1145/3278721.3278736

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation. In W.-T. Fu, S. Pan, O. Brdiczka, P. Chau, & G. Calvary (Eds.), *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 263–274). ACM. https://doi.org/10.1145/3301275.3302316

Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., Riener, A., & Riedl, M. O. (2021). Operationalizing Human-Centered Perspectives in Explainable AI. In Y. Kitamura, A. Quigley, K. Isbister, & T. Igarashi (Eds.), *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). ACM. https://doi.org/10.1145/3411763.3441342

Eibeck, A., Chadzynski, A., Lim, M. Q., Aditya, K., Ong, L., Devanand, A., Karmakar, G., Mosbach, S., Lau, R., Karimi, I. A., Foo, E. Y. S., & Kraft, M. (2020). A Parallel World Framework for scenario analysis in knowledge graphs. *Data-Centric Engineering*, *1*. https://doi.org/10.1017/dce.2020.6

Eiras-Franco, C., Guijarro-Berdiñas, B., Alonso-Betanzos, A., & Bahamonde, A. (2019). A scalable decision-tree-based method to explain interactions in dyadic data. *Decision Support Systems*, *127*, 113141. https://doi.org/10.1016/j.dss.2019.113141

Elster, S. (1997). *Implementierung der DIN EN ISO 9000 ff. in international orientierten Handelsunternehmen unter besonderer Berücksichtigung der heicko Schraubenvertriebs- GmbH* [Diplom Kaufmann Thesis]. Westfälische Wilhelms Universität, Münster.

Elster, S. (2005). *Implementierung von Business Intelligence Systemen unter besonderer Berücksichtigung des Projekts Ulrich Walter GmbH* [Diplom Wirtschaftsinformatiker Thesis]. Westsächsische Hochschule Zwickau.

Elster, S. (2009). *Performance Management in Supply Chain Approaches* [Master Thesis]. Westfälische Wilhelms Universität, Münster.

Elster, J. (2009). *Reason and Rationality*. Princeton University Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=355028 https://doi.org/10.1515/9781400833177

Elster, J. (2015). *Explaining social behavior: More nuts and bolts for the social sciences* (Second edition). Cambridge University Press. https://doi.org/10.1017/CBO9781107763111

Elster, J., & Geitel, H. F. (1897): *Zusammenstellung der Ergebnisse neuerer Arbeiten über atmosphärische Elekticität*, Wolfenbüttel

Ethiraj, S. K., Gambardella, A., & Helfat, C. E. (2017). Reviews of strategic management research. *Strategic Management Journal*, *38*(1), 3. https://doi.org/10.1002/smj.2606

EUR-Lex - 52021PC0206 - EN - EUR-Lex. (2023, June 18). https://eur-lex.europa.eu/legal-content/DE/TXT/?uri=CELEX%3A52021PC0206

Falter, W., Keller, A., Nickel, J.-P., & Meincke, H. (2017). *Chemie 4.0: Wachstum durch Innovation in einer Welt im Umbruch*. Endbericht. VCI, Deloitte. https://www.vci.de/vci/downloads-vci/publikation/vci-deloitte-studie-chemie-4-punkt-0-langfassung.pdf

Fernández, R. R., Martín de Diego, I., Aceña, V., Fernández-Isabel, A., & Moguerza, J. M. (2020). Random forest explainability using counterfactual sets. *Information Fusion*, *63*, 196–207. https://doi.org/10.1016/j.inffus.2020.07.001

Fettke, P., & Loos, P. (2004). Referenzmodellierungsforschung: Langzeitfassung eines Aufsatzes. http://www.isym.bwl.uni-mainz.de

Fischer, D., & Breitenbach, J. (2020). *Die Pharmaindustrie*. Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-662-61035-0

Fischer, A., & Funke, J. (2016). Entscheiden und Entscheidungen: Die Sicht der Psychologie. In S. Kirste (Ed.), *Recht und Philosophie: Band 1. Interdisziplinarität in den Rechtswissenschaften: Ein interdisziplinärer und internationaler Dialog* (pp. 217–229). Duncker & Humblot.

Foot, P. (1967). *The Problem of Abortion and the Doctrine of the Double Effect*. https://philpapers.org/rec/footpo-2

Fox, M., Long, D., & Magazzeni, D. (2017, September 29). *Explainable Planning*. https://arxiv.org/pdf/1709.10256

Frankfurter Allgemeine Zeitung (2022, March 27). Künstliche Intelligenz: Wie Menschen und Algorithmen gemeinsam bessere Entscheidungen treffen. *Frankfurter Allgemeine Zeitung*. https://www.faz.net/aktuell/wirtschaft/digitec/kuenstliche-intelligenz-computer-koennen-menschen-gut-ergaenzen-17913366.html

Franzen, A. (2019). Antwortskalen in standardisierten Befragungen. In N. Baur & J. Blasius (Eds.), *Springer eBook Collection. Handbuch Methoden der empirischen Sozialforschung* (2nd ed., pp. 843–854). Springer VS. https://doi.org/10.1007/978-3-658-21308-4_58

Frey, C. B., & Osborne, M. A. (2017). The future of employment: How susceptible are jobs to computerisation? *Technological Forecasting and Social Change*, *114*, 254–280. https://doi.org/10.1016/j.techfore.2016.08.019

Fu, W.-T., Pan, S., Brdiczka, O., Chau, P., & Calvary, G. (Eds.) (2019). *Proceedings of the 24th International Conference on Intelligent User Interfaces*. ACM.

Funkat, D. (1995). *Die Erklärungskomponente in hybriden Expertensystemen mit numerischer und symbolischer Wissensverarbeitung: Entwurf und Realisierung der Erklärungskomponente für ein Expertensystem in der Diabetestherapie - DIABETEX*. Zugl.: Ilmenau, Techn. Univ., Diss., 1994 (1. Aufl.). Cuvillier.

Furman, J., Marchant, G., Price, H., & Rossi, F. (Eds.) (2018). *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Furman, J., & Orszag, P. (2018). 1. A Firm-Level Perspective on the Role of Rents in the Rise in Inequality. In *Toward a Just Society* (pp. 19–47). Columbia University Press. https://doi.org/10.7312/guzm18672-003

Furman, J., & Orszag, P. R. (2018). *Slower Productivity and Higher Inequality: Are They Related?* https://doi.org/10.2139/ssrn.3191984

Fürnkranz, J., Kliegr, T., & Paulheim, H. (2020). On cognitive preferences and the plausibility of rule-based models. *Machine Learning*, *109*(4), 853–898. https://doi.org/10.1007/s10994-019-05856-5

Futia, G., & Vetrò, A. (2020). On the Integration of Knowledge Graphs into Deep Learning Models for a More Comprehensible AI—Three Challenges for Future Research. *Information*, *11*(2), 122. https://doi.org/10.3390/info11020122

G. Schuh, P. Scholz, T. Leich, & R. May (2020). Identifying and Analyzing Data Model Requirements and Technology Potentials of Machine Learning Systems in the Manufacturing Industry of the Future. In *2020 61st International Scientific Conference on Information Technology and Management Science of Riga Technical University (ITMS)*.

Galster, M., & Avgeriou, P. (2011). Empirically-grounded reference architectures. In I. Crnkovic (Ed.), *ACM Conferences, Proceedings of the joint ACM SIGSOFT conference -- QoSA and ACM SIGSOFT symposium -- ISARCS on Quality of software architectures -- QoSA and architecting critical systems -- ISARCS* (pp. 153–158). ACM. https://doi.org/10.1145/2000259.2000285

Garcez, A. d [Artur d'Avila], Gori, M., Lamb, L. C [Luis C.], Serafini, L., Spranger, M., & Tran, S. N. (2019). *Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning*.

Garcez, A. d [Artur d'Avila], & Lamb, L. C [Luis C.]. (2020, December 10). *Neurosymbolic AI: The 3rd Wave*. https://arxiv.org/pdf/2012.05876

Garcez, A. d [Artur d'Avila], & Lamb, L. C [Luís C.] (2023). Neurosymbolic AI: The 3rd wave. *Artificial Intelligence Review*, 1–20. https://doi.org/10.1007/s10462-023-10448-w

Garnelo, M., & Shanahan, M. (2019). Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, *29*, 17–23. https://doi.org/10.1016/j.cobeha.2018.12.010

Gaur, A., & Kumar, M. (2018). A systematic approach to conducting review studies: An assessment of content analysis in 25 years of IB research. *Journal of World Business*, *53*(2), 280–289. https://doi.org/10.1016/j.jwb.2017.11.003

Ge, Z. (2017). Review on data-driven modeling and monitoring for plant-wide industrial processes. *Chemometrics and Intelligent Laboratory Systems*, *171*, 16–25. https://doi.org/10.1016/j.chemolab.2017.09.021

Ge, Z., Song, Z., Ding, S. X., & Huang, B. (2017). Data Mining and Analytics in the Process Industry: The Role of Machine Learning. *IEEE Access*, *5*, 20590–20616. https://doi.org/10.1109/access.2017.2756872

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems / Aurélien Géron* (Second Edition). O'Reilly.

Gharbi, M., Koschel, A., & Rausch, A. (2019). *Software Architecture Fundamentals: A Study Guide for the Certified Professional for Software Architecture® – Foundation Level – iSAQB compliant*. dpunkt.verlag. http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1298524

Gharbi, M., Koschel, A., Rausch, A., & Starke, G. (2020). *Basiswissen für Softwarearchitekten: Aus- und Weiterbildung nach iSAQB-Standard zum Certified Professional for Software Architecture - Foundation Level* (4., überarbeitete und aktualisierte Auflage). *Basiswissen*. dpunkt.verlag. http://www.content-select.com/index.php?id=bib_view&ean=9783969100127

Ghosal, S., Blystone, D., Singh, A. K., Ganapathysubramanian, B., Singh, A., & Sarkar, S. (2018). An explainable deep machine vision framework for plant stress phenotyping. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(18), 4613–4618. https://doi.org/10.1073/pnas.1716999115

Gianfagna, L., & Cecco Di, A. (2021). *Explainable AI with Python* (1st ed. 2021). https://doi.org/10.1007/978-3-030-68640-6

Gigerenzer, G., & Selten, R. (Eds.). (2002). *Bounded rationality: The adaptive toolbox* (1st MIT Press pbk. ed.). MIT Press. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=78072

Giglietto, F., & Rossi, L. (2012). Ethics and Interdisciplinarity in Computational Social Science. *Methodological Innovations Online*, *7*(1), 25–36. https://doi.org/10.4256/mio.2012.003

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018, May 31). *Explaining Explanations: An Overview of Interpretability of Machine Learning*. https://arxiv.org/pdf/1806.00069

Gobet, F., & Chassy, P. (2009). Expertise and Intuition: A Tale of Three Theories. *Minds and Machines*, *19*(2), 151–180. https://doi.org/10.1007/s11023-008-9131-5

Goodman, B., & Flaxman, S. (2017). European Union Regulations on Algorithmic Decision-Making and a "Right to Explanation". *AI Magazine*, *38*(3), 50–57. https://doi.org/10.1609/aimag.v38i3.2741

Goos, M., & Manning, A. (2007). Lousy and Lovely Jobs: The Rising Polarization of Work in Britain. *Review of Economics and Statistics*, *89*(1), 118–133. https://doi.org/10.1162/rest.89.1.118

Gordon, R. (2017). *The rise and fall of American growth: The US standard of living since the civil war.* (2017). https://www.degruyter.com/document/doi/10.1515/9781400888955/html

Görz, G., Schmid, U., & Braun, T. (Eds.). (2021). *Handbuch der künstlichen Intelligenz* (6. Auflage). De Gruyter Oldenbourg.

Graus, E., Özgül, P., & Steen, S. (2021). *Künstliche Intelligenz: Die Zukunft der Arbeit anhand von Erkenntnissen aus der Unternehmenspraxis gestalten*. https://cris.maastrichtuniversity.nl/files/77657433/aiconomics_policybrief_german.pdf

Grennan, L., Kremer, A., Zipparo, P., & Singla, A. (2022). *Why businesses need explainable AI—and how to deliver it.* McKinsey Global Institute. https://www.mckinsey.com/capabilities/quantumblack/our-insights/why-businesses-need-explainable-ai-and-how-to-deliver-it#/

Grennan, J., & Michaely, R. (2020). Artificial Intelligence and High-Skilled Work: Evidence from Analysts. *SSRN Electronic Journal.* Advance online publication. https://doi.org/10.2139/ssrn.3681574

Growiec, J. (2022). *Accelerating economic growth: Lessons from 200,000 years of technological progress and human development*. *Frontiers in economic history*. Springer International Publishing AG. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=7081882

Gruetzemacher (2019). A Holistic Framework for Forecasting Transformative AI. *Big Data and Cognitive Computing*, *3*(3), 35. https://doi.org/10.3390/bdcc3030035

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*, *51*(5), 1–42. https://doi.org/10.1145/3236009

Gunning, D. DARPA's explainable artificial intelligence (XAI) program. In https://doi.org/10.1145/3301275.3308446

Gunning, D., & Aha, D. (2019). Darpa's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, *40*(2), 44–58. https://doi.org/10.1609/aimag.v40i2.2850

Gupta, R., Liu, Y [Yan], Tang, J., & Prakash, B. A. (Eds.) (2020). *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM.

Gutiérrez, G., & Philippon, T. (2017). *Declining Competition and Investment in the U.S*. Cambridge, MA. https://doi.org/10.3386/w23583

Häder, M. (2019). *Empirische Sozialforschung: Eine Einführung* (4. Auflage). *Springer eBook Collection*. Springer VS. https://doi.org/10.1007/978-3-658-26986-9

Halpern, J. Y. (2015, May 1). *A Modification of the Halpern-Pearl Definition of Causality*. http://arxiv.org/pdf/1505.00162v1

Halpern, J. Y., & Pearl, J. (2005). Causes and Explanations: A Structural-Model Approach. Part I: Causes. *The British Journal for the Philosophy of Science*, *56*(4), 843–887. https://doi.org/10.1093/bjps/axi147

Hammer, R. (2015). *Unternehmensplanung: Planung und Führung* (9., überarbeitete und erweiterte Auflage). De Gruyter Oldenbourg.

Hanna, R. (2023). *Hinton & Me: Don't Pause Giant AI Experiments, Ban Them. https://bobhannahbob1.medium.com/hinton-me-dont-pause-giant-ai-experiments-ban-them-52c60581ae1d*

Hanschke, I. (2016). *Enterprise Architecture Management - einfach und effektiv: Ein praktischer Leitfaden für die Einführung von EAM* (2., überarbeitete Auflage). Hanser.

Hansen, L. K., & Rieger, L. (2019). Interpretability in Intelligent Systems – A New Concept?, *11700*, 41–49. https://doi.org/10.1007/978-3-030-28954-6_3

Harfouche, A. L., Jacobson, D. A., Kainer, D., Romero, J. C., Harfouche, A. H., Scarascia Mugnozza, G., Moshelion, M., Tuskan, G. A., Keurentjes, J. J. B., & Altman, A. (2019). Accelerating Climate Resilient Plant Breeding by Applying Next-Generation Artificial Intelligence. *Trends in Biotechnology*, *37*(11), 1217–1235. https://doi.org/10.1016/j.tibtech.2019.05.007

Haritsa, J., Roy, S., Gupta, M., Mehrotra, S., Srinivasan, B. V., & Simmhan, Y. (Eds.) (2021). *8th ACM IKDD CODS and 26th COMAD*. ACM.

Harrison, S., Bardzell, S., Neustaedter, C., & Tatar, D. (Eds.) (2019). *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion*. ACM.

Hars, A. (1994). *Referenzdatenmodelle: Grundlagen Effizienter Datenmodellierung*. *Schriften Zur EDV-Orientierten Betriebswirtschaft Ser*. Springer Gabler. in Springer Fachmedien Wiesbaden GmbH. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6586178

Haugeland, J. (1993). *Artificial intelligence: The very idea* (6. print). *Bradford books*. MIT Press.

Helal, A. (2021). Data Lakes Empowered by Knowledge Graph Technologies. In G. Li (Ed.), *ACM Digital Library, Proceedings of the 2021 International Conference on Management of Data* (pp. 2884–2886). Association for Computing Machinery. https://doi.org/10.1145/3448016.3450584

Hempel, C. G., & Oppenheim, P. (1948). Studies in the Logic of Explanation. *Philosophy and Science*, *15*, 135–175.

Henderson, J. C., & Venkatraman, H. (1999). Strategic alignment: Leveraging information technology for transforming organizations. *IBM Systems Journal*, *38*(2.3), 472–484. https://doi.org/10.1147/SJ.1999.5387096

Hevner et al. (2004, March). Design Science in IS Research. *MIS Quarterly*, pp. 75-105.

High-Level Expert Group on Artificial Intelligence. Ethics guidelines for trustworthy AI, *2019*. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai

Hilton, D. J. (1991). A Conversational Model of Causal Explanation. *European Review of Social Psychology*, *2*(1), 51–81. https://doi.org/10.1080/14792779143000024

Hoddinot, S. N., Bass, M. J. (1986). *The dillman total design survey method.* (1986). https://www.ncbi.nlm.nih.gov/pmc/articles/pmc2328022/

Hoffman, R. R., & Klein, G. (2017). Explaining Explanation, Part 1: Theoretical Foundations. *IEEE Intelligent Systems*, *32*(3), 68–73. https://doi.org/10.1109/MIS.2017.54

Hohman, F., Head, A., Caruana, R., DeLine, R., & Drucker, S. M. (2019). Gamut. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–13). ACM. https://doi.org/10.1145/3290605.3300809

Hollenberg, S. (2016). *Fragebögen: Fundierte Konstruktion, sachgerechte Anwendung und aussagekräftige Auswertung*. *essentials*. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=4443072

Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the United States of America*, *79*(8), 2554–2558. https://doi.org/10.1073/pnas.79.8.2554

Hruschka, P. (2019). *Business Analysis und Requirements Engineering: Produkte und Prozesse nachhaltig verbessern* (2., aktualisierte Auflage). Hanser. https://doi.org/45589

http://arxiv.org/abs/1910.12507. (2019).

Hu, Z. F., Kuflik, T., Mocanu, I. G., Najafian, S., & Shulner Tal, A. (2021). Recent Studies of XAI - Review. In J. Masthoff, E. Herder, N. Tintarev, & M. Tkalčič (Eds.), *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization* (pp. 421–431). ACM. https://doi.org/10.1145/3450614.3463354

Hulbert, D. G. (2000). An Interpreting Expert System for Advanced Control. *IFAC Proceedings Volumes*, *33*(18), 253–258. https://doi.org/10.1016/S1474-6670(17)37152-5

Ilkou, E., & Koutraki, M. (2020). Symbolic Vs Subymbolic AI Methods: Friends or Enemies? In *CSSA'20: Workshop on Combining Symbolic and Subymbolic Methods and their Applications co-located with CIKM2020*. https://www.researchgate.net/profile/eleni_ilkou2/publication/345243725_symbolic_vs_subymbolic_ai_methods_friends_or_enemies

Ilvento, C. (2019, June 1). *Metric Learning for Individual Fairness*. http://arxiv.org/pdf/1906.00250v2

Ishiguro, K. (2021). *Klara and the sun* (First edition).faber.

Israelsen, B. W. (2017, August 1). *"I can assure you [$\ldots$] that it's going to be all right" -- A definition, case for, and survey of algorithmic assurances in human-autonomy trust relationships*. https://arxiv.org/pdf/1708.00495

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2020). Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government. *Social Science Computer Review*, 089443932098011. https://doi.org/10.1177/0894439320980118

Janzen, S., Gdanitz, N., Abdel Khaliq, L., Munir, T., Franzius, C., & Maass, W. (2023). *Proceedings of the 56th Annual Hawaii International Conference on System Sciences: January 3-6, 2023*.

https://scholarspace.manoa.hawaii.edu/items/c4c09f4f-a33f-4556-9ad2-b34499787f4c/full https://doi.org/102456

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How can I choose an explainer? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (the, pp. 805–815). ACM. https:/ /doi.org/10.1145/3442188.3445941

Johnson, W. L. (1994). *Agents that Learn to Explain Themselves*. https://www.aaai.org/papers/aaai/1994/aaai94-193.pdf

Jonathan Gillham. (2017). *The macroeconomic impact of artificial intelligence.* https://doi.org/10.13140/RG.2.2.21506.38083

Jones, O., & Gatrell, C. (2014). Editorial: The Future of Writing and Reviewing for IJMR. *International Journal of Management Reviews*, *16*(3), 249–264. https://doi.org/10.1111/ijmr.12038

Kagermann, H.;Wahlster, W.; Helbig, J. (2013): *Umsetzungsempfehlungen für das Zukunftsprojekt Industrie 4.0 ; Deutschlands Zukunft als Produktionsstandort sichern ; Abschlussbericht des Arbeitskreises Industrie 4.0*. Forschungsunion, Berlin, 2013. http://digital.bib-bvb.de/view/bvb_single/single.jsp?dvs=1687257824358~198&locale=de_DE&VIEWER_URL=/view/bvb_single/single.jsp?&DELIVERY_RULE_ID=39&bfe=view/action/singleViewer.do?dvs=&frameId=1&usePid1=true&usePid2=true

Kahn, H., & Wiener, A. J. (1967). *The year 2000; a framework for speculation on the next thirty-three years*. Macmillan.

Kahneman, D [Daniel]. (2013). *Thinking, fast and slow* (First paperback edition). *Psychology/economics*. Farrar Straus and Giroux.

Kahneman, D [Daniel], Slovic, P., & Tversky, A [Amos]. (2018). *Judgment under uncertainty: Heuristics and biases* (28. Auflage). Cambridge University Press.

Kahneman, D [Daniel], & Tversky, A [Amos] (1982). The Psychology of Preferences. *Scientific American*, *246*(1), 160–173. https://doi.org/10.1038/scientificamerican0182-160

Kaplan, R. S. (2008). *The Execution Premium: Linking Strategy to Operations for Competitive Advantage*. Harvard Business Review Press.

Kaplan, R. S., & Norton, D. P. (2009). *The balanced scorecard: Translating strategy into action* [Nachdr.]. Harvard Business School Press.

Kastens, U., & Kleine Büning, H. (2021). *Modellierung: Grundlagen und formale Methoden* (5., aktualisierte Auflage). *Hanser eLibrary*. Hanser. https://www.hanser-elibrary.com/doi/book/10.3139/9783446469563 https://doi.org/10.3139/9783446469563

Katz, M., Sohrabi, S., & Udrea, O. (2020). Top-Quality Planning: Finding Practically Useful Sets of Best Plans. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(06), 9900–9907. https://doi.org/10.1609/aaai.v34i06.6544

Kautz, H. A. (2022). The third AI summer: AAAI Robert S. Engelmore Memorial Lecture. *AI Magazine*, *43*(1), 105–125. https://doi.org/10.1002/aaai.12036

Kejriwal, M. (2021). Link Prediction Between Structured Geopolitical Events: Models and Experiments. *Frontiers in Big Data*, *4*, 779792. https://doi.org/10.3389/fdata.2021.779792

Kenny, E. M., & Keane, M. T. (2021). Explaining Deep Learning using examples: Optimal feature weighting methods for twin systems using post-hoc, explanation-by-example in XAI. *Knowledge-Based Systems*, *233*, 107530. https://doi.org/10.1016/j.knosys.2021.107530

Kepczynski, R., Jandhyala, R., Sankaran, G., & Dimofte, A. (2018). *Integrated Business Planning*. Springer International Publishing. https://doi.org/10.1007/978-3-319-75665-3

Khalil, O. E. M. (1993). Artificial Decision-Making and Artificial Ethics: A Management Concern. *Journal of Business Ethics*, *12*(4), 313–321. http://www.jstor.org/stable/25072403

KI-Gesetz: erste Regulierung der künstlichen Intelligenz | Aktuelles | Europäisches Parlament. (2023). https://www.europarl.europa.eu/news/de/headlines/society/20230601STO93804/ki-gesetz-erste-regulierung-der-kunstlichen-intelligenz

King, P. L. (2019). *Lean for the Process Industries: Dealing with Complexity, Second Edition* (2nd ed.). Productivity Press. https://www.taylorfrancis.com/books/9780429400155

Kirsch, W., & Roventa, P. (Eds.). (1983). *Bausteine eines Strategischen Managements*. De Gruyter. https://doi.org/10.1515/9783110857252

Kirste, S. (Ed.). (2016). *Recht und Philosophie: Band 1*. *Interdisziplinarität in den Rechtswissenschaften: Ein interdisziplinärer und internationaler Dialog*. Duncker & Humblot.

Kitamura, Y., Quigley, A., Isbister, K., Igarashi, T., Bjørn, P., & Drucker, S. (Eds.) (2021). *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. ACM.

Klein, R., Becker, F., Tweardy, J.R., & Schühly, A. (2021). Intelligence gathering: Bringin AI technology into strategic planning. CFO Insights (June 2021). Deloitte.

Klein, R., & Scholl, A. (2011). *Planung und Entscheidung: Konzepte, Modelle und Methoden einer modernen betriebswirtschaftlichen Entscheidungsanalyse* (2. Aufl.). *Vahlens Handbücher der Wirtschafts- und Sozialwissenschaften*. Vahlen.

Knedlová, J., Bílek, O., Sámek, D., & Chalupa, P. (2017). Design and construction of an inspection robot for the sewage pipes. *2261-236X*, *121*, 1006. https://doi.org/10.1051/matecconf/201712101006

Kraus, S. (Ed.) (2019). *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI-19)*. International Joint Conferences on Artificial Intelligence.

Kraus, T., Ganschow, L., Eisenträger, M., & Wischmann, S. (2022). Explainable AI. https://www.digitale-technologien.de/DT/Redaktion/EN/Downloads/Publikation/KI_Inno_Xai_Studie.pdf?__blob=publicationFile&v=1

Kugel, Robert, Ventana Research. (2023). *Integrated Business Planning: Reimagining Planning and Budgeting*. Independent.

Kundisch, D., & Dzoienziol, J. (2008). Decision Support for Financial Planning. *Journal of Decision Systems*, *17*(2), 175–209. https://doi.org/10.3166/jds.17.175-209

Kunisch, S., Menz, M., Bartunek, J. M., Cardinal, L. B., & Denyer, D. (2018). Feature Topic at Organizational Research Methods. *Organizational Research Methods*, *21*(3), 519–523. https://doi.org/10.1177/1094428118770750

Kurakin, A., Goodfellow, I., & Bengio, S. (2016, November 4). *Adversarial Machine Learning at Scale*. http://arxiv.org/pdf/1611.01236v2

Kurzweil, R. (2018). *The singularity is near: When humans transcend biology* (This impression 2018). Duckworth.

Lakkaraju, H., & Bastani, O. (2020). "How do I fool you?". In A. Markham, J. Powles, T. Walsh, & A. L. Washington (Eds.), *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (pp. 79–85). ACM. https://doi.org/10.1145/3375627.3375833

Lamas, D., Sarapuu, H., Šmorgun, I., & Berget, G. (Eds.) (2020). *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society*. ACM.

Lang, J. (Ed.) (2018). *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18): Stockholm, 13-19 July 2018*. International Joint Conferences on Artificial Intelligence.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, 103473. https://doi.org/10.1016/j.artint.2021.103473

Lankhorst, M. (2017). *Enterprise architecture at work: Modelling, communication and analysis* (Fourth edition). *The Enterprise engineering series*. Springer. https://doi.org/10.1007/978-3-662-53933-0

Laurent Moussiegt, Lea Samek, & OECD. (2023). *Oecd Science, Technology and Industry Working Papers* (2023/01). OECD. https://www.oecd-ilibrary.org/science-and-technology/identifying-artificial-intelligence-actors-using-online-data_1f5307e7-en https://doi.org/10.1787/1f5307e7-en

Le Matt. (2019). *Long Papers,, Anna Korhonen, David R.*

Lecun, Y., Bottou, L [L.], Bengio, Y [Y.], & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE, 86*(11), 2278–2324. https://doi.org/10.1109/5.726791

Li, G. (Ed.) (2021). *Proceedings of the 2021 International Conference on Management of Data*. *ACM Digital Library*. Association for Computing Machinery.

Li, M., Plunkett Tost, L., & Wade-Benzoni, K. (2007). The dynamic interaction of context and negotiator effects. *International Journal of Conflict Management*, *18*(3), 222–259. https://doi.org/10.1108/10444060710825981

Liang, T.-P., Robert, L., Sarker, S., Cheung, C. M., Matt, C., Trenz, M., & Turel, O. (2021). Artificial intelligence and robots in individuals' lives: how to align technological possibilities and ethical issues. *Internet Research*, *31*(1), 1–10. https://doi.org/10.1108/INTR-11-2020-0668

Liao, Q. V., Gruen, D., & Miller, S. (2020). Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems.* ACM. https://doi.org/10.1145/3313831.3376590

Liaskos, S., McIlraith, S. A., Sohrabi, S., & Mylopoulos, J. (2011). Representing and reasoning about preferences in requirements engineering. *Requirements Engineering*, *16*(3), 227–249. https://doi.org/10.1007/s00766-011-0129-9

Lin, Y.-S., Lee, W.-C., & Celik, Z. B. (2021). What Do You See? In F. Zhu, B. Chin Ooi, & C. Miao (Eds.), *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining* (pp. 1027–1035). ACM. https://doi.org/10.1145/3447548.3467213

Ling, R. C., & Goddard, W. E. (op. 1988). *Orchestrating success: Improve control of the business with sales & operations planning*. Oliver Wight Ltd. Publications.

Lipton, Z. C. (2016, June 10). *The Mythos of Model Interpretability*. https://arxiv.org/pdf/1606.03490

Liu Qi. (2019). *Hyperbolic Graph Neural Networks. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019.*

Liu, H., Zhong, C., Alnusair, A., & Islam, S. R. (2021). FAIXID: A Framework for Enhancing AI Explainability of Intrusion Detection Results Using Data Cleaning Techniques. *Journal of Network and Systems Management*, *29*(4). https://doi.org/10.1007/s10922-021-09606-8

Loecker, J. de, & Eeckhout, J. (2017). *The Rise of Market Power and the Macroeconomic Implications*. Cambridge, MA. https://doi.org/10.3386/w23687

Lopes, B. G. C. O., Soares, L. S., Prates, R. O., & Gonçalves, M. A. (2021). Analysis of the User Experience with a Multiperspective Tool for Explainable Machine Learning in Light of Interactive Principles. In I. T. Monteiro, K. R. Da Hora Rodrigues, T. de Gois Ribeiro Darin, A. P. Freire, & M. P. Mota (Eds.), *Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems* (pp. 1–11). ACM. https://doi.org/10.1145/3472301.3484360

Lopez-Gazpio, I., Maritxalar, M., Gonzalez-Agirre, A., Rigau, G., Uria, L., & Agirre, E. (2017). Interpretable semantic textual similarity: Finding and explaining differences between sentences. *Knowledge-Based Systems*, *119*, 186–199. https://doi.org/10.1016/j.knosys.2016.12.013

López-Martínez, F., Núñez-Valdez, E. R., García-Díaz, V., & Bursac, Z. (2020). A Case Study for a Big Data and Machine Learning Platform to Improve Medical Decision Support in Population Health Management. *Algorithms*, *13*(4), 102. https://doi.org/10.3390/a13040102

Loureiro, S. M. C., Guerreiro, J., & Tussyadiah, I. (2020). Artificial intelligence in business: State of the art and future research agenda. *Journal of Business Research.* Advance online publication. https://doi.org/10.1016/j.jbusres.2020.11.001

Luhn, H. P. (1958). A Business Intelligence System. *IBM Journal of Research and Development*, *2*(4), 314–319. https://doi.org/10.1147/rd.24.0314

Lundberg, S., & Lee, S.-I. (2017, May 22). *A Unified Approach to Interpreting Model Predictions*. http://arxiv.org/pdf/1705.07874v2

Maartje M.A. de Graaf, & Bertram Malle. (2017). *How people explain action (and AIS should too)*. https://www.researchgate.net/publication/320930548_How_people_explain_action_and_AIS_should_too

Macharzina, K., & Wolf, J. (2023). *Unternehmensführung: Das internationale Managementwissen : Konzepte - Methoden - Praxis* (12., überarbeitete und erweiterte Auflage). Springer Gabler. https://doi.org/10.1007/978-3-658-41053-7

Machi, L. A., & McEvoy, B. T. (2016). *The literature review: Six steps to success* (Third Edition). Corwin.

Magdalena, L. (2019). Semantic interpretability in hierarchical fuzzy systems: Creating semantically decouplable hierarchies. *Information Sciences*, *496*, 109–123. https://doi.org/10.1016/j.ins.2019.05.016

Makarius, E. E., Mukherjee, D., Fox, J. D., & Fox, A. K. (2020). Rising with the machines: A sociotechnical framework for bringing artificial intelligence into the organization. *Journal of Business Research*, *120*, 262–273. https://doi.org/10.1016/j.jbusres.2020.07.045

Mangalathu, S., Hwang, S.-H., & Jeon, J.-S. (2020). Failure mode and effects analysis of RC members based on machine-learning-based SHapley Additive exPlanations (SHAP) approach. *Engineering Structures*, *219*, 110927. https://doi.org/10.1016/j.engstruct.2020.110927

Manu, J. (2022). *Modern TIME SERIES FORECASTING WITH PYTHON: Explore industry-ready time series forecasting using modern machine learning and deep learning* (1st edition). PACKT PUBLISHING LIMITED. https://learning.oreilly.com/library/view/-/9781803246802/?ar

Manyika, J. (2017). *A future that works: automation, employment, and productivity*. McKinsey Global Institute. https://www.voced.edu.au/content/ngv:75268

March, J. G. (1987). Ambiguity and accounting: The elusive link between information and decision making. *Accounting, Organizations and Society*, *12*(2), 153–168. https://doi.org/10.1016/0361-3682(87)90004-3

March, J. G. (2013). *Handbook of Organizations*. *Routledge Library Editions: Organizations*. Taylor and Francis.

March, S. T., & Smith, G. F. (1995). Design and natural science research on information technology. *Decision Support Systems*, *15*(4), 251–266. https://doi.org/10.1016/0167-9236(94)00041-2

Marcin SZCZEPANSKI. (2019). *Economic impacts of artificial intelligence (AI)*. EPRS: European Parliamentary Research Service. https://policycommons.net/artifacts/1334867/economic-impacts-of-artificial-intelligence-ai/1940719/

Marcus, G. (2020, February 14). *The Next Decade in AI: Four Steps Towards Robust Artificial Intelligence*. https://arxiv.org/pdf/2002.06177

Marcus, G., & Davis, E. (2021). Insights for AI from the human mind. *Communications of the ACM*, *64*(1), 38–41. https://doi.org/10.1145/3392663

Markham, A., Powles, J., Walsh, T., & Washington, A. L. (Eds.) (2020). *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. ACM.

Markin, S. (2021). *Sap Integrated Business Planning: Functionality and Implementation* (3rd ed.). Rheinwerk Publishing Inc. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6884699

Masthoff, J., Herder, E., Tintarev, N., & Tkalčič, M. (Eds.) (2021). *Adjunct Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization*. ACM.

Matta, R. de, & Miller, T [Tan] (2018). A Strategic Manufacturing Capacity and Supply Chain Network Design Contingency Planning Approach. In *2018 IEEE Technology and Engineering Management Conference (TEMSCON): 28 June-1 July 2018* (pp. 1–6). IEEE. https://doi.org/10.1109/TEMSCON.2018.8488401

Mazzanti, S. (2020, April 1). SHAP Values Explained Exactly How You Wished Someone Explained to You. *Towards Data Science*. https://towardsdatascience.com/shap-explained-the-way-i-wish-someone-explained-it-to-me-ab81cc69ef30

McCarthy, J., & Hayes, P. (1969). *Some philosophical problems from the standpoint of artificial intelligence*. https://philpapers.org/rec/MCCSPP

McCorduck, P. (2018). *Machines who think: A personal inquiry into the history and prospects of artificial intelligence*. *An A K Peters book*. CRC Press.

McCulloch, W. S., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *The Bulletin of Mathematical Biophysics*, *5*(4), 115–133. https://doi.org/10.1007/BF02478259

McGrath, S. K., & Whitty, S. J. (2018). Accountability and responsibility defined. *International Journal of Managing Projects in Business*, *11*(3), 687–707. https://doi.org/10.1108/IJMPB-06-2017-0058

McGuinness, D. L., & Da Pinheiro Silva, P. (2004). Explaining answers from the Semantic Web: the Inference Web approach. *Journal of Web Semantics*, *1*(4), 397–413. https://doi.org/10.1016/j.websem.2004.06.002

McKinsey Global Institute. (2018). *Notes from the AI frontier: modelling the impact of ai on the world economy.* McKinsey Global Institute. https://t.ly/7myh

McKinsey Global Institute. (2021). *The state of AI in 2021*. McKinsey Global Institute. https://www.mckinsey.com/capabilities/quantumblack/our-insights/global-survey-the-state-of-ai-in-2021

Mennatallah El-Assady, Wolfgang Jentner, Rebecca Kehlbeck, Udo Schlegel, & Daniel Keim (2019). Towards XAI: Structuring the Processes of Explanations. In *Human-Centered Machine Learning Perspectives Workshop*. https://www.researchgate.net/profile/mennatallah-el-assady/publication/332802468_towards_xai_structuring_the_processes_of_explanations

Menzel, C., & Winkler, C. (2019). *Zur Diskussion der Effekte Künstlicher Intelligenz in der wirtschaftswissen-schaftlichen Literatur*.

Meske, C., Bunde, E., Schneider, J., & Gersch, M. (2020). Explainable Artificial Intelligence: Objectives, Stakeholders, and Future Research Opportunities. *Information Systems Management*, 1–11. https://doi.org/10.1080/10580530.2020.1849465

Michalski, R. S. (1983). A THEORY AND METHODOLOGY OF INDUCTIVE LEARNING. *Machine Learning*, 83–134. https://doi.org/10.1016/B978-0-08-051054-5.50008-X

Miller, T [Tim] (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T [Tim], Howe, P., & Sonenberg, L. (2017, December 2). *Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*. https://arxiv.org/pdf/1712.00547

Minsky, M. Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy.

Mittelstadt, B., Russell, C., & Wachter, S. (2019). Explaining Explanations in AI. In *Proceedings of the Conference on Fairness, Accountability, and Transparency* (pp. 279–288). ACM. https://doi.org/10.1145/3287560.3287574

Mittelstraß, J., Blasche, S., & Wolters, G. (Eds.). (1995). *Enzyklopädie Philosophie und Wissenschaftstheorie* (Korrigierter Nachdr). Metzler.

Molnar, C. (2019). *Interpretable machine learning: A guide for making Black Box Models interpretable*. Lulu.

Monopolkommission (2018). Algorithms and collusion: Excerpt from Chapter I of the XXII. Biennial Report of the Monopolies Commission (“Competition 2018”) in accordance with Section 44 Paragraph 1 Sentence 1 of the German Act against Restraints of Competition.

Montavon, G., Samek, W., & Müller, K.-R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing*, *73*, 1–15. https://doi.org/10.1016/j.dsp.2017.10.011

Monteiro, I. T., Da Hora Rodrigues, K. R., Gois Ribeiro Darin, T. de, Freire, A. P., & Mota, M. P. (Eds.) (2021). *Proceedings of the XX Brazilian Symposium on Human Factors in Computing Systems*. ACM.

Montero-Manso, P., & Hyndman, R. J. (2020, August 2). *Principles and Algorithms for Forecasting Groups of Time Series: Locality and Globality*. https://arxiv.org/pdf/2008.00444

Moore, J. D., & Paris, C. L. (1991). Requirements for an expert system explanation facility. *Computational Intelligence*, *7*(4), 367–370. https://doi.org/10.1111/j.1467-8640.1991.tb00409.x

Mössner, U. (1982). *Planung flexibler Unternehmensstrategien*. Zugl.: München, Techn. Univ., Diss., 1981. *Hochschulschriften zur Betriebswirtschaftslehre: Vol. 3*. Florentz.

Murzin, D. (2022). *Chemical reaction technology* (2nd edition). *De Gruyter graduate*. De Gruyter. https://www.degruyter.com/isbn/9783110712520

Nagpal, K., Foote, D., Tan, F., Liu, Y [Yun], Chen, P.-H. C., Steiner, D. F., Manoj, N., Olson, N., Smith, J. L., Mohtashamian, A., Peterson, B., Amin, M. B., Evans, A. J., Sweet, J. W., Cheung, C., van der Kwast, T., Sangoi, A. R., Zhou, M., Allan, R., . . . Mermel, C. H. (2020). Development and Validation of a Deep Learning Algorithm for Gleason Grading of Prostate Cancer From Biopsy Specimens. *JAMA Oncology*, *6*(9), 1372–1380. https://doi.org/10.1001/jamaoncol.2020.2485

Nakagawa, E. Y., Guessi, M., Maldonado, J. C., Feitosa, D., & Oquendo, F. (2014). Consolidating a Process for the Design, Representation, and Evaluation of Reference Architectures. In *2014 IEEE/IFIP Conference on Software Architecture (WICSA 2014): Sydney, Australia, 7 - 11 April 2014* (pp. 143–152). IEEE. https://doi.org/10.1109/WICSA.2014.25

Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., & Ritz, A. (2020). Transforming the Language of Life. In *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics.* ACM. https://doi.org/10.1145/3388440.3412467

Naser, M. Z. (2021). An engineer's guide to eXplainable Artificial Intelligence and Interpretable Machine Learning: Navigating causality, forced goodness, and the false perception of inference. *Automation in Construction*, *129*, 103821. https://doi.org/10.1016/j.autcon.2021.103821

Newell, A., & Simon, H. A [H. A.]. (1988). GPS, A PROGRAM THAT SIMULATES HUMAN THOUGHT. In *Readings in Cognitive Science* (pp. 453–460). Elsevier. https://doi.org/10.1016/b978-1-4832-1446-7.50040-6

Newell, A., & Simon, H. A [Herbert Alexander]. (2019). *Human problem solving*. Echo Point Books & Media.

Nilsson, N. J. (1979). *Problem-solving methods in artifical intelligence: Tab* (11. print). *McGraw-Hill computer science series*. McGraw-Hill.

Niu, N., Da Xu, L., & Bi, Z. (2013). Enterprise Information Systems Architecture—Analysis and Evaluation. *IEEE Transactions on Industrial Informatics*, *9*(4), 2147–2154. https://doi.org/10.1109/tii.2013.2238948

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction*, *27*(3-5), 393–444. https://doi.org/10.1007/s11257-017-9195-0

OECD Productivity Working Papers. (2016). https://doi.org/10.1787/63629cc9-en

Oecd Science, Technology and Industry Policy Papers (Vol. 101). (2021). OECD. https://www.oecd-ilibrary.org/science-and-technology/management-skills-and-productivity_007f399e-en https://doi.org/10.1787/007f399e-en

Oecd Science, Technology and Industry Policy Papers (Vol. 103). (2021). OECD. https://www.oecd-ilibrary.org/science-and-technology/the-role-of-innovation-and-human-capital-for-the-productivity-of-industries_197c6ae9-en https://doi.org/10.1787/197c6ae9-en

Oecd Science, Technology and Industry Policy Papers (Vol. 120). (2021). OECD. https://www.oecd-ilibrary.org/science-and-technology/the-human-capital-behind-ai_2e278150-en https://doi.org/10.1787/2e278150-en

Oecd Science, Technology and Industry Policy Papers (Vol. 120). (2021). OECD. https://www.oecd-ilibrary.org/science-and-technology/the-human-capital-behind-ai_2e278150-en https://doi.org/10.1787/2e278150-en

Oecd Science, Technology and Industry Working Papers (2021/05). (2021). OECD. https://www.oecd-ilibrary.org/science-and-technology/burning-glass-technologies-data-use-in-policy-relevant-analysis_cd75c3e7-en https://doi.org/10.1787/cd75c3e7-en

OECD, & Standards, Centre for the Study of Living. (2017). *International Productivity Monitor*. OECD Publishing; Centre for the Study of Living Standards. https://doi.org/10.1787/9789264279179-en

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y [Yoshua]. (2019, May 24). *N-BEATS: Neural basis expansion analysis for interpretable time series forecasting*. https://arxiv.org/pdf/1905.10437

Pal Singh Toor, T., & Dhir, T. (2011). Benefits of integrated business planning, forecasting, and process management. *Business Strategy Series*, *12*(6), 275–288. https://doi.org/10.1108/17515631111185914

Paredes, J. N., Teze, J. C. L., Simari, G. I., & Martinez, M. V. (2021). *On the Importance of Domain-specific Explanations in AI-based Cybersecurity Systems (Technical Report).*

Pasquale Minervini, Sebastian Riedel, Pontus Stenetorp, Edward Grefenstette, & Tim Rocktäschel (2020). Learning Reasoning Strategies in End-to-End Differentiable Proving. *International Conference on Machine Learning*, 6938–6949. https://proceedings.mlr.press/v119/minervini20a.html

Pearl, J. (2009). Causal inference in statistics: An overview. *Statistics Surveys*, *3*(none). https://doi.org/10.1214/09-SS057

Pearl, J. (2009). *Causality: Models, reasoning and inference* (Second edition, reprinted with corrections 2021). Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pearl, J. (2018). *Theoretical Impediments to Machine Learning With Seven Sparks from the Causal Revolution.*

Pearl, J. (2019). The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, *62*(3), 54–60. https://doi.org/10.1145/3241036

Pearl, J. (2019). *The Book of Why: The New Science of Cause and Effect. Pearl, Judea, and Dana Mackenzie. 2018. Hachette UK.* (2018). http://bayes.cs.ucla.edu/why/jmde-why-review2018.pdf

Pelachaud, C., Martin, J.-C., Buschmeier, H., Lucas, G., & Kopp, S. (Eds.) (2019). *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents*. ACM.

Pereira, D. F., Oliveira, J. F., & Carravilla, M. A. (2020). Tactical sales and operations planning: A holistic framework and a literature review of decision-making models. *International Journal of Production Economics*, *228*, 107695. https://doi.org/10.1016/j.ijpe.2020.107695

Pescholl, A. (2011). *Adaptive Entwicklung eines Referenzmodells für die Geschäftsprozessunterstützung im technischen Großhandel* (1. Auflage, digitale Originalausgabe). GRIN Verlag.

Peterson, G. D., Cumming, G. S., & Carpenter, S. R. (2003). Scenario Planning: a Tool for Conservation in an Uncertain World. *Conservation Biology*, *17*(2), 358–366. https://doi.org/10.1046/j.1523-1739.2003.01491.x

Petropoulos, G. Artificial intelligence: how to get the most from the labour-productivity boost: Analysis 02/2023.

Petropoulos, G., Marcus, J. S., Moës, N., & Bergamini, E. (2019). *Digitalisation and European welfare states*. *Bruegel blueprint series: volume 30*. Bruegel. https://bruegel.org/2019/07/digitalisation-and-european-welfare-states/

Pino, R., Fernández, I., La Fuente, D. de, Parreño, J., & Priore, P. (2010). Supply chain modelling using a multi-agent system. *Journal of Advances in Management Research*, *7*(2), 149–162. https://doi.org/10.1108/09727981011084968

Poli, R., & Seibt, J. (Eds.). (2010). *Theory and applications of ontology*. *Philosophical perspectives*. Springer. https://doi.org/10.1007/978-90-481-8845-1

Pretorius, L., & Pretorius, M. (Eds.) (2021). *Mot for the world of the future: 30th Annual Conference of the International Association for Management of Technology (IAMOT 2021) : Cairo, Egypt, 19-23 September 2021*. Curran Associates Inc.

Problemfelder und Entwicklungstendenzen der Planungspraxis. (1983). In W. Kirsch & P. Roventa (Eds.), *Bausteine eines Strategischen Managements* (pp. 309–354). De Gruyter. https://doi.org/10.1515/9783110857252-017

Publishing, O. (2017). *Oecd digital economy outlook 2017*. OECD Publishing. https://doi.org/10.1787/9789264276284-en

Raphael, B. (1976). *The thinking computer: Mind inside matter*. *A series of books in psychology*. Freeman.

Reidt, A., Pfaff, M., & Krcmar, H. (2018). Der Referenzarchitekturbegriff im Wandel der Zeit. *HMD Praxis Der Wirtschaftsinformatik*, *55*(5), 893–906. https://doi.org/10.1365/s40702-018-00448-8

Reussner, R., & Hasselbring, W. (Eds.). (2009). *Handbuch der Software-Architektur* (2., überarbeitete und erweiterte Auflage). dpunkt.verlag.

Rezaei, M., Chaharsooghi, S. K., Kashan, A. H., & Babazadeh, R. (2020). A new approach based on scenario planning and prediction methods for the estimation of gasoil consumption. *International Journal of Environmental Science and Technology*, *17*(6), 3241–3250. https://doi.org/10.1007/s13762-019-02583-1

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, February 16). *"Why Should I Trust You?": Explaining the Predictions of Any Classifier*. https://arxiv.org/pdf/1602.04938

Rich, E. (1988). *Artificial intelligence* (Internat. ed., 8. print). *McGraw-Hill series in artificial intelligence*. McGraw-Hill.

Richard E. Fikes, & Nils J. Nilsson (1971). Strips: A new approach to the application of theorem proving to problem solving. *Artificial Intelligence*, *2*(3), 189–208. https://doi.org/10.1016/0004-3702(71)90010-5

Rohlfing, K. J. (2021). Under Co-construction: Toward the Social Design of Explainable AI Systems. In C. Schulte, B. A. Becker, M. Divitini, & E. Barendsen (Eds.), *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1* (p. 1). ACM. https://doi.org/10.1145/3430665.3460419

Roodman. (2020). *On the probability distribution of long-term changes in the growth rate of the global economy: An outside view*. https://www.openphilanthropy.org/wp-content/uploads/modeling-the-human-trajectory.pdf

Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, *65*(6), 386–408. https://doi.org/10.1037/h0042519

Rosenfeld, A., & Richardson, A. (2020). Wy, Who, What, When and How about Explainability in Human-Agent Systems, 2161–2163.

Rousseau, D. M., Manning, J., & Denyer, D. (2008). 11 Evidence in Management and Organizational Science: Assembling the Field's Full Weight of Scientific Knowledge Through Syntheses. *Academy of Management Annals*, *2*(1), 475–515. https://doi.org/10.5465/19416520802211651

Rowley, J., & Slack, F. (2004). Conducting a literature review. *Management Research News*, *27*(6), 31–39. https://doi.org/10.1108/01409170410784185

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Pinguin.

Russell, S. J., & Norvig, P. (2012). *Künstliche Intelligenz: Ein moderner Ansatz* (3., aktualisierte Aufl.). *it Informatik*. Pearson.

Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th Global Edition).

Ryan, M., & Stahl, B. C. (2021). Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications. *Journal of Information, Communication and Ethics in Society*, *19*(1), 61–86. https://doi.org/10.1108/JICES-12-2019-0138

Rynes, S. L., & Bartunek, J. M. (2017). Evidence-Based Management: Foundations, Development, Controversies and Future. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*(1), 235–261. https://doi.org/10.1146/annurev-orgpsych-032516-113306

Sachan, S., Yang, J.-B., Xu, D.-L., Benavides, D. E., & Li, Y. (2020). An explainable AI decision-support-system to automate loan underwriting. *Expert Systems with Applications*, *144*, 113100. https://doi.org/10.1016/j.eswa.2019.113100

Sallaba, M., Esser, R., Fach, P., & Salzmann, O. (2020). State of AI in Enterprise (3rd. ed.).

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., & Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning* (Vol. 11700). Springer International Publishing. https://doi.org/10.1007/978-3-030-28954-6

Santana-Mancilla, P. C., & Valderrama, E. (Eds.) (2019). *Proceedings of the IX Latin American Conference on Human Computer Interaction*. ACM.

SAP Deutschland AG & Co. KG. (2009). *Innovative Gestaltung von Geschäftsprozessen in der Prozessindustrie: Marktumfeld, Herausforderungen, Vorgehensmodell, Praxisbeispiele, Handlungsempfehlungen* (1. Aufl.). *SAP*. dpunkt-Verl.

Saunders, M. (2023). *Research Methods for Business Students* (9th ed.). Pearson. https://elibrary.pearson.de/book/99.150005/9781292402734

Schank, R. C. (1986). *Explanation patterns: Understanding mechanically and creatively* (Online-Ausg). L. Erlbaum Associates. http://site.ebrary.com/lib/alltitles/Doc?id=10751992

Schelter, S., & Stoyanovich, J. (2020). Taming Technical Bias in Machine Learning Pipelines. *Bulletin of the Technical Committee on Data Engineering*, *43*(4). https://par.nsf.gov/biblio/10287316

Schneeweiss, C. (2010). *Distributed Decision Making* (2., nd ed. Softcover version of original hardcover edition 2003). Springer Berlin.

Schnell, R., Hill, P. B., & Esser, E. (2022). *Methoden der empirischen Sozialforschung* (12., aktualisierte und erweiterte Auflage). *De Gruyter Studium*. De Gruyter Oldenbourg. https://www.degruyter.com/isbn/9783110752991

Schoemaker, P. (1995). Scenario Planning: A Tool for Strategic Thinking. *Sloan Management Review*, *36*, 25–40.

Schoemaker, P. J. H., & Russo, J. E. (2019). Decision-Making. In M. Augier & D. J. Teece (Eds.), *Springer eBook Collection. The Palgrave Encyclopedia of Strategic Management* (pp. 1–5). Palgrave Macmillan. https://doi.org/10.1057/978-1-349-94848-2_341-1

Schön, D. A. (1983). *The reflective practitioner: How professionals think in action*. Basic Books. http://www.loc.gov/catdir/enhancements/fy0832/82070855-d.html

Schulte, C., Becker, B. A., Divitini, M., & Barendsen, E. (Eds.) (2021). *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*. ACM.

Schutt, R., & O'Neil, C. (2014). *Doing data science*. O'Reilly Media. https://search.ebscohost.com/login.aspx?direct=true&scope=site&db=nlebk&db=nlabk&AN=650570

Schwaiger, R. (Ed.). (2019). *Neuronale Netze programmieren mit Python* (1. Auflage). Rheinwerk Computing. https://www.rheinwerk-verlag.de/neuronale-netze-programmieren-mit-python/?v=4590

Seghrouchni, A. E. F., Sukthankar, G., Liu, T.-Y., & van Steen, M. (Eds.) (2020). *Companion Proceedings of the Web Conference 2020*. ACM.

Sekaran, U. and Bougie, R. (2016) *Research Methods for Business A Skill-Building Approach. 7th Edition, Wiley & Sons, West Sussex. - References - Scientific Research Publishing.* (2023, June 18). https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/referencespapers.aspx?referenceid=2371540

Selmi, M. H., Jemai, Z., Gregoire, L., & Dallery, Y. (2021). Integrated Business Planning Process: Link Between Supply Chain Planning and Financial Planning. In A. Dolgui, A. Bernard, D. Lemoine, G. von Cieminski, & D. Romero (Eds.), *Springer eBook Collection: Vol. 632. Advances in Production Management Systems. Artificial Intelligence for Sustainable and Resilient Production Systems: Ifip WG 5.7 International Conference, APMS 2021, Nantes, France, September 5–9, 2021, Proceedings, Part III* (1st ed., Vol. 632, pp. 149–158). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-85906-0_17

Shah, C., Belkin, N. J., Byström, K., Huang, J., & Scholer, F. (Eds.) (2018). *Chiir'18: Proceedings of the 2018 Conference on Human Information Interaction and Retrieval : March 11-15, 2018, New Brunswick, NJ, USA*. ACM Association for Computing Machinery.

Sharda, R., Delen, D., & Turban, E. (2020). *Analytics, data science, & artificial intelligence: Systems for decision support* (Eleventh edition). Pearson.

Shiebler, D. (2023, June 6). *Understanding Neural Networks with Layerwise Relevance Propagation and Deep Taylor Series*. https://danshiebler.com/2017-04-16-deep-taylor-lrp/

Short, J. (2009). The Art of Writing a Review Article. *Journal of Management*, *35*(6), 1312–1317. https://doi.org/10.1177/0149206309337489

Shrestha, Y. R., Ben-Menahem, S. M., & Krogh, G. von (2019). Organizational Decision-Making Structures in the Age of Artificial Intelligence. *California Management Review*, *61*(4), 66–83. https://doi.org/10.1177/0008125619862257

Shrestha, Y. R., Krishna, V., & Krogh, G. von (2021). Augmenting organizational decision-making with deep learning algorithms: Principles, promises, and challenges. *Journal of Business Research*, *123*, 588–603. https://doi.org/10.1016/j.jbusres.2020.09.068

Silver, D. L. On_Common_Ground_Neural_Symbolic_Integration_And_Lifelong_Machine_Learning, *August 2013*.

Simon, H. A [Herbert A.]. (1957). *Models of man; social and rational: Models of man; social and rational*. Wiley.

Simon, H. A [Herbert A.] (1977). What Computers Mean for Man and Society. *Science*, *195*(4283), 1186–1191. http://www.jstor.org/stable/1743571

Simon, H. A [Herbert Alexander]. (1978). *The sciences of the artificial* (Paperback ed., 6. print). *Karl Taylor Compton lectures: Vol. 1968*. M.I.T. Pr.

Simon, H. A [Herbert A.] (1980). The Behavioral and Social Sciences. *Science*, *209*, 72–78. Th

Simon, H. A [Herbert Alexander]. (1996). *The sciences of the artificial* (3. ed. [Nachdr.]. MIT Press.

Simon, H. A [Herbert A.]. (1997). *Models of bounded rationality*. MIT Press.

Simon, H. A [Herbert A.]. (2019). *The sciences of the artificial* (Third edition [2019 edition]. The MIT Press. https://doi.org/10.7551/mitpress/12107.001.0001

Simsek, S., Albizri, A., Johnson, M., Custis, T., & Weikert, S. (2021). Predictive data analytics for contract renewals: a decision support tool for managerial decision-making. *Journal of Enterprise Information Management*, *34*(2), 718–732. https://doi.org/10.1108/JEIM-12-2019-0375

Singh, R., Vatsa, M., & Ratha, N. (2021). Trustworthy AI. In J. Haritsa, S. Roy, M. Gupta, S. Mehrotra, B. V. Srinivasan, & Y. Simmhan (Eds.), *8th ACM IKDD CODS and 26th COMAD* (pp. 449–453). ACM. https://doi.org/10.1145/3430984.3431966

Sohrabi, S., Katz, M., Hassanzadeh, O., Udrea, O., & Feblowitz, M. D. (2018). Ibm Scenario Planning Advisor: Plan Recognition as AI Planning in Practice. In J. Lang (Ed.), *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18): Stockholm, 13-19 July 2018* (pp. 5865–5867). International Joint Conferences on Artificial Intelligence. https://doi.org/10.24963/ijcai.2018/864

Sohrabi, S., Udrea, O., Riabov, A., & Hassanzadeh, O. (2020). Interactive Planning-Based Hypothesis Generation with LTS+ +. In M. Vallati & D. Kitchin (Eds.), *Springer eBook Collection. Knowledge Engineering Tools and Techniques for AI Planning* (1st ed., pp. 189–207). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-38561-3_10

Solow R. (1987). We'd better watch out. *New York Times Book Review*, *36*. https://cir.nii.ac.jp/crid/1571698599544098816

Sowa, J. (2006). *Concept mapping*. http://www.jfsowa.com/talks/cmapping.pdf

Sowa, J. F. (2010). The Role of Logic and Ontology in Language and Reasoning. In R. Poli & J. Seibt (Eds.), *Theory and applications of ontology. Philosophical perspectives* (pp. 231–263). Springer. https://doi.org/10.1007/978-90-481-8845-1_11

Spiegelhalter, D. J., & Knill-Jones, R. P. (1984). Statistical and Knowledge-Based Approaches to Clinical Decision-Support Systems, with an Application in Gastroenterology. *Journal of the Royal Statistical Society. Series a (General)*, *147*(1), 35. https://doi.org/10.2307/2981737

Spinner, T., Schlegel, U., Schafer, H., & El-Assady, M. (2020). Explainer: A Visual Analytics Framework for Interactive and Explainable Machine Learning. *IEEE Transactions on Visualization and Computer Graphics*, *26*(1), 1064–1074. https://doi.org/10.1109/tvcg.2019.2934629

Stachowiak, H. (1983). *Kritische Information: Vol. 101*. *Modelle, Konstruktion der Wirklichkeit*. Fink. http://nbn-resolving.de/urn/resolver.pl?urn=urn:nbn:de:bvb:12-bsb00048243-2

Statistisches Bundesamt (Ed.). *Klassifikationen: Gliederung der Klassifikation der Wirtschaftszweige, Ausgabe 2008 (WZ 2008)*. Statistisches Bundesamt.

Stegmüller, W. (1977). 'The Problem of Causality'. In *Collected Papers on Epistemology, Philosophy of Science and History of Philosophy* (pp. 26–43). Springer, Dordrecht. https://doi.org/10.1007/978-94-010-1132-7_2

Stepin, I., Alonso-Moral, J. M., Catala, A., & Pereira-Fariña, M. (2022). An empirical study on how humans appreciate automated counterfactual explanations which embrace imprecise information. *Information Sciences*, *618*, 379–399. https://doi.org/10.1016/j.ins.2022.10.098

Stock, W. G. (2013). *Handbook of Information Science*. *Knowledge and Information*. Walter de Gruyter GmbH. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=893252

Stone, M., Aravopoulou, E., Ekinci, Y., Evans, G., Hobbs, M., Labib, A., Laughlin, P., Machtynger, J., & Machtynger, L. (2020). Artificial intelligence (AI) in strategic marketing decision-making: a research agenda. *The Bottom Line*, *33*(2), 183–200. https://doi.org/10.1108/BL-03-2020-0022

Stuart E. Dreyfus, & Hubert L. Dreyfus. (1980). *A Five-Stage Model of the Mental Activities Involved in Directed Skill Acquisition*. https://www.researchgate.net/publication/235125013_A_Five-Stage_Model_of_the_Mental_Activities_Involved_in_Directed_Skill_Acquisition

Sufi, F. K. (2022). AI-Tornado: An AI-based Software for analyzing Tornadoes from disaster event dataset. *Software Impacts*, *14*, 100357. https://doi.org/10.1016/j.simpa.2022.100357

Suresh, H., Gomez, S. R., Nam, K. K., & Satyanarayan, A. (2021). Beyond Expertise and Roles: A Framework to Characterize the Stakeholders of Interpretable Machine Learning and their Needs. In Y. Kitamura, A. Quigley, K. Isbister, T. Igarashi, P. Bjørn, & S. Drucker (Eds.), *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (pp. 1–16). ACM. https://doi.org/10.1145/3411764.3445088

Suresh, H., & Guttag, J. A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle. *EAAMO*, 1–9. https://doi.org/10.1145/3465416.3483305

Squicciarini, M. & OECD. (2022). *Oecd Science, Technology and Industry Working Papers* (2022/06). OECD. https://www.oecd-ilibrary.org/science-and-technology/identifying-and-characterising-ai-adopters_154981d7-en https://doi.org/10.1787/154981d7-en

Surie, G. (2021). Digital TECHNOLOGIES AND THE WORLD OF THE FUTURE: Implications FOR THE MANAGEMENT OF TECHNOLOGY. In L. Pretorius & M. Pretorius (Eds.), *Mot for the world of the future: 30th Annual Conference of the International Association for Management of Technology (IAMOT 2021) : Cairo, Egypt, 19-23 September 2021* (pp. 173–189). Curran Associates Inc. https://doi.org/10.52202/060557-0012

Swartout, W. R., & Moore, J. D. (1993). Explanation in Second Generation Expert Systems. In J.-M. David, J.-P. Krivine, & R. Simmons (Eds.), *Second Generation Expert Systems* (pp. 543–585). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-77927-5_24

Takeuchi, H., Ito, Y., Nishiyama, R., & Isomura, T. (2021). Modeling of Machine Learning Projects Using ArchiMate. In (pp. 222–231). Springer, Singapore. https://doi.org/10.1007/978-981-16-3264-8_21

Teo, T. S., & King, W. R. (1996). Assessing the impact of integrating business planning and IS planning. *Information & Management*, *30*(6), 309–321. https://doi.org/10.1016/S0378-7206(96)01076-2

Teso, S., & Kersting, K. (2019). Explanatory Interactive Machine Learning. In V. Conitzer, G. Hadfield, & S. Vallor (Eds.), *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 239–245). ACM. https://doi.org/10.1145/3306618.3314293

Thampi, A. (2022). *Interpretable AI: Building Explainable Machine Learning Systems*. Manning. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=7030591

Thayyib, P. V., Mamilla, R., Khan, M., Fatima, H., Asim, M., Anwar, I., Shamsudheen, M. K., & Khan, M. A. (2023). State-of-the-Art of Artificial Intelligence and Big Data Analytics Reviews in Five Different Domains: A Bibliometric Summary. *Sustainability*, *15*(5), 4026. https://doi.org/10.3390/su15054026

The Use of Explanations in Knowledge-Based Systems: Cognitive Perspectives and a Process-Tracing Analysis (2000). *Journal of Management Information Systems*, *17*(2), 153–179. https://doi.org/10.1080/07421222.2000.11045646

Thellman, S., & Ziemke, T. (2021). The Perceptual Belief Problem. *ACM Transactions on Human-Robot Interaction*, *10*(3), 1–15. https://doi.org/10.1145/3461781

Tiddi, I [I.]. (2016). *Explaining data patterns using knowledge from the web of data*. https://search.proquest.com/openview/0931db29b2638bbb146b35364da38a1f/1?pq-origsite=gscholar&cbl=51922

Tiddi, I [Ilaria], Bastianelli, E., Bardaro, G., d'Aquin, M., & Motta, E. (2017). An ontology-based approach to improve the accessibility of ROS-based robotic systems. In *Proceedings of the Knowledge Capture Conference.* ACM. https://doi.org/10.1145/3148011.3148014

Tiddi, I [Ilaria], Lécué, F [Freddy], & Hitzler, P [Pascal] (Eds.). (2020). *Studies on the semantic web: volume 047. Knowledge graphs for explainable artificial intelligence: Foundations, applications and challenges*. IOS Press; AKA.

Tiddi, I [I.], Lécué, F [F.], & Hitzler, P [P.]. (2020). *Knowledge graphs for eXplainable artificial intelligence: Foundations, applications and challenges*. *Studies on the semantic web: v. 047*. IOS Press.

Tiddi, I [Ilaria], & Schlobach, S. (2022). Knowledge graphs as tools for explainable machine learning: A survey. *Artificial Intelligence*, *302*, 103627. https://doi.org/10.1016/j.artint.2021.103627

Toorajipour, R., Sohrabpour, V., Nazarpour, A., Oghazi, P., & Fischl, M. (2021). Artificial intelligence in supply chain management: A systematic literature review. *Journal of Business Research*, *122*, 502–517. https://doi.org/10.1016/j.jbusres.2020.09.009

Tranfield, D., Denyer, D., & Smart, P [Palminder] (2003). Towards a Methodology for Developing Evidence-Informed Management Knowledge by Means of Systematic Review. *British Journal of Management*, *14*(3), 207–222. https://doi.org/10.1111/1467-8551.00375

Tripicchio, P., & D'Avella, S. (2020). Is Deep Learning ready to satisfy Industry needs? *Procedia Manufacturing*, *51*, 1192–1199. https://doi.org/10.1016/j.promfg.2020.10.167

Tsichritzis, D. (1997). The Dynamics of Innovation. In *Beyond Calculation* (pp. 259–265). Springer, New York, NY. https://doi.org/10.1007/978-1-4612-0685-9_19

Tushman, M. L., & Nadler, D. A. (1978). Information Processing as an Integrating Concept in Organizational Design. *Academy of Management Review*, *3*(3), 613–624. https://doi.org/10.5465/amr.1978.4305791

Tversky, A [A.], & Kahneman, D [D.] (1974). Judgment under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124

V. Samawi, M. Akram, & A. Abeer (2013). Arabic expert system shell. *Int. Arab J. Inf. Technol.* https://www.semanticscholar.org/paper/Arabic-expert-system-shell-Samawi-Akram/5687b93b9e05dafd57b009e11aa1f5f1021c14e1

Vallati, M., & Kitchin, D. (Eds.). (2020). *Springer eBook Collection*. *Knowledge Engineering Tools and Techniques for AI Planning* (1st ed. 2020). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-38561-3

van Lamsweerde, A. (2013). *Requirements engineering: From system goals to UML models to software specifications* (Repr). Wiley.

van Leeuwen, C., Elprama, S. A., Jacobs, A., Heyman, R., Pierson, J., & Duysburgh, P. (2020). Unethically Me: Explaining Artificial Intelligence's Results by Being Unethical. In D. Lamas, H. Sarapuu, I. Šmorgun, & G. Berget (Eds.), *Proceedings of the 11th Nordic Conference on Human-Computer Interaction: Shaping Experiences, Shaping Society* (pp. 1–3). ACM. https://doi.org/10.1145/3419249.3420065

van Lent, M., & Fisher, W. (2004). *An explainable artificial intelligence system for small-unit tactical behavior*. https://www.aaai.org/papers/iaai/2004/iaai04-019.pdf

Varian, H. (2019). Artificial Intelligence, Economics, and Industrial Organization. In A. Agrawal, J. Gans, & A. Goldfarb (Eds.), *National Bureau of Economic Research conference report. The economics of artificial intelligence: An agenda* (pp. 399–422). The University of Chicago Press. https://doi.org/10.7208/chicago/9780226613475.003.0016

Vasconcelos. A., Sousa P., & Tribolet, J. (2005). *Information System Architecture Evaluation: From Software to Enterprise Level Approaches*. https://www.researchgate.net/profile/jose-tribolet/publication/228967139_information_system_architecture_evaluation_from_software_to_enterprise_level_approaches

VDE Verband der Elektrotechnik Elektronik Informationstechnik e.V. (2020). *Entwicklung und Vertrauenswürdigkeit von autonom/kognitiven Systemen* (Draft - application reference VDE -AR-E 2842-61-1:2020-07). VDE. https://www.dke.de/de/normen-standards/dokument?id=7141809&type=dke%7Cdokument

Vetrò, A., Santangelo, A., Beretta, E., & Martin, J. C. de (2019). AI: from rational agents to socially responsible agents. *Digital Policy, Regulation and Governance*, *21*(3), 291–304. https://doi.org/10.1108/DPRG-08-2018-0049

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

Vitruvius. (2001). *Ten books on architecture* (I. D. Rowland, Trans.) (1. paperback ed.). Cambridge Univ. Press.

Vivacqua, A. S., Stelling, R., Garcia, A. C. B., & Gouvea, L. C. (2019). Explanations and sensemaking with AI and HCI. In P. C. Santana-Mancilla & E. Valderrama (Eds.), *Proceedings of the IX Latin American Conference on Human Computer Interaction* (pp. 1–4). ACM. https://doi.org/10.1145/3358961.3359004

vom Brocke, J. (2003). *Referenzmodellierung: Gestaltung und Verteilung von Konstruktionsprozessen*. Zugl.: Münster, Univ., Diss., 2002. *Advances in information systems and management science: Vol. 4*. Logos-Verl. http://bvbr.bib-bvb.de:8991/F?func=service&doc_library=BVB01&doc_number=010411452&line_number=0002&func_code=DB_RECORDS&service_type=MEDIA

vom Brocke, J., Hevner, A., & Maedche, A. (Eds.). (2020). *Springer eBook Collection*. *Design Science Research. Cases* (1st ed. 2020). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-46781-4

vom Brocke, J., Hevner, A., & Maedche, A. (2020). Introduction to Design Science Research. In J. vom Brocke, A. Hevner, & A. Maedche (Eds.), *Springer eBook Collection. Design Science Research. Cases* (1st ed., pp. 1–13). Springer International Publishing; Imprint Springer. https://doi.org/10.1007/978-3-030-46781-4_1

vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., & Cleven, A. (2015). Standing on the Shoulders of Giants: Challenges and Recommendations of Literature Search in Information Systems Research. *Communications of the Association for Information Systems*, *37*. https://doi.org/10.17705/1CAIS.03709

W Flores, A., Bechtel, K., & Lowenkamp, C. (2016). False Positives, False Negatives, and False Analyses: A Rejoinder to “Machine Bias: There’s Software Used Across the Country to Predict Future Criminals. And it’s Biased Against Blacks.”. *Federal Probation*, *80*.

Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, *7*(2), 76–99. https://doi.org/10.1093/idpl/ipx005

Wack, P. (1985). Shhoting the rapids. *Harvard Business Review*, *63*(6), 139–150.

Wallkötter, S., Tulli, S., Castellano, G., Paiva, A., & Chetouani, M. (2021). Explainable Embodied Agents Through Social Cues. *ACM Transactions on Human-Robot Interaction*, *10*(3), 1–24. https://doi.org/10.1145/3457188

Walls, J. G., Widmeyer, G. R., & El Sawy, O. A. (1992). Building an Information System Design Theory for Vigilant EIS. *Information Systems Research*, *3*(1), 36–59. https://doi.org/10.1287/isre.3.1.36

Wamba-Taguimdje, S.L., Fosso Wamba, S., Kala Kamdjoug, J. R., & Tchatchouang Wanko, C. E. (2020). Influence of artificial intelligence (AI) on firm performance: the business value of AI-based transformation projects. *Business Process Management Journal*, *26*(7), 1893–1924. https://doi.org/10.1108/BPMJ-10-2019-0411

Wang, J.Z., Hsieh, S.T., & Hsu, P.Y. (2012). Advanced sales and operations planning framework in a company supply chain. *International Journal of Computer Integrated Manufacturing*, *25*(3), 248–262. https://doi.org/10.1080/0951192X.2011.629683

Wang, G., Liu, X., Wang, Z., & Yang, X. (2020). Research on the influence of interpretability of artificial intelligence recommendation system on users' behavior intention. In *Proceedings of the 2020 4th International Conference on Electronic Information Technology and Computer Engineering* (pp. 762–766). ACM. https://doi.org/10.1145/3443467.3443850

Wang, D., Yang, Q., Abdul, A., & Lim, B. Y. (2019). Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.* ACM. https://doi.org/10.1145/3290605.3300831

Weitz, K., Schiller, D., Schlagowski, R., Huber, T., & André, E. (2019). "Do you trust me?". In C. Pelachaud, J.-C. Martin, H. Buschmeier, G. Lucas, & S. Kopp (Eds.), *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (pp. 7–9). ACM. https://doi.org/10.1145/3308532.3329441

Welge, M. K. (2017). *Strategisches Management: Grundlagen - Prozess - Implementierung* (7., überarbeitete und aktualisierte Auflage). *SpringerLink Bücher*. Springer Gabler. https://doi.org/10.1007/978-3-658-10648-5

Wiegand, G., Schmidmaier, M., Weber, T., Liu, Y [Yuanting], & Hussmann, H. (2019). I Drive - You Trust. In S. Brewster, G. Fitzpatrick, A. Cox, & V. Kostakos (Eds.), *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1–6). ACM. https://doi.org/10.1145/3290607.3312817

Wieringa, R. J. (2014). *Design Science Methodology for Information Systems and Software Engineering*. Springer Berlin Heidelberg. http://nbn-resolving.org/urn:nbn:de:bsz:31-epflicht-1614141

Wild, J. (1980). *Grundlagen der Unternehmungsplanung* (4th ed.). *Wv Studium Ser: v.26*. VS Verlag fur Sozialwissenschaften GmbH. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=6673700

Wilde, T., & Hess, T. (2007). Forschungsmethoden der Wirtschaftsinformatik. *WIRTSCHAFTSINFORMATIK*, *49*(4), 280–287. https://doi.org/10.1007/s11576-007-0064-z

Willms, P., & Brandenburg, M. (2019). Emerging trends from advanced planning to integrated business planning. *IFAC-PapersOnLine*, *52*(13), 2620–2625. https://doi.org/10.1016/j.ifacol.2019.11.602

Winograd, T. (2004). Procedures as a Representation for Data in a Computer Program for Understanding Natural Language.

Winston, P. H. (1979). *Artificial intelligence* (2. printing). Addison-Wesley.

Wittgenstein, L. (2014). *Tractatus Logico-Philosophicus: German and English*. *International library of psychology, philosophy, and scientific method*. Taylor and Francis. https://ebookcentral.proquest.com/lib/kxp/detail.action?docID=1694408

Wolf, C. T. (2019). Explainability scenarios. In W.-T. Fu, S. Pan, O. Brdiczka, P. Chau, & G. Calvary (Eds.), *Proceedings of the 24th International Conference on Intelligent User Interfaces* (pp. 252–257). ACM. https://doi.org/10.1145/3301275.3302317

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., Platen, P. v., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., . . . Rush, A. M. (2019, October 9). *HuggingFace's Transformers: State-of-the-art Natural Language Processing*. http://arxiv.org/pdf/1910.03771v5

Wraith, S. M., Aikins, J. S., Clancey, W. J., Fagan, L. M., van Melle, W. J., Buchanan, B. G., Davis, R., Scott, A. C., Shortliffe, E. H., Axline, S. G., Hannigan, J. F., Yu, V. L., & Cohen, S. N. (1976). Computerized consultation system for selection of antimicrobial therapy. *American Journal of Health-System Pharmacy*, *33*(12), 1304–1308. https://doi.org/10.1093/ajhp/33.12.1304

Yang, T., Yi, X., Lu, S., Johansson, K. H., & Chai, T. (2021). Intelligent Manufacturing for the Process Industry Driven by Industrial Artificial Intelligence. *Engineering*, *7*(9), 1224–1230. https://doi.org/10.1016/j.eng.2021.04.023

Zhang, Y., Hong, D., McClement, D., Oladosu, O., Pridham, G., & Slaney, G. (2021). Grad-CAM helps interpret the deep learning models trained to classify multiple sclerosis types using clinical brain magnetic resonance imaging. *Journal of Neuroscience Methods*, *353*, 109098. https://doi.org/10.1016/j.jneumeth.2021.109098

Zhao, T., Huang, Y., Yang, S., Luo, Y., Feng, J., Wang, Y., Yuan, H., Pan, K., Li, K., Li, H., & Zhu, F [Fu] (2019). Mathgraph: A Knowledge Graph for Automatically Solving Mathematical Exercises. In (pp. 760–776). Springer, Cham. https://doi.org/10.1007/978-3-030-18576-3_45

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2015, December 14). *Learning Deep Features for Discriminative Localization*. https://arxiv.org/pdf/1512.04150

Zhu, F [Feida], Chin Ooi, B., & Miao, C. (Eds.) (2021). *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. ACM.

GLOSSARY

TERM	DEFINITION
A	
Artificial Intelligence	The study of how to make computers do intelligent things that only people can do (until now)
Asset utilization	Optimising fixed assets and reducing working capital
B	
Balanced scorecard	Tool/ methodology in strategic management and strategic planning
Black-box model	Intransparent AI or especially machine learning model
Business Analytics	The process of transforming data into insights to improve business decisions
Business Intelligence	The process of transforming data into information and information into knowledge that can be used to increase a company's competitiveness
C	
Capital efficiency	Optimising fixed assts and reduce working capital
Comprehensibility	Making sense of the message no matter how it is conveyed
Corporate Planning	Scenario planning and integrated corporate planning
D	
Decision Support Systems	Systems to support decision-making, which include models, data manipulation and the ability to handle uncertainty and risk
Decomposability	Breaking up into independent modules
Deep learning	Machine learning technique that teaches computers to do what comes naturally to humans: learn by example
Deep Neural Networks	An artificial neural network (ANN) with multiple layers between the input and output layers. There are different types of neural networks but they always consist of the same components: neurons, synapses, weights, biases, and functions
Definition of Accountability	An assurance that an individual or organisation is evaluated on its performance or behavior related to something for which it is responsible. The term is related to responsibility but is regarded more from the perspective of oversight
E	
Expert System late	A computer program that uses artificial intelligence (AI) technologies to simu- the judgment and behavior of a human or an organisation that has expertise and experience in a particular field.
Explainable Artificial Intelligence	Artificial intelligence in which humans can understand the reasoning behind decisions or predictions made by the AI; also known as *understandable artificial intelligence* or *explainable machine learning*
F	
Feedback loops	The part of a system in which some portion (or all) of the system's output is used as input for future operations

G

Governance	All the practices, processes and policies that help one guide a business in the right direction
Glass-box model	A model that is transparent to the user, in which all the features and the model parameters are known to the user.

H

Homo oeconomicus	The concept of the individual assuming that man, as a rational being, always seeks to maximise profits and make choices for the economic value (utility) of the results of those choices
Hybrid approach	Taking two (or more) different project management methodologies and combining them to create an entirely new method and give a clear roadmap to the team with roles and responsibilities, deadlines, and expectations

I

Integrated Business Planning	A cloud-based supply chain planning solution that scales to accommodate business growth and integrates with other systems
Interpretability	The extent to which a cause and effect can be observed within a system

K

Knowledge Based Systems	Computer programs that use a centralised repository of data known as a knowledge base to provide a method for problem-solving

L

Layer Relevance Propagation	A technique that brings such explainability and scales to potentially highly complex deep neural networks, which operates by propagating the prediction backward in the neural network, using a set of purposely designed propagation rules

M

Machine Learning	A subset of artificial intelligence which focuses on teaching computers how to learn from data and improve as they gain experience
MYCIN	One of the first expert systems for the diagnostic treatment of bacterial infections

N

N/A	Not applicable

O

Opacity	Lack of transparency

P

Predictive Analytics	A branch of advanced analytics that makes predictions about future outcomes using historical data combined with statistical modeling, data mining techniques and machine learning
Prescriptive Analytics	The use of advanced processes and tools to analyse data and content to recommend the optimal course of action or strategy moving forward
Process Industries	Those industries where the primary production processes are either continuous, or occur on a batch of materials that is indistinguishable

PYTHON	A high-level, general-purpose programming language

R

Re-Fish	Reference architecture built as a composition in honour of Marian Rejewski, the leading Polish scientist who solved the Enigma code and Babelfish,
Reference architecture	The field of software architecture provides a template solution for an architecture for a domain
Reference model	An abstract framework or domain-specific ontology consisting of an interlinked set of clearly defined concepts produced by an expert or body of experts to encourage clear communication
Reliability	The quality of being dependable, trustworthy, or of performing consistently well
Robustness	The capability of performing without failure under a wide range of possible conditions

S

Scenario Planning	Making assumptions on what the future is going to be and how s business environment will change overtime in light of that future
SHapley Additive exPlanations	A method to explain individual predictions
Simulatability	The condition of being, or the extent to which something can be simulated
Stage Gate Model	A value-creating business process and risk model designed to quickly and profitably transform an organization's best new ideas into new outcomes
Standard Model	The model which includes members of several classes of elementary particles, which in turn can be distinguished by other characteristics, such as color change

T

Transparency	*In a business or governance context* → being open and honest.

U

Understandability	The concept that a system is presented in a way that can be easily comprehended to make operations become a straightforward process

APPENDIX

Appendix A- Presentation for Architecture Evaluation

1. Introduction and Presentation to the Topics

This part of the survey was about introducing the topic. A PowerPoint presentation was shown (see additional files- DVD - Evaluation presentation) - the first two pages of this presentation introduced to the topic and motivation and relevance- The architectural views of the Re_fish reference architecture was then presented. Afterwards, the participants were asked to answer the questionnaires. The author was always available for questions or explanations.

2. Objectives of the Interview

The objectives of the evaluation of the architecture have already been presented in chapter 5.3. The aim is not to evaluate the quality of the individual systems, but rather the projection of the architecture in terms of its effectiveness and impact on the application or instantiation for a specific architecture.

3. Methodology

In addition to the Powerpoint presentation mentioned above, the questionnaire was sent by using MS Forms. The questions of the questionnaire are presented in Appendix B.

4. Approach

a) Introduction;

b) Explanation of the scope and the objectives of the research;

c) Sending out the questionnaire via MS Forms about the quality of the Re_fish reference architecture;

d) Summary, open questions and explanation of next steps.

Appendix B- Survey Questions for Architecture Evaluation

1.Name
2.Profession
3.Company
4.No. of years' experience with corporate planning and decision making
5.No. of years' experience with IT architecture
6.No. of years with Business Analytics
7.No. of years with Artificial Intelligence
8.No. of years with Project Management

Reference architecture quality attribute 1: Usability

The reference architecture is presented in two different ways - once with the modelling language Archimate and once as a simple diagram type. - This presentation is for information purposes. You can leave a comment below. All relevant information (diagrams, excerpts etc. can be found under the link below) - Slide Deck Complete Overview of Re_fish Reference Architecture:
https://sync.luckycloud.de/f/3d519a96aaa940d58f29/

Archimate:

https://pubs.opengroup.org/architecture/archimate3-doc/ch-Technology-Layer.html

Feel free to add any comments - here and in the following
U1: After the reference architecture was explained to you based on the description, you were able to easily understand the architecture model.

U2: The model clearly defines all four levels of architecture (no separate view for data architecture) when considered as a blueprint for future instantiation for implementation.

U3: If you select one of the business actors or stakeholders, you can see which business processes it accesses and which application components. and which application components. Slide 10- 14

U4: The business process steps are easy to follow. Slide 13

U5: The explanations are presented to the user in a human-readable form. The user also has the possibility to ask further follow-up questions, similar to a ChatBot. Do you rate this type of interface as sufficient for the use case?. If no, please comment, what is missing - Slide 17, 18

U6: One of the findings of the thesis was, that causal explanations- following the "ladder of causality" are sufficient in planning situations (other situations maybe require different explanations, of course). Therefore,

the explanations are also displayed for the user in a causality diagram with cause-and-effect chains in a tree diagram. This diagram adapts to the further questions or builds up accordingly. Do you rate this type of interface as sufficient for the application? If no, please comment, what is missing. - Slide 16

U7: The explanations of the machine learning models also include the presentation of the results of the respective explanatory model used (LIME, SHAP, ELI5 etc.) Do you rate this type of presentation as sufficient for the application? If no, please comment, what is missing. - Slide 14

Quality Attribute 2: Performance

In this section, the reference architecture is evaluated in terms of performance requirements such as system responsiveness. As this is a reference architecture, the assessment can therefore only be transformative - in the sense of anticipating the properties to be evaluated in the "instantiated" architecture/implementation.

P1: The reference architecture uses a Knowledge Graph data bank (e.g., NEO4J) as a knowledge base. This knowledge base contains the semantic data used in addition to the respective data of the subsymbolic model and also the data provenance. In addition, the results of the subsymbolic explainer (LIME, ELI5, etc.) are also stored here. this means that the explainer module provides a comprehensive explanation in the event of queries to the system. - Slide 14, 15

Do you think this approach is sufficient? If no, please comment briefly.

P2: One of the main components of Re_Fish is the Data Services component. This module is about collecting data from different pre-systems - streaming data, structured data, unstructured data, etc. - and putting it into context so that it can be used for explanations. For example, in scenario planning. - Slide 15- 18

Do you think this approach is sufficient? If no, please comment briefly.

P3: One of the outstanding features of Re_Fish is the interaction of symbolic and non-symbolic AI to provide the user with a comprehensive explanation that also allows the data provenance to be shown. - Slide 17

Do you think this approach is sufficient? If no, please comment briefly.

Quality Attribute 3: Reliability

This section evaluates the reference system architecture from the perspective of reliability requirements, such as fault tolerance, recoverability, overall data reliability.

R1: One of the results of the thesis is that the explainability of an AI must already be guaranteed in the design and throughout the entire life cycle. In addition to the architecture, this also includes comprehensive lifecycle management. A component for transparency is also the tracker component, which makes it possible to track the status of the non-symbolic machine learning model and to detect deviations if necessary. By separating

development, testing and production, it can be ensured that no biased model or its results end up in production.- Slide 15

Quality Attribute 5: Security (and Compliance)

This section evaluates the reference system architecture from the perspective of security requirements, such as user roles and authorizations.

S1: In the architecture model, user roles can be clearly distinguished. - Slide 13- 15

S2: One of Re_Fish's key requirements is to address society's growing concerns in AI by ensuring ethical principles and compliance with regulations and standards (GDPR) throughout the lifecycle of the AI model. One of the main components of Re_Fish to ensure this requirement is the use of a separate audit module that contains secure, proprietary (i.e., separation of concerns) access to all information, logging (tracker), metadata etc. of the AI model, but also of Re_Fish itself. - Slide 14,15

Quality Attribute 6: Functionality

This section evaluates the reference system architecture from the perspective of functionality requirements, such as alignment with business needs, interoperability, integration.

F1: The architecture properly describes the application and infrastructure components that support the explainability of the AI models. - Slide 13

F2: The dialogue components of the auditor and the business user are separated. the business users are divided into different groups, knowledge engineer, planner etc. The users can thus be assigned to groups via their roles, which then receive the necessary authorisations. The activities of the users are recorded in the tracker. Slide 15- 18

Quality Attribute 7: Modifiability

In this section, the reference architecture is evaluated from the point of view of modifiability with regard to changed requirements and, for example, with regard to cost- and time-effective changes.

Q1: The reference architecture can be used flexibly to derive an instance for specific use cases and thus enables implementation through appropriate adaptation (instantiation) to changing business requirements or even completely different situational contexts.

Q2: It can be assumed that the architecture can be implemented in the foreseeable future using existing technologies and reusing existing components.

Q3: It is, assuming that one does not have to start completely from scratch, that the architecture will be implemented at a foreseeable cost.

STATUTORY DECLARATION

I declare that I have authored this thesis independently, that I have not used other than the declared sources / resources, and that I have explicitly marked all material which has been quoted either literally or by content from the used sources.