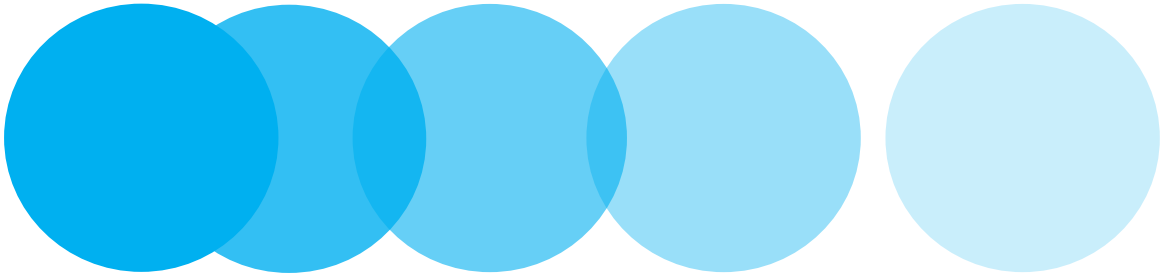


Design guidelines to protect stakeholders' values in AI systems

Based on a use case situated in the Japanese life insurance industry

Shan Amin

TU Delft

Design guidelines to protect stakeholders' values in AI systems

Based on a use case situated in the Japanese life insurance Industry

Thesis report

by

Shan Amin

to obtain the degree of Master of Science
at the Delft University of Technology
to be defended publicly on 26th of January 2024

Thesis committee:

Chair:	Dr. Y.A. (Aaron) Ding
First supervisor:	Dr.mr. S. (Sander) Renes
Second supervisor:	Dr.ir. R.I.J. (Roel) Dobbe

Place:	Faculty of Technology, Policy and Management, Delft
Project Duration:	August, 2023 - January, 2024
Student number:	5391504

Acknowledgement

As I present this master's thesis, "Design guidelines to protect stakeholders' values in AI systems - Based on a use case in the Japanese life insurance industry," I am filled with a sense of accomplishment. This work is a pivotal part of my journey in the MSc Complex Systems Engineering and Management program at Delft University of Technology. It represents the culmination of my academic pursuits and a significant personal and professional growth chapter.

It was an extraordinary opportunity to conduct this research, particularly in Japan. Hence, I express my heartfelt gratitude to the organization that provided me with this opportunity. The experience of working in such a dynamic and innovative environment was nothing short of amazing. It offered me a unique perspective on the practical applications of my research, enriched by Japan's cultural and technological advancements.

My gratitude extends to my colleagues, whose assistance and collaboration have been instrumental throughout this journey. Their willingness to help whenever needed, coupled with their insightful perspectives and shared knowledge, greatly contributed to the depth and quality of my research.

I owe a special debt of gratitude to my supervisor, Sander Renes. Thank you for your guidance, the enlightening feedback, and the brainstorming sessions that helped refine my ideas. Your expertise and encouragement were priceless in navigating the complexities of my research topic. I want to thank my second supervisor, Roel Dobbe, for the meetings, valuable feedback, and innovative ideas that greatly enhanced my work. Your insights have been a guiding light in this academic endeavor. And to Aaron Ding, my chair of the committee, your feedback and oversight have significantly contributed to the rigor and quality of this thesis.

Embarking on this topic was both challenging and exciting. The complex interplay of AI, risk management, and value-driven design in the Japanese life insurance industry presented a unique challenge. However, the process was also incredibly rewarding. The opportunity to delve into such a dynamic field, combining technical complexity with real-world implications, was a truly enriching experience. It allowed me to apply my academic knowledge and grow as a researcher and professional.

Accomplishing this thesis also means the end of a chapter of my time at TU Delft. The challenges faced, the knowledge gained, and the relationships forged during this process have shaped me in ways I had not anticipated. I hope this thesis will contribute meaningfully to the field and perhaps, in some small way, pave the path for future exploration and innovation in AI risk management, design for values, and socio-technical system design.

The last thing I want to acknowledge is that a few sentences are refined with AI-based language tools to support clarity and conciseness.

I hope you will enjoy reading it!

Shan Amin,

January 8th, 2024

Executive summary

Integrating artificial intelligence (AI) into Japan's life insurance sector marks a significant move towards data-centric precision, reflecting the nation's shift towards Society 5.0. In this domain, AI is revolutionizing decision-making processes and enhancing operational efficiency, with applications ranging from fraud detection in credit card systems to predictive underwriting. While AI offers notable benefits, it also introduces risks, including privacy breaches, algorithmic biases, and inadequate human supervision. The Japanese government underscores these concerns and initiates guidelines and frameworks for societal protection. However, a gap persists in the insurance industry's adoption of these protective frameworks, particularly in identifying and implementing social norms. In addition, this lack of clear guidelines in high-stakes domains, such as the life insurance sector, poses the following problem: identifying which social norms need protection and how to integrate them into AI systems to protect stakeholders' values. Neglecting these norms risks reputational damage and presents a complex socio-technical dilemma, placing a substantial responsibility on AI system developers.

A socio-technical approach is needed to address the problem associated with AI systems in Japan's life insurance domain. Such an approach is needed as it looks into the development and use of AI systems and how it affects the contextual environment. In addition, the approach considers stakeholders' diverse perspectives and concerns on AI, whose perceptions of harm are deeply intertwined with their norms, values, and rules. To effectively resolve these challenges, the industry requires a guide to identify and protect values from various stakeholder viewpoints. This contributes practically to the safe design, development, and deployment of AI systems with a continuous improvement strategy. Before a guide can be designed, two knowledge gaps must first be addressed. One is missing a value framework where little research has been done to translate high-level values into Japanese life insurance industry requirements. The second scientific knowledge gap concerns a lack of a standard process for translating identified values into practical organizational guidelines.

The deliverable of this research is a design guide that can be applied throughout the lifecycle of an AI system to protect stakeholders' values. From the conception of an AI initiative to its operational deployment, a combination of design science research and design for values is chosen to answer the following question:

What design guidelines can developers in Japan's life insurance domain follow to control AI systems while protecting stakeholders' values?

To answer the main question, first, the environment was explored using a use case, namely predictive underwriting, guiding the research through its various stages. Exploring the environment resulted in goals and constraints guiding the selection and composition of the value framework and design guidelines. This first step highlights the need for integrating laws and stakeholder input to define the goals and restrictions of a system. Next, using different information sources and methods, 13 values were identified from society, legal and industry, and business perspectives. The research used an integrative literature review to understand the societal view, text, and content; a comparison analysis to draw an understanding of the legal and industry documents; and a questionnaire to understand the business view. The following 13 values emerged from the analysis: *trust, explainability, understandability, transparency, privacy, security, robustness, fairness, usability, accountability and responsibility, effectiveness, and continuous improvement.* In addition, the four informal social institutions gained from the integrative literature review must be considered during the design process of the value framework and guidelines, namely *contributing to community, wholeness, sincerity, and sensitivity.* These 13 values and four social institutions collectively create the framework's foundation.

Next, an empirical study is conducted to create norms for each value. Eight participants completed questionnaires to identify the values they wanted to protect from their expert role. Next, workshops are conducted to identify the risks and constraints of the AI system. Following the workshops, semi-structured interviews are conducted to gather information missing from the workshops. Concepts from system safety theory supported these methods and served to identify the norms for a risk-based approach. Elements

used for system safety were the safety constraints and the hierarchical safety control structure. These efforts resulted in 54 norms, which underwent a four-step refinement process to make them manageable in an organizational context. These steps were adjusted from Garst et al.'s (2022) steps for selecting relevant topics. The steps for converging the norms are:

1. **Define:** Norms are defined from the empirical data collection with the support of the system safety concepts of Leveson (2012).
2. **Categorization:** Norms are categorized as 23 process- or 31 assessment-focused. The framework of Mäntymäki et al. (2022) provided structure for the IT infrastructure, which resulted in 18 data and 13 AI assessment norms.
3. **Determining the information sources:** Selecting information sources representing relevant stakeholder criteria who are interested in the AI system. The selection resulted in three stakeholder criteria to examine the assessment norms.
4. **Selecting relevant topics:** After examining the stakeholder criteria to the assessment norms, 11 data and 10 AI assessment norms remained, which must undergo the next phase, namely setting design requirements.

In the next research phase, it is shown how design requirements were created. This phase highlighted the need for expert consultations to understand how norms can be actionable and controllable in an organizational setting while defining scope, method, and objective. The combination of these design requirements led to the formulation of control mechanisms for specific norms, with consideration of potential trade-offs.

The design guidelines were created through backward traceability, chronologically reflecting the findings per research phase. In addition, the combination of the steps taken in this research and the identified 23 process norms is input for the following ten design guidelines, validated by experts:

- **DG 1:** Create a multi-disciplinary team to set up the project and go through the control cycle.
- **DG 2:** Define the scope and business purpose.
- **DG 3:** Identify the metric(s) for the effectiveness of the AI system.
- **DG 4:** Identify stakeholder criteria and assess the design.
- **DG 5:** Assess the design of the system.
- **DG 6:** Review assessment with experts and set requirements
- **DG 7:** Assess the development of the system.
- **DG 8:** Review assessments with experts and set control mechanisms.
- **DG 9:** Monitor risks through communication channels
- **DG 10:** Create continuous improvement within the AI system and its governance.

The research results in guidelines that foster safe socio-technical AI systems' design, development, and deployment, driven by continuous improvement. Following these guidelines protects stakeholders' values, as they are integrated into the guideline design. The guidelines are an effort to combine socio-cultural influences from society, institutions provided by authorities, and expert knowledge from an organizational point of view. The significance of these design guidelines extends to addressing scientific knowledge gaps and industry needs. In addressing the problem statement, the practical contribution lies in guiding the identification of risks, ensuring safety, and fostering continuous improvement. In this sense, safety means safeguarding the specified values from business, societal, and stakeholder views. This guide is tailored for AI system developers, offering clear directives on actions, stakeholder engagement, timing, and assessment criteria throughout the AI lifecycle to protect stakeholders' values.

This research contributes to the scientific knowledge base by addressing the scientific knowledge gaps. The first scientific knowledge gap is that there is little research on consolidating these high-level principles and requirements into the Japanese industry context. This research provides a value framework based on the predictive underwriting use case, which can be built upon through further research. The second scientific knowledge gap is the lack of a standard process to translate high-level values into practical guidelines. This research contributes by introducing an initial process that could be repeated and improved in similar AI use cases. The initial process consists of a combination of design for values, complemented

by system safety concepts, reporting standardization steps, and an AI governance concept. The research provides empirical evidence that these concepts are useful for identifying and implementing social values through a safety lens in an organizational context.

This study highlights several possibilities for further research and development. First, since the current guidelines are founded on a single-use case, it is recommended to replicate this process across various use cases. Replicating the process would facilitate gradual adjustments to the guidelines and reflect on the concepts from the literature used. The second recommendation is to reevaluate the value framework, incorporating perspectives from a broader range of stakeholders. As the method and guidelines are empirically tested with internal stakeholders from the organization, they lack a broader perspective of agents, customers, legal entities, and society. Adding these stakeholder's perspectives in an empirical setting could create a complex challenge but provide a more robust value framework and a better-focused process. Third, exploring and delineating the interaction dynamics between Japanese society and AI systems is recommended. The value framework provides the first steps to build further research on it. Last, this research recommends extending the examination of system theoretic hazards analysis, which is part of system safety. This research only focused on the institutional constraints and hazards, missing an in-depth analysis of the technical constraints and hazards. Further research would provide a concrete analysis of the conflicts that could arise within the organizational setting and offer possibilities to explore trade-offs.

Contents

Nomenclature

AI	Artificial Intelligence
AML	Anti-Money Laundering
APPI	Act on Protection of Personal Information
CAS	Corporate audit service
D&AI	Data & Artificial Intelligence
DfV	Design for Values
DSR	Design science research
EIOPA	European Insurance and Occupational Pensions Authority
EU	European Union
FERM	Financial risk management
FP	False Positive
FPR	False Positive Rate
FSA	Financial Service Agency
GIAJ	General Insurance Association of Japan
HR	Human Resource
IRM	Information risk management
LIAJ	Life insurance association of Japan
MCDA	Multi-criteria decision-making analysis
ML	Machine Learning
OECD	Organization for Economic Cooperation and Development
ORM	Operational risk management
PPV	Predictive Parity
PU	Predictive Underwriting
SME	Small medium enterprise
SP	Social Principles of human-centric AI stated by the Japanese government
TN	True Negative
TP	True Positive

List of Figures

List of Tables

Part I

Problem definition

1

Introduction

Integrating artificial intelligence (AI) in the Japanese insurance industry transforms the sector's approach towards data-driven precision, aligning with Japan's general transition to Society 5.0 (Maier et al., 2020; Mourtzis et al., 2022). This transition implies that AI systems could reshape the industry landscape, excelling beyond human accuracy in decision-making and streamlining workflows with remarkable efficiency (Alzubi et al., 2018; Ruf & Detyniecki, 2021). These AI systems are pivotal in high-stakes insurance processes, navigating the complexities of data-intensive tasks. The utilization of this technology spans multiple domains, including high-frequency sales trading, credit card system monitoring for fraudulent claims, predictive underwriting, and cybersecurity threat detection (Khambatta et al., 2021). Implementing AI systems can enhance decision-making processes and improve efficiency, accuracy, and cost-effectiveness.

Despite these great opportunities, recent research highlights major AI risk issues, such as privacy concerns, harm through biased systems, and a lack of human oversight (Katirai, 2023; Mamiko, 2020; Wirtz et al., 2022). The Japanese government has acted in response to the development of AI as a first step towards safeguarding society from potential risks. Moreover, the Japanese government published the Social Principles of Human-Centric AI (Social Principles) guidelines for AI implementation in society (Habuka, 2023). However, this framework has not been reflected in the insurance industry's local norms and values, reflecting that the life insurance industry does not fully integrate AI within business processes within a high-stake domain (Burston et al., 2020; Miyashita, 2016; Radu, 2021). Also, the current framework and guidelines are open to multiple interpretations within the socio-technical context of the insurance industry, which leaves room for high-risk exposure. In summary, this guideline shows that Japan has taken steps towards highlighting AI risks. However, the insurance industry has not yet adopted these value frameworks, which remains a significant issue.

The absence of guidelines and lack of understanding of the social norms in high-stake domains in Japan, such as the life insurance industry, raises the problem of not knowing what social norms to safeguard and how to safeguard them. The Rikunabi case underscores the need for Japanese organizations to protect social norms within the AI landscape. Rikunabi is a large job-seeking platform operated by a Japanese company, Recruitment Career. This platform utilized algorithmic scores to predict the probability of job applicants declining employment offered and selling this information to employers (Cyphers & Rodriguez, 2021). Despite acting within legal bounds, the Japanese public and regulators deemed the actions unethical, resulting in significant reputational harm, renaming the organization, and rebuilding its operations and business from scratch. The Rikunabi case highlights the strong prevailing culture of informal social rules, which are highly regarded in every sector. It highlights the urge for organizations in Japan, such as the life insurance industry, to understand what social norms to safeguard and how to translate them into the practical socio-technical environment. In addition, the design and control of AI systems pose a multi-faceted socio-technical challenge that places responsibility on AI system developers (Ruf & Detyniecki, 2021).

The problem statement of this thesis is the challenge of determining which social norms should be protected and implemented to control the technical operations of AI systems in organizational settings. Literature provides ways to tackle this issue. Van de Poel (2013) demonstrates from a philosophical perspective how to identify social values and translate them into design requirements. Nevertheless, this approach lacks the translation from high-level values into practical guidelines from a protection perspective (Aizenberg &

Van den Hoven, 2020). Leveson (2012) presents safety concepts that offer possibilities for translating values into controls. Using this approach results in identifying a wide range of potential risks. Furthermore, Garst et al. (2022) propose several steps to converge the identified topics, in our case risks, into a contextually relevant selection that is controllable for organizations. Mäntymäki et al. (2022) established an AI governance framework to contextualize and incorporate these points within the organization. However, academic literature lacks a standard process for translating identified values into practical organizational guidelines. Therefore, this thesis offers an initial setup for identifying and translating social values into the operational activities of the Japanese insurance industry. The initial setup focuses on AI systems' design, development, and deployment stages to safeguard these social values.

In this thesis, we use a use case to shed light on when an AI system is (not) complying with social values norms in its perceived socio-technical context. To control AI systems according to social norms, we need to take two steps: identifying and operationalizing values in an organizational context. First, we need to create a framework that fits the specific context. This creation can be supported by using existing concepts from the traditional "design for values" theory to build a value framework. Second, we must ensure that AI developers can use this framework in an organizational setting. By taking these two steps, this thesis provides three contributions. The first scientific contribution is a repeatable and improvable process to identify social norms and implement controls in the technical operations of AI systems to protect stakeholders' values. The second scientific and practical contribution involves establishing an initial value framework specifically tailored for the Japanese life insurance sector, which can serve as a foundation for subsequent research and development. Last, the practical contribution is guidelines for AI system developers to use through the AI lifecycle to identify risks that could harm stakeholders' values.

1.1. Context

The following sub-chapters motivate the need for this study by providing more relevant context.

1.1.1. The use case: Predictive underwriting in the Japanese life insurance industry

Japanese insurance organizations are navigating through the complexities of integrating AI into decision-making. The complexity is particularly apparent due to their responsibility for handling sensitive data that can significantly impact customers (Burston et al., 2020). This research utilizes a use case from a Japanese life insurance organization that offers small and medium enterprises (SMEs) life insurance products, such as bereavement support. The use case illustrates the challenges of managing sensitive customer data and the potential direct effects on SME operations.

The link between the protection of stakeholders' values and the use case becomes evident in the underwriting process for SMEs. This process requires a systematic socio-technical approach to balance innovative AI applications with societal, regulatory, and business expectations (Bossen, 2018). The socio-technical approach explains the interactions among various technical, process, and institutional components. The predictive underwriting (PU) system, which evaluates application risks, is the use case as it showcases the tension between leveraging AI for efficiency and accuracy, which must align with diverse stakeholder values while controlling potential risks and regulatory compliance (Mullins et al., 2021). The use case sheds light on what norms are (not) accepted according to stakeholders' values and how to control AI systems within the practical organizational environment. The details of the PU system are described in Chapter 3.

1.1.2. The Japanese enforcement culture

Applying international value frameworks to a specific use case in Japan is challenging because of local values and context differences. For example, one of the challenges is the collision and difference in privacy and cultural enforcement regimes between Japan and other jurisdictions (Wang, 2020). The EU, for example, views data protection and privacy as fundamental rights, regarding such rights as a more dominant reason for regulation than economic incentives (McGeveran, 2016). This enforcement mechanism starkly contrasts with the Japanese government's approach, which views personal data as an economic commodity and protects a restricted range of personal information. Greenleaf and Shimpo (2014) differentiate between two types of cultural enforcement mechanisms: hard power and soft power. Hard power, used by the EU, is stated by Miyashita (2016), such as a solid regulatory body, law enforcement, and punishment. The challenge remains that the soft power mechanism could change per situation and

domain, as cultural values and social norms vary by socio-technical context. The Japanese regulator uses soft power, which coerces firms to protect society through cultural values and social norms (Miyashita, 2016). According to Miyashita, firms tend to adhere to ministry instructions due to their apprehension of violating social norms. According to the Japanese viewpoint, the potential consequences of breaching social norms, such as the loss of social trust and business reputation, are significantly more significant than the financial penalty (Wang, 2020). These differences make applying European frameworks directly to the Japanese context impossible.

1.1.3. The importance of reputation and social norms in Japan

As mentioned in Chapter 1.1.2, Japan operates its enforcement through a soft power mechanism based on reputation (Wang, 2020). In Japan, reputational value serves as a soft enforcement mechanism, where unforeseen occurrences in the IT environment possess the potential to either enhance or damage a company's reputation (Ishihara, 2006). The Japanese market can be differentiated into "domestic" and "overseas," with our focus on the overseas segment.

An overseas company entering the Japanese market must consider reputation management issues. Regarding overseas companies entering the Japanese market, the proverb "When in Rome, do as the Romans do" (Ishihara, 2006, p. 448) is applicable. The proverb refers to companies' ability to fully understand public values to operate without failing in the Japanese insurance market. A positive impact on corporate reputation is envisaged if the company applies social norms. The use case is focused on an overseas company, highlighting the importance of considering public values when using AI systems.

1.2. Scientific knowledge gap

The rapid advancement of AI across various sectors has ushered in an era of unprecedented technological capabilities and ethical complexities (Aizenberg & Van Den Hoven, 2020). Particularly in the life insurance industry, where decision-making bears significant consequences, the integration of AI presents challenges (Mullins et al., 2021). Furthermore, while the potential of AI is evident, numerous organizations have difficulty realizing the expected advantages. (Makarius et al., 2020). Moreover, AI systems may cause harm by violating local social values and norms, which can cause reputational damage to the organization. Therefore, there is an urge to understand the applicable social norms to control the AI lifecycle. Three gaps must be addressed to unlock this urge's potential fully.

1.2.1. A missing value framework for the Japanese life insurance industry

The first gap is the missing value framework regarding using AI in the life insurance industry. Currently, no unified set of values can direct the use of AI on both a global and national scale. These value frameworks are essential for ethically developing models and systems in decision-making processes with high stakes, such as the insurance industry (Kurshan et al., 2020; Truby et al., 2020). In response, numerous reports on how AI should be governed to protect fundamental rights have been published recently (Dobbe et al., 2021). For example, the EU's Assessment List for Trustworthy Artificial Intelligence specifies seven requirements for AI (AI HLEG, 2021). However, these guidelines are viewed from a Western perspective, and whether these guidelines work for the Japanese insurance industry is still being determined (van den Hoven et al., 2015; Wirtz et al., 2020). In addition, the Japanese government has published Social Principles for Human-Centered AI. These guidelines are seen from a high-level perspective but lack a translation to specific industry requirements. As Aizenberg and Van den Hoven (2020) state, local social norms must be translated into structured institutions to establish norms that align with the societal context. As seen from the current literature on design for values, there is little research on consolidating these high-level principles and requirements into the Japanese industry context.

This thesis aims to fill this practical and scientific gap by creating a value framework that gives insights into the applicable social norms for AI system developers. This value framework can add to the scientific knowledge base, enriching our understanding of global AI developments and diverse impacts.

1.2.2. The translation of the value framework into practical guidelines

The second gap is integrating AI to translate identified values into practical guidelines for AI system developers. Extensive literature discusses methods for converting social norms into structured guidelines that reflect ethical considerations (Shilton, 2012; Sapraz & Han, 2022; Van de Poel, 2018). Van de Poel

(2013) and Aizenberg & Van Den Hoven (2020) detail the processes for identifying social and ethical values and incorporating them into design requirements. Vries (2009) adopts a traditional system engineering approach to derive design requirements from stakeholder values. However, these methodologies often lack a translation to a practical method for AI system developers (van den Hoven et al., 2015).

Furthermore, the literature around the design of guidelines is a novel topic. Methnani et al. (2023) offer insights on converting high-level ethical guidelines into practical operational requirements, highlighting key characteristics in this process. However, as Mittelstadt (2019) notes, there is a need for further research and empirical data to implement AI systems effectively within their specific contexts. Wolters (2022) contributes to this area by providing a methodology for applying guidelines, particularly focusing on technical vulnerabilities from a socio-technical perspective. Nevertheless, there remains a significant gap in the literature, particularly in incorporating social norms into practical guidelines (Aizenberg & Van den Hoven, 2020).

The thesis aims to fill this practical gap by designing guidelines through insights from empirical data and knowledge from academic research. The insights serve the scientific knowledge base with guidelines based on a use case, which can be used as a best practice in subsequent research.

1.2.3. A reusable method to design guidelines for AI

The third scientific gap is a lack of a standardized method to identify, translate, and integrate social norms into practical guidelines from a safety perspective to control the AI lifecycle. The problem statement calls for guidance in safeguarding stakeholder values during AI integration by identifying relevant norms and embedding them into system design in an organizational context. The third gap can be bridged by applying closely related concepts from traditional literature to the problem synthesized, namely design for values, system safety, reporting standards, and AI governance.

The traditional theory of design for values primarily aids in pinpointing and translating values into system functionalities from a philosophical angle (Van de Poel, 2015). Nevertheless, it falls short in applying these values from a practical safety viewpoint, an aspect of which this thesis aims to protect stakeholders' values (Aizenberg & Van den Hoven, 2020). Leveson (2012) argues that systems safety theory adopts a systems theory perspective on causality to identify potential accidents within complex sociotechnical systems. The relationship between system safety and design for values is linked through the concepts of norms and identifying safety constraints. In addition, systems theory considers hierarchical structures, wherein each level imposes safety constraints on the activity of the level beneath it, thereby enabling or regulating behavior at lower levels based on the presence or absence of safety constraints at higher levels. This construction resembles Van de Poel's (2013) hierarchy of values. By employing system safety concepts, specifically safety constraints and safety control structures, this research adapts values into actionable norms (safety constraints), approaching the context of the whole system to protect stakeholders' values (Leveson, 2012).

Additionally, to ensure that AI systems remain operational and relevant for organizational use, the methodology narrows down to pertinent norms using steps outlined by Garst et al. (2022). Garst et al. (2022) introduce a six-step process, which presents an approach for selecting relevant subjects in organizations' environmental, social, and governance reports. The steps Garst et al. (2022) take are reusable and restructured to fit the research purpose.

The last step includes the integration of relevant norms into daily operations, aligning with organizational AI governance. Mäntymäki et al. (2022) reviewed several academic pieces of literature and defined AI governance from a governance perspective. The definition combines other literature where the technical, ethical, regulatory, organizational, and cultural components occur. The framework was selected because it closely relates to the use case and emphasizes straightforward guidelines for structuring and dividing the identified norms into data and AI governance.

The research aims to introduce an initial repeatable process by combining several concepts from the literature in the design for values, system safety, report standards, and AI governance frameworks. The topics are chosen because they are closely related to the problem statement: AI system developers in the Japanese life insurance industry face a challenge in determining which social norms should be protected and implemented to control the technical operations of AI systems in organizational settings. The initial method is tested and validated in a practical environment through a case study. This study employs various

concepts from existing literature to examine the practical applicability of this approach in an organizational context and validates its use through empirical insights. The primary utility of the method emerges in addressing unresolved problems characterized by undefined values, which need systematic control in high-stake domains, as exemplified by the life insurance industry in Japan.

1.3. Research objective and main question

This thesis delves into the problem of safeguarding values and norms within the AI lifecycle in Japan's life insurance industry. The research aims to develop design guidelines for developers of AI systems, enabling them to control the design, development, and deployment processes. These guidelines are systematically structured to protect stakeholder values through the various AI system design, development, and deployment stages. The design of these guidelines provides a method for future use cases, as exemplified by the PU system use case. This leads to the main question of this thesis:

What design guidelines can developers in Japan's life insurance domain follow to control AI systems while protecting stakeholders' values?

The results of these steps are guidelines that foster safe socio-technical AI systems' design, development, and deployment, driven by continuous improvement. Following these guidelines protects stakeholders' values, as they are integrated into the guideline design. Reaching this objective contributes to the practical and scientific fields.

The first contribution is filling the practical and scientific gap of a missing value framework for the Japanese life insurance industry. Creating the framework takes a social, legal, industry, and business perspective and translates these values into actionable norms focusing on safety. Also, creating a value framework draws on understanding which values must be protected and what is seen as an implication of these values for the Japanese life insurance industry. The value framework shows which aspects the Japanese life insurance industry can shed light on. These insights are valuable for the organization where the use case is practiced by understanding the impact of certain design choices on their stakeholders and the organization. The value framework highlights the interaction between stakeholders' expectations and AI systems' functionalities.

The second practical contribution is translating the value framework into practical guidelines. As Japan has a prevailing culture regarding social norms, this thesis adds value to the practical gap in guiding AI system developers in Japanese life insurance organizations through the complex socio-technical field of integrating AI systems. The guidelines guide the AI system design, development, and deployment process, focusing on value protection. The guidelines provide insights to problem owners regarding the necessary steps, controls, expertise, and resources required at specific times during the process. In addition, the guidelines are aligned with society, law, industry, and business objectives.

Last, the thesis makes a scientific contribution by providing a reusable method through synthesizing different methodologies from existing literature into a systematic process, from value identification to guideline formulation. Reusability is based on testing different concepts in a real-life practical environment and validating these findings with experts related to the problem and use case. To resolve the problem statement and reach the research objective, it integrates theoretical concepts from design for values, safety constraints, safety control structures, AI governance frameworks, and reporting standards to facilitate practical implementation.

1.4. Thesis structure

After introducing the thesis and the context of "Design guidelines to protect stakeholders' values in AI systems - Based on a use case in the Japanese Life Insurance Industry," Chapter 2 outlines the research approach and methods. Chapter 3 examines the environment of predictive underwriting, exploring AI's scope, goals, and relevant legislation. Chapter 4 establishes a knowledge base around societal, legal, industry, and business values. Chapter 5 integrates different types of theories, translating identified values into safety norms and establishing a methodological framework. This framework is then applied to collect empirical data and operationalize the norms with the support of corporate AI governance frameworks and reporting standards. Chapter 6 showcases the translation of these norms into concrete design requirements, illustrating their practical implementation. Chapter 7 uses all the insights from the previous sub-questions to create design guidelines. The thesis culminates in Chapter 8, thoroughly discussing the research findings

and the artifact, addressing limitations and recommendations, and providing a reflective overview of the entire research process.

Part II

Approach and method

2

Research approach and methods

This chapter outlines the research approach used to answer the main question: *"What design guidelines can developers in Japan's life insurance domain follow to control AI systems while protecting stakeholders' values?".* Furthermore, the research approach is delineated in sub-questions.

2.1. Research strategy

The study attempts to answer the main research question through a combination of design science research (DSR) and design for values (DfV) approaches. The DSR methodology provides a structured method that complements the DfV approach by providing a clear process for identifying, translating, and integrating values within their specific context. This structure ensures the inclusivity of different sources at different levels, which allows that subjectivity to be minimized by using triangulation.

2.2. Design science research

The DSR approach was chosen because the research is based on a real-life use case conducted by a life insurance company in Japan. In addition, this research faces a design problem, wherefore an artifact will be built. DSR provides a framework that combines literature and practice. According to Johannesson and Perjons (2014), DSR is the scientific study and creation of artifacts to solve real problems of public interest. DSR attempts to produce information that adds to existing scientific knowledge and uses actual data from local practice, as shown in Figure 2.1. The limitations of DSR lie in its applicability to other contexts, which may limit the generalizability and transferability of research findings (Jacob et al., 2021).

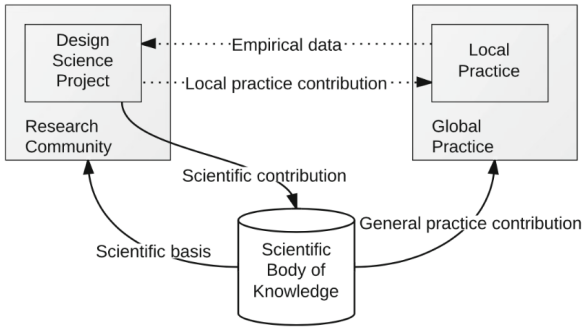


Figure 2.1: The design research science stated by Johannesson and Perjons (2014).

The artifact to be developed in this study is a methodological guide. This guide will serve as a systematic method for developers, offering them an approach during the design, development, and deployment of AI systems. It aims to safeguard stakeholder values by mitigating risks throughout the project lifecycle. The method will synthesize knowledge from diverse information sources, converting it into actionable guidelines and processes that address the identified problem (Johannesson & Perjons, 2014).

The three-phase perspective of design science research (DSR) can guide the systematic conduct of DSR, which is necessary for success (Hevner, 2007). The application domain, including stakeholders,

institutions, and technical systems, is described in the first cycle, called the relevance cycle, along with the challenges and opportunities of the domain. It connects the environment to the design. The second cycle, the rigor cycle, attempts to gather information from previous experience, knowledge, existing artifacts, and scientific ideas and methodologies relevant to the scope. The relevance cycle's needs and the rigor cycle's design and assessment theories and methods inform the third cycle, the design cycle, which creates and evaluates the artifact. Hevner (2007) argues that iterative changes in design and subsequent implementation rely heavily on real-world design testing in the context of the local practice environment. This research concludes by presenting design guidelines, validating the method with stakeholders during field testing, and drawing a full understanding of the knowledge base and local practice context.

2.3. Design for values

The DfV approach of Van den Hoven, Vermaas, and Van de Poel (2015) supports the sub-questions because the research focuses on translating local values to design requirements within a specific context. This approach focuses on developing design requirements that arise from values that can be transcribed into norms for the socio-technical context from a philosophical angle. The Japanese government also employs this strategy to identify the social human-centered AI principles, which serve as a document of Japanese AI values and guidelines (Cabinet Secretariat: Council for Social Principles of Human-centric AI, 2019). Moreover, this approach forms the basis for the chosen sub-questions in Table 2. The DSR three-cycle view provides a structure for the method explained in Chapter 4. Figure 2 provides an overview of the entire research approach. Limitations of this approach could include framing and prioritizing values and subjectivity through emotions (Friedman et al., 2021). Therefore, this research included stakeholders from different divisions to minimize this risk.

DfV is a design approach aimed at integrating values in all stages of the design process. The approach holds that moral values can be systematically specified and captured for practical purposes in a specification schema, a decomposition of ethical considerations, a values hierarchy, and a hierarchical structure of values, norms, and design requirements (Veluwenkamp & van den Hoven, 2023). This approach could be compared to a roadmap, as shown in Figure 2.2, where the top layer represents values. The bottom layer consists of specific design requirements, and in between, it contains the identified norms.

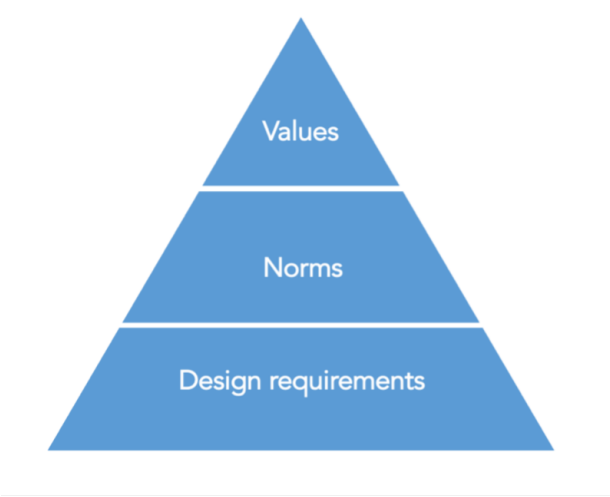


Figure 2.2: The three layers of value hierarchy (van de Poel, 2013).

Identifying relevant values and selecting those on which the design process should concentrate is the initial step in any systematic attempt to incorporate values into the design of new technologies (van de Poel, 2015; van de Poel, 2013; Michelfelder & Doorn, 2020). Second, the specification of values refers to the translation of values into norms and design requirements. Van de Poel (2013) identified norms as guidelines and may refer to the properties, attributes, or capabilities the designed object should have. Such norms may include what are sometimes called goals (for example, striving for "no discrimination in the output of our AI models") and constraints (for example, "no sensitive attributes can be used in the model"). These final norms are inputs to the design requirements. The design requirements make the

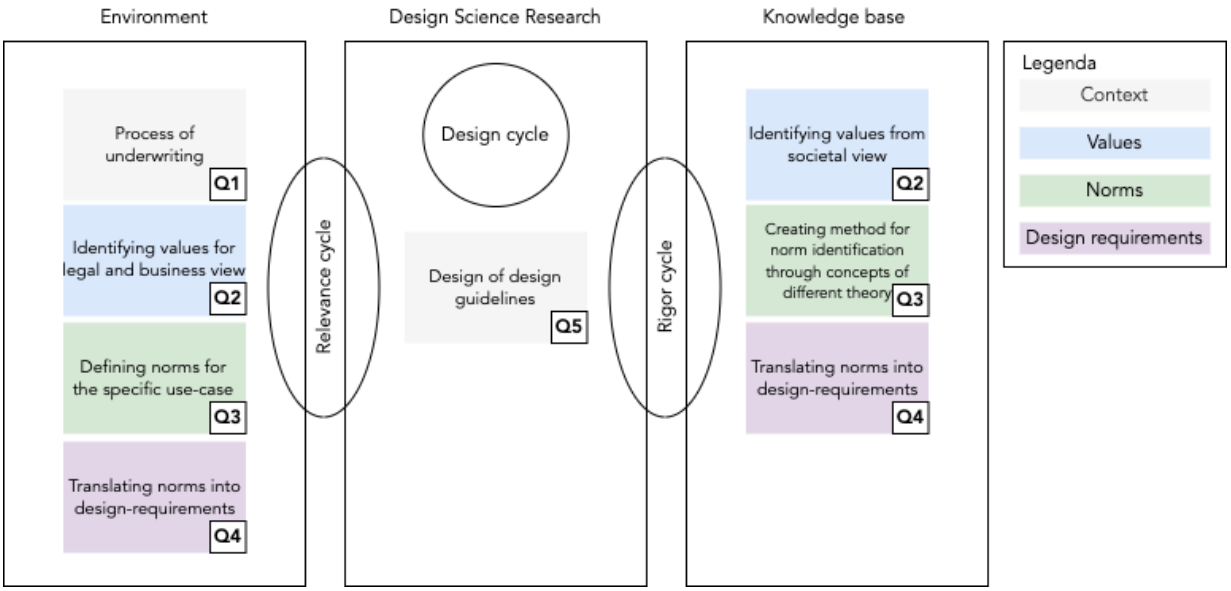


Figure 2.3: Design Science Research approach for this research adapted from Hevner (2007).

norms measurable.

When using the DfV, the possible limitations must also be considered. The values hierarchy of van de Poel (2013) is presented in a deductive manner, which means that lower-level norms or design requirements may need to be logically derived from higher-level identified values. This theory may turn out differently during the norms and requirements design process. The lower levels are more concrete or specific, and their formulation requires consideration of the context or design project for which the value hierarchy is being constructed (Veluwenkamp & van den Hoven, 2023; van de Poel, 2013).

2.4. Design approach

The problem statement and previously discussed research approach were used in formulating the sub-questions. The DSR framework is used as a basis to give structure to the sequence of questions. In addition, DSR provides DfV input for gaining knowledge about the environment before we start identifying values. Also, it gathers all the information gained from the DfV steps to draw guidelines in the design phase. Figure 2.3 shows this structure with the corresponding research questions. This figure shows that the first sub-questions focus on the use case to scope the design process. Sub-question two identifies the values applicable to the use case in the Japanese insurance industry. Sub-question three focuses on identifying and specifying the norms and risks. Sub-question four specifies the norms as measurable design requirements. The last sub-question is the design of design guidelines, where the results of the previous phases are combined.

2.5. Research phases and sub-questions

The research follows the three phases of the value hierarchy. Each stage answers one or more sub-questions that together answer the main question. Figure 2.4 shows the research phrases connected to each sub-question.

2.5.1. Phase 1: Exploring the environment

Research phase one emphasizes acquiring an in-depth understanding of the environment in which organizations utilize AI. Understanding the environment involves desk and field research to comprehend the drivers behind AI adoption. A high-risk use case involving customer data and directly impacting customers is selected for analysis.

Sub question 1 – What is the goal of the predictive underwriting system? This sub-question explores the motivations and restrictions behind adopting AI in underwriting processes. It presents a use case

Main question	*What design guidelines can developers in Japan's life insurance domain follow to control AI systems while protecting stakeholders' values?*	
Q Nr.	**Phase**	**Sub-question**
Q1	*Phase 1*	What is the goal of the predictive underwriting system?
Q2	*Phase 2*	Which values apply to the use case?
Q3	*Phase 3*	Which norms emerge from the use case?
Q4	*Phase 4*	Which design requirements can be derived from the norms?
Q5	*Phase 5*	What design guidelines can guide the development of AI systems?

Figure 2.4: Sub-questions per research phase.

focused on predictive underwriting to explain the socio-technical environment and identify the driving values businesses seek to achieve through this adoption. The inquiry delves into the manual activities, the impact of AI within underwriting, the stakeholders involved, and the regulatory environment. These insights define the research scope and serve as a foundational baseline, informing the development of the value framework and guiding the formulation of relevant guidelines.

2.5.2. Phase 2: Identifying and defining values

This phase is dedicated to identifying and defining the application domain's values, considering varied stakeholder perspectives encompassing societal, legal, and business viewpoints. Information is gathered from the knowledge and environment base. Completing this phase means establishing a value framework grounded in these diverse perspectives.

Sub question 2 – Which values apply to the use case? By answering this sub-question, AI system developers know what values to protect concerning the use case. This sub-question delves into the different perspectives of the stakeholders in the use case. The stakeholder groups are identified in the previous sub-question. The objective is to identify the relevant values specific to the use case. First, societal values are identified through the academic knowledge base. Consequently, a document analysis of already-existing frameworks and guidelines is conducted to explore legal and industry values. The business values are questioned through questionnaires. The identified values are the fundamental building blocks for the value framework and serve as input for the subsequent sub-questions.

2.5.3. Phase 3: Defining norms that derive from values

In the third phase of this research, the knowledge base and environment base are combined, specifically translating values into norms. The translation of values into norms involves the application of theoretical concepts that are examined through empirical research. This empirical study involves engaging stakeholders in practical settings to gather empirical data and identify any constraints that may impede the achievement of the identified values.

Sub question 3 – Which norms emerge from the use case? By answering this sub-question, AI system developers understand the scope and extent of the value being considered. Therefore, a few steps must be taken to translate the values into norms. First, this research delves into system safety theory, which employs the concepts of safety constraints and safety control structure to ensure safety in complex socio-technical systems. The problem statement's emphasis on protecting social values and norms inherent in AI-driven high-stakes decision-making is central to the study. This sub-question aims to develop a methodology that will be instrumental in collecting empirical data in the subsequent phase of the research. The concepts chosen are safety constraints and safety control structures from system safety theory. Thereby enhancing the understanding of system safety in the context of AI applications.

In addressing this sub-question, the study systematically collects empirical data during the environmental phase via questionnaires, workshops, and semi-structured interviews. The primary objective is to identify a broad range of risks associated with the AI lifecycle, drawing on insights from various experts. This compilation of data facilitates the formulation of norms through safety constraints. Post-data collection, these norms are refined and made practical through a methodology endorsed by selected literature in reporting standards and AI governance. The outcome of this phase is a consolidated set of norms, which will inform the design guidelines and serve as a foundation for transforming these norms into specific design requirements in the subsequent sub-question.

2.5.4. Phase 4: Specify design requirements from the defined norms

The concluding stage of the design for values approach involves defining the design requirements. This process entails a bottom-up assessment to ascertain whether the identified design requirements effectively embody the previously discovered values. The use case is a practical test for reflecting on and verifying the alignment between the design requirements and their context.

Sub question 4 – Which design requirements can be derived from the norms? In Sub-question 4, the research builds on the foundations laid by Sub-question 3, employing the established environment to operationalize norms. This phase involves evaluating the use case to render the set norms actionable. Establishing design requirements is enriched through expert interviews, ensuring these requirements support the application environment. The academic knowledge base supplements instances in which experts lack the understanding or ability to apply or resolve certain requirements. An exemplar of an identified value, accompanied by corresponding norms, is illustrated to demonstrate the process of translating norms into design requirements. In addition, conflicts and relations are mapped out to identify controls and make trade-offs. AI developers can design controls for the corresponding AI system by answering this sub-question.

2.5.5. Phase 5: Designing of design guidelines

The culmination of this research aligns with the Design Science Research (DSR) design cycle step. In this phase, all gather insights to inform the creative process to design a guide for AI developers to control AI systems and protect stakeholders' values.

Sub question 5 – What design guidelines can guide the development of AI systems? This final step synthesizes the insights accrued, laying the groundwork for the creative development of guidelines. These guidelines are tailored for AI systems within a specific contextual framework, underpinning the design process with a rigorous, research-informed foundation.

2.6. Method

The research methods are explained per research phase. This chapter clarifies the methodologies used for each sub-question, data collection, analysis of data, and results.

2.6.1. Phase 1: Desk- and field research

The primary focus is delineating the system boundaries and explaining the driving factors and restrictions in adopting and implementing AI in underwriting processes. The output of this approach is an understanding of the drive behind using AI in this use case. This focus is achieved by exploring the technological, process, and institutional lenses. This three-fold approach provides a holistic view of the predictive underwriting model, ensuring that all relevant aspects are considered. The technological perspective focuses on the AI systems' technical aspects; the process lens examines the workflows and processes; and the institutional perspective sheds light on the regulatory and organizational contexts within which these systems operate.

Understanding the motivation behind AI in the underwriting process, reviewing (gray) literature, conducting document analysis, and working with stakeholders are all ways to respond to sub-question one. Grey literature from the organization and legal entities is utilized to explore the use case manual, AI use, and regulatory environment. The collaborations with stakeholders are instrumental in identifying the impacts of AI systems on traditional manual underwriting processes and understanding the dynamic interactions between technology, processes, and institutional aspects. Academic literature complements areas with insufficient information. It provides theoretical underpinnings and comparative analysis, enhancing the robustness of the conclusions.

2.6.2. Phase 2: Integrative literature review, content- and comparison Analysis, and questionnaires

Phase 2 delves into societal, legal and industry, and business values. Bauer and Herder's (2009) framework provides a structure for how the various stakeholder groups should be interpreted.

Given the nascent nature of AI applications in Japan's life insurance industry, this study employs an integrative approach to identify societal values. As limited literature about societal values is available, a literature review offers possibilities to combine different topics to create an understanding of the topic

(Torraco, 2005). Appendix A shows the list of selected literature.

Next, legal documents are selected based on their relevance to the use case. The primary analytical tool employed is text analysis conducted via ATLAS.ti software. This process involves three key steps:

- **Value selection via word frequency:** We utilize word frequency tables to discern which values are predominantly mentioned in legal texts. Word frequency was not considered in the value selection. All values named are included.
- **Content analysis**: Subsequently, the identified values serve as a basis for coding across all documents, a step to draw an understanding of the nuances of these values as articulated by Weber (1990).
- **Comparative evaluation**: The documents, now coded, are compared based on these values. This comparison elucidates the significance of these values in risk identification and sheds light on the expectations of regulators and the industry from companies. The values consistently cited across Japanese documentation are then integrated into our value framework.

Furthermore, these identified values from legal texts inform the questionnaires distributed to organizational management. Due to participant time constraints, the questionnaire method was selected, utilizing the PROMETHEE approach to capturing human perceptions and identifying the most suitable alternatives based on problem understanding (Morfoulaki & Papathanasiou, 2021). The top three values that the participants prioritized and supported with arguments were arbitrary choices to include in the framework. A limitation of this method is that the value of the choice lies in the qualitative substantiation. Therefore, the questionnaire includes an option for participants to substantiate their choice.

The culmination of this phase is a set of values related to the use case. These insights are instrumental in addressing the subsequent sub-questions of the study.

2.6.3. Phase 3: Literature review, questionnaires, workshops, and semi-structured interviews

This sub-question thoroughly reviews system theory literature, focusing on concepts relevant to our problem statement. The literature review delves into the role of system safety and how it could aid DfV in the context of AI systems. The emphasis here is on understanding how safety considerations can inform the translation of values into norms, particularly from a risk-based and socio-technical perspective. This theory explores the interaction of safety, values, and norms within AI systems. The output is a method to provide structure in the workshop and semi-structured interviews. During the workshop, all participants utilized a standardized template. Distinct questions were formulated for each expert during the semi-structured interviews, considering the questionnaire outcomes and workshops. This method's limitation is the required time to prepare and analyze semi-structured interviews.

Next, empirical qualitative data is gathered from experts for risk identification and setting norms. Eight participants were asked in this stage, all with a relevant role to the use case. The method employs the same multi-actor decision-making approach using the PROMETHEE format for sub-question three, where participants prioritize values. These prioritized values are instrumental in mapping safety control structures and linking risks and values. After the questionnaires and the workshop, the process also includes conducting semi-structured interviews for deeper insights into workshop responses. These three methods guide experts from different domains through translating values into norms to cover the risks.

The process of reporting standards by which Garst et al. (2022) suggest a structured method to converge to a set of norms applicable to the use case. The direct outcome of this sub-question is a set of norms that must be assessed. Next, Mäntymäki et al. (2022) provide definitions for AI and data governance that support this research in categorizing norms within the framework. The result of this phase is a value framework based on a safety perspective gathered from empirical data and academic literature. This framework encapsulates values and norms, ensuring practical relevance and operational feasibility.

2.6.4. Phase 4: Semi-structured interviews

This phase involves collecting qualitative data via semi-structured interviews with experts in data science and underwriting. The focus of these interviews is to understand control mechanisms in manual processes and, with the aid of academic literature, explore how these can be adapted for AI systems. This process

supports the operationalization design requirements specific to AI in insurance by providing practical examples of setting these requirements, focusing on trade-offs, and developing control mechanisms.

2.6.5. Phase 5: Combining insights

The last phase serves as a synthesis of all the information gathered from the preceding sub-questions. It involves compiling and analyzing the outputs of norms and values identified in legal, societal, and business contexts and integrating these into the design guidelines. Therefore, backward traceability is used to implement the findings of the steps taken in this research. This thorough collection of insights shapes the process's form and structure, ensuring that the design guidelines are well-informed by thoroughly understanding the various factors and perspectives discovered in the earlier stages of the research.

2.6.6. Data gathering and processing

Data gathering for the knowledge base will be done through Scopus and Google Scholar. From the knowledge base, we gather information for the first, second, third, fourth, and fifth sub-questions. The document analysis is processed through a systematic coding scheme through ATLAS.ti. Professionals and researchers in various disciplines of study utilize the software ATLAS.ti for coding (Friese et al., 2018). This data can be processed by organizing it in a word table, structured with rows and columns of interest. The narrative data is subsequently displayed within the cells of the table.

Data gathering for the environment base will be done through questionnaires, workshops, and semi-structured interviews. All interviews are recorded and processed through Copilot. To process this data, Microsoft 365 Copilot is used. Microsoft 365 Copilot is an internal organizational system to keep participants' data safe during processing (Microsoft, n.d.). Copilot functions on a natural language processing system. By inserting the coding scheme, Copilot provides insights from the interview arguments by quoting from the participants. The answers are checked and validated manually.

2.6.7. Research flow diagram

The research flow is visualized in Figure 2.5. This research flow shows the five phases, with the data collection, input, methods, and tools needed to collect the data. In addition, the sub-questions divide the chapters.

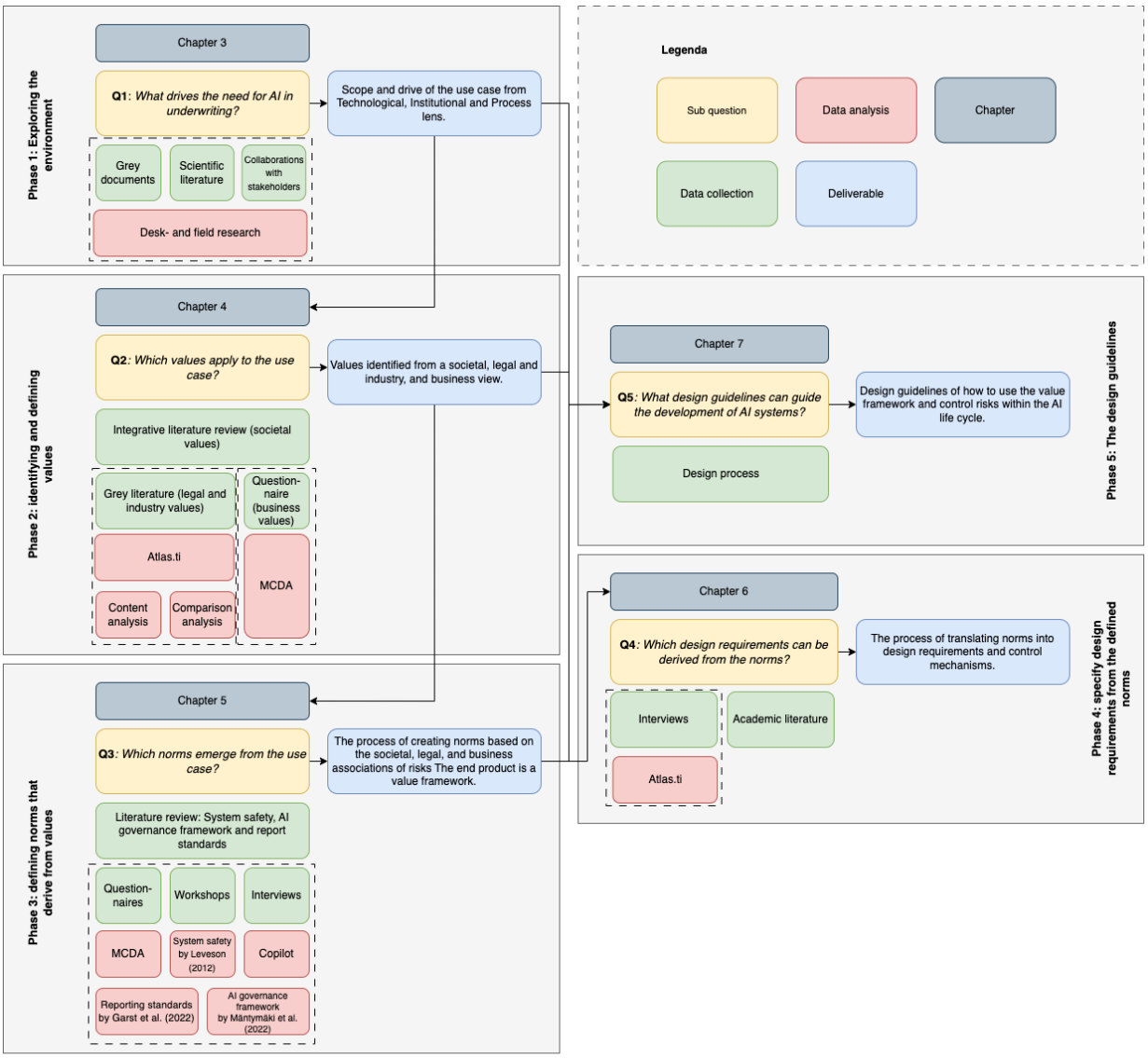


Figure 2.5: Research flow diagram.

Part III

Research

3

Predictive underwriting in insurance

This chapter explores the application environment in the first phase of this research. The research uses the PU system as a guiding example to identify the relevant values requiring protection. The use case helps inform the design of the value framework and ensures it remains closely aligned with real-world context. The chapter maps the PU landscape by investigating the manual purpose and use of underwriting practices, existing legislation around underwriting, and the current AI techniques used. This chapter concludes with Section 3.4, which addresses the first sub-question:

"Q1: What is the goal of the predictive underwriting system?"

By answering the sub-question, we gain insight into the specific organizational objective being pursued and system boundaries, contributing to the research by understanding the goals and restrictions of the system.

3.1. The manual underwriting processes

Currently, life insurance has become increasingly vital in the restitution of human activities. Insurance is the primary method for mitigating the potential for loss and uncertainty (Avraham, 2017). Financial support is granted in business and human existence to ensure protection against severe and devastating losses (Toshmurzaevich, 2020).

Financial security is ensured by calculating the risk for every client. Underwriting processes are part of the financial security process. They are established to cover the insurer's activities, assess the risks assumed for insurance, ascertain suitable rates and conditions, and assemble a profitable insurance portfolio. Highly qualified specialists (underwriters) are essential to carry out this activity. These specialists can solve complex problems associated with evaluating insurance objects and determining the probability of risks occurring. Toshmurzaevich (2020) asserts that the underwriting process in life insurance encompasses a series of measures to establish risk acceptance limits. The purpose of this process is also to guarantee adequate coverage of the insurer's risks. The process is designed to ensure the provision of insurance services based on a contractual agreement that satisfies the requirements of both the insurer and the insured. The contractual agreement encompasses the risks covered by insurance, the insurance rates, and the deductible amount.

Life insurance underwriting is a careful and complex process that involves assessing applicants based on predetermined criteria. According to Glenn (2003), this process can be simplified as:

"Each company has a set of criteria that applicants must meet in order to be placed into a certain plan. The underwriter compares the applicants to a matrix, and if they fit into it, they are accepted. It's all very scientific, anybody could do it if they had the specialized and arcane knowledge to understand the complicated categories that insurers use. If the application is denied, it is because the applicant's characteristics don't match up with the underwriting guidelines that were created by actuaries using very sophisticated data sets." (Glenn, 2003, p. 133)

The use case centers on leveraging the existing client database to up-sell or cross-sell products or services. Up-selling entails presenting a more extensive product variant the client is already interested in (Maier et al., 2020). Life insurance entails promoting and acquiring a policy with an increased coverage limit or supplementary advantages. Cross-selling refers to selling an additional product or service to existing

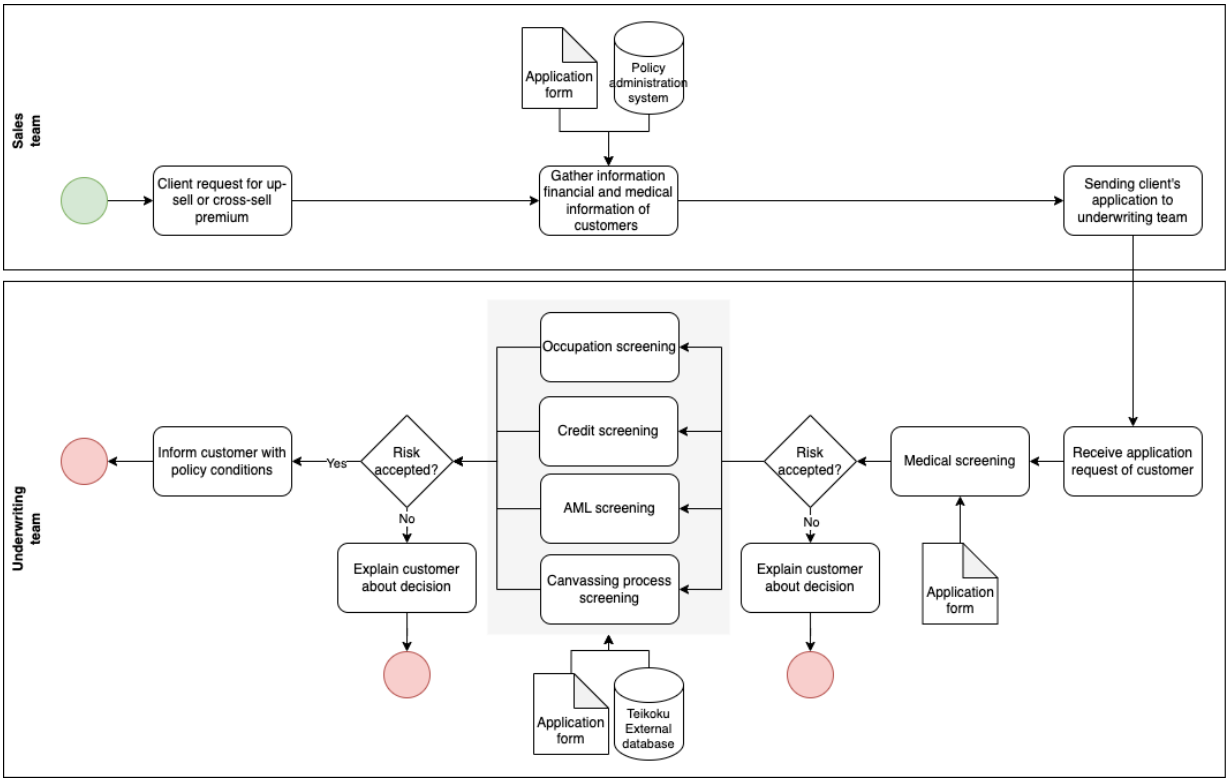


Figure 3.1: Manual risk acceptance process flow for sales and underwriting (created from document analysis and field research).

business clients. This selling strategy entails providing clients with a life insurance policy and the option to acquire supplementary products, such as health insurance or an annuity, within the life insurance industry. Underwriting practices are linked to executing up-selling and cross-selling strategies, serving various purposes (Cummins & Weiss, 2013). These strategies facilitate risk diversification. By cross-selling diverse products, such as life and health insurance, insurers can establish multiple revenue streams while distributing risk across various insurance products. Furthermore, larger product portfolios enable insurers to access more extensive data, enhancing their ability to conduct detailed risk assessments. The data provided allows insurance companies to create tailored offers that closely align with an individual's health, financial circumstances, and other pertinent underwriting factors.

Figure G.3 shows the process of underwriting use cases. The diagram illustrates that the process starts at the sales department, where the client requests to up-sell or cross-sell the premium product. This activity can be stimulated through two methods: either the agent receives advice from the insurer to up-sell or cross-sell a specific product, or the client requests an up-selling or cross-selling of their premium. In both instances, the applicant must complete an application form containing financial and medical details already known to the insurer (Aggour & Cheetham, 2005). However, any changes in these circumstances will be updated accordingly. The applicant's information is used to estimate mortality risk. Estimating mortality risk is an aspect of the underwriting process for various forms of life insurance. Actuaries are responsible for calculating the expenses related to insuring against the risk of death throughout the policy and converting it into a sequence of premium payments. Subsequently, the underwriters view these applications and rate the risk of insuring each person through medical and financial acceptance guidelines. A threshold determines the risks the insurer can and cannot cover. If these requirements cannot be met, the process concludes by providing the client with an explanation for the rejection of their application.

Once the application is deemed acceptable based on medical criteria, it proceeds to the subsequent screening stage. This stage encompasses occupation screening, credit screening, anti-money laundering (AML) screening, and canvassing process screening, all conducted simultaneously:

- Occupation screening assesses the potential risks and life expectancy associated with an individual's

activities.

- The credit screening process primarily assesses the client's financial capacity to afford a specific product based on its cost.
- The AML screening assesses the potential affiliation of the client with a terrorist organization or any involvement in malicious activities.
- The canvassing process examines the client's application for completeness and accuracy and determines if any additional information is required to decide. Additionally, senior underwriters provide additional guidance in cases of ambiguity or concern. Additional information is gathered through external Teikoku databases containing the client's financial information to check their eligibility.

Traditionally, underwriting for life insurance significantly relied on expert judgment, supported by medical and financial impairment manuals outlining guidelines for categorizing individuals into broad mortality risk classes. Complex cases are presented to the advising doctor. The manuals utilized point-based systems, where medical studies were used to assign values to different medical and behavioral attributes through debits and credits (Maier, 2019). The combination of various attributes, including specific medical conditions and family medical histories, formed an assessment that determined risk classes and corresponding premiums.

3.2. Formal and informal institutions in underwriting

The Financial Services Agency (FSA) is Japan's regulatory authority for insurance and reinsurance businesses. One of the FSA's regulations is the Insurance Business Act, which regulates the life insurance industry. The Insurance Business Act included regulations such as canvassing, AML, and terrorist financing. In addition, the government enforces the APPI to control citizens' data. These regulations are seen as formal institutions. Furthermore, the informal market structure impacts the Japanese life insurance industry. This chapter explains the formal and informal institutions that are important to consider when identifying system boundaries.

3.2.1. Insurance canvassing

The Business Act shed light on the canvassing process, also known as the solicitation process, in the Japanese government and industry documents. The FSA stated that the purpose of this Act is to protect policyholders by ensuring sound and appropriate business operations of those conducting insurance business and fairness in insurance canvassing, thereby contributing to the stability of citizens' lives and the sound growth of the national economy (Financial Service Agency, 2021).

The FSA obliged insurance companies to establish a controlled environment for insurance canvassing, wherefore guidelines are published in the 'Comprehensive Guidelines for Supervision for Insurance Companies' (2021) report. This report includes regulations that the canvassing process for older people and people with disabilities must be explained when rejected (Financial Service Agency, 2021, p. 245). In addition, this target group may not be discriminated against by their age or disabilities (Financial Service Agency, 2021, p. 304). In addition, an insurer may not approach clients without objective grounds, also called 'cold-calling' (Financial Services Agency, 2023).

Another entity in the industry is the General Insurance Association of Japan (GIAJ). According to the guidelines set by GIAJ, insurers must explain the risk underwriting to each client (General Life Insurance Association of Japan, 2020, p. 53). Despite being guidelines, these rules are expected to be adhered to by all organizations.

3.2.2. Anti-Money Laundry & Counter Terrorist Financing

The Business Act includes the Act on Prevention of Transfer of Criminal Proceeds, known as the Criminal Proceeds Act, influenced by the Financial Action Task Force (FATF) guidelines for Anti-Money Laundering (AML) and counter-terrorist financing (CTF) (Hiroshi, 2019). This Act forms a key part of Japan's AML framework. Financial institutions, including insurance companies, must follow a risk-based approach to assess and mitigate AML risks. This approach focuses on client Due Diligence, transaction monitoring, and other compliance measures. However, it is important to note that this law is not directly relevant to the PU system as these processes are parallel, and the outcomes are unrelated.

3.2.3. Act on the Protection of Personal Information

The Japanese government implemented the Protection of Personal Information Act, also known as the Amended APPI, on May 30, 2017, to safeguard clients when they provide sensitive personal information. The legislation defines ”special care-required personal information” as information about an individual’s race, creed, medical history, or other relevant factors. It establishes that obtaining such information without the individual’s consent is legally prohibited (Nishikino & Kanazawa, 2018). The laws prioritize protecting personal information and emphasize the potential utilization of anonymized personal data without individual consent, provided that appropriate safety management measures have been established.

The underwriting process utilizes ”special care-required personal information,” namely financial and medical data. According to Rothstein and Joly (2009), using data in the life insurance underwriting process is deemed acceptable when employed solely for this specific purpose. To justify using this data, life insurers should establish significant causal and medical evidence that links attributes to mortality risk (Maier et al., 2020). Laboratory tests and health questionnaires play a role in assessing mortality risk in life insurance underwriting, as they have a longstanding precedent and are grounded in medical principles.

3.2.4. The informal Japanese life insurance market structure

In addition to formal institutions, one must also consider traditional informal institutions. Sales practices in the life insurance industry often involve insurance agents, considered business partners of the insurance organizations. These agents work closely with clients to approach them and facilitate sales (Ishihara, 2006). Hence, agents have the highest frequency of interaction with clients. The agents collaborate with life insurance organizations to develop sales strategies while maintaining regular client contact.

In the given use case, agents collaborate with clients to finalize a product transaction successfully. The process’s initial stage involves the client applying for a paper application form. The agents can guide the clients in filling out these application forms. Nevertheless, information may be left behind by this flow during agent-client interaction. Incomplete or inaccurate information the agent collects can result in erroneous risk assessments during the underwriting practices of the life insurance organization.

Additionally, this can lead to delays in the approval process, inaccurately priced policies, or the issuance of policies that would have otherwise been rejected or modified. Hence, establishing effective communication channels and control mechanisms between agents and underwriters is paramount. These control mechanisms underscore the significance of accurate information exchange in guaranteeing data quality.

3.3. AI in underwriting practices

The purpose of AI in underwriting will be elucidated in the subsequent sections. This section will analyze the techniques employed in the use case and evaluate their impact on the manual process from technical and stakeholder perspectives.

3.3.1. The purpose of the AI system in underwriting processes

The main driver for using AI lies in the economic stability of the organization. Given the complexity and breadth of information required to make accurate risk assessments and the need for rapid decision-making, the process is often time-consuming and susceptible to errors. Manual processes hinder insurers’ ability to accurately calculate mortality risk and maximize product pricing efficiency (Maier et al., 2020; Fang et al., 2016). The digitization that characterizes our current state, in which businesses, governments, households, and individuals generate enormous amounts of data, is one external factor contributing to this limitation (Kumar et al., 2019). To handle large amounts of information effectively, it is crucial to incorporate and leverage AI technologies actively.

The predictive underwriting system (an AI system) streamlines lead generation in the underwriting process for the sales department to achieve accuracy and efficiency. The PU system facilitates the first step of the process by generating leads for clients to engage in up-selling or cross-selling activities, as shown in Figure 3.2. This process entails utilizing the client organization’s preexisting data to develop tailored offers. Once a client decides to request a product, the process starts.

The AI system facilitates the underwriting process by automating two activities: medical screening and financial screening. In doing so, AI’s predictive algorithms differentiate between client profiles when screening medical and financial risks during underwriting. The system effectively distinguishes between

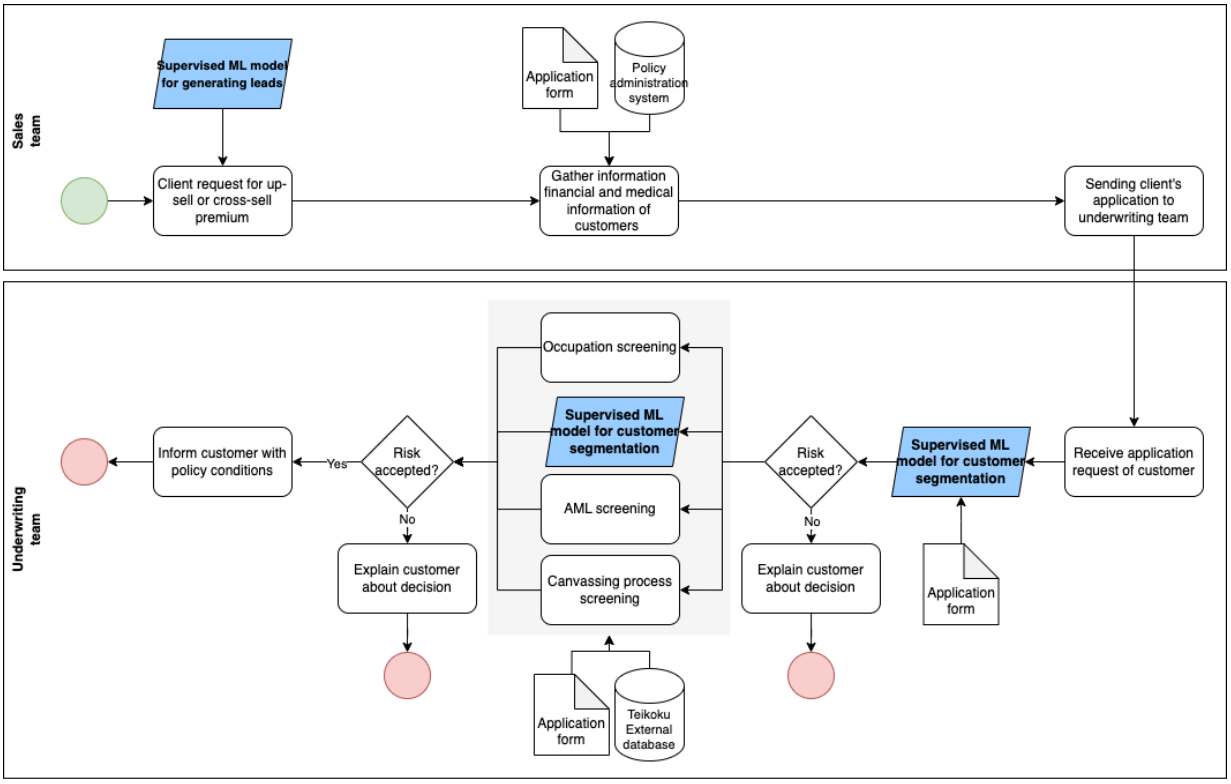


Figure 3.2: The risk acceptance process of sales and underwriting departments with the integrated AI models.

'standard clients,' who typically adhere to conventional risk thresholds, and 'high-risk clients,' who may present greater risks due to various factors. High-risk clients undergo manual evaluation as a step in the differentiation process to help find and manage complex risks, while standard-risk clients go through streamlined, automated processing. The AI system described here serves two purposes for the underwriting process:

1. Generating leads for sales activities to drive business growth: AI algorithms can tailor policies to meet each client's specific needs by analyzing a wide range of data sources and using big data (Maier, 2020).
2. Improving the efficiency of the underwriting process: Improving efficiency has a positive effect on business performance. Consequently, calculating risks could take weeks but can be completed in minutes using AI systems (Ceylan, 2022). Using AI systems for risk calculations creates significant cost savings for the insurer, contributing to underwriting and claim processing efficiency.

For both departments, the accuracy of the outcomes holds significant importance (Jaiswal, 2023; Kelley et al., 2018). In sales and underwriting, it is recommended that the client and the agent choose the correct product. An erroneous recommendation with a sufficient explanation can harm the trust between the insurer, agent, and client (Papenmeier et al., 2022). Furthermore, incorrect recommendations can be detrimental to the organization's financial health.

The primary motivation for employing a PU system is to enhance risk acceptance accuracy and work process efficiency. However, there is a trade-off between these two objectives. For instance, while a PU system might efficiently classify a few individuals as high-risk, this could compromise accuracy. The essential requirement within the use case is that the system's accuracy surpasses manual processes. Moreover, the system's utility depends on its ability to underwrite tasks more efficiently than the manual process. Consequently, the impact of the PU system lies in aiding client segmentation and lead generation in sales. The overarching goal is to achieve a more accurate sales and underwriting process, with an underlying need for efficiency in these operations.

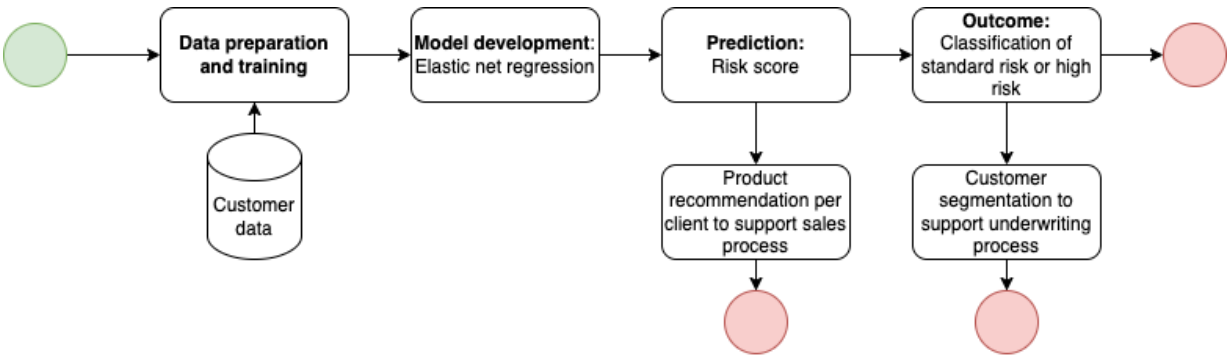


Figure 3.3: High-level workflow process of the PU system focused on the technical data flow, processing, and output.

3.3.2. The technical components of the AI system

Machine learning techniques are commonly employed in underwriting practices, including supervised, unsupervised, and natural language processing. The primary focus of the PU use case revolves around an AI model that exclusively employs supervised machine learning (ML) techniques. Supervised learning models can calculate predictions given a set of target variables based on another set of observations. Figure 3.3 illustrates the overall high-level workflow of the PU model.

The workflow commences by inputting financial and medical client data, which is subsequently prepared and trained. The method then uses elastic net regression and various decision tree methods to predict a risk score for every client. The score is a proactive tool recommending clients engage in up-selling or cross-selling premium products.

Once the score has been calculated, a threshold is established to differentiate between standard and high-risk applications. The financial and medical screening processes can be eliminated for existing standard-risk clients, as indicated by the model outcomes. The clients categorized as high-risk undergo a manual process with additional expertise for risk assessment.

3.3.3. The need of stakeholders from the PU system

Once the AI system's manual process, connected institutions, and purpose have been established, it becomes possible to describe the needs of stakeholders based on the system's purpose. Table 3.1 shows the needs of each stakeholder. The summarized information can be divided into several stakeholder views: stakeholder (agent and client), business, legal and industry, and societal. The business view is formally focused on accuracy, serving clients' needs, and a complaint system. The stakeholder's (agents and clients) view is focused on the need for trustworthy products and services and understanding the outcomes. Society's view could be redirected from legal entities that state guidelines to provide ethical and responsible guidance. The developer is excluded as they create the system. These insights highlight the need for a deeper understanding of societal, legal and industry, and business perspectives, as the current insights are overly general. Therefore, the following chapters will shed light on the values of these different stakeholder groups.

Stakeholders	Stakeholders' need from the system
End-user (underwriter)	Accuracy in the system outcomes, efficiency improvements to reduce manual workload, and understanding of the system input, output, and methodology (End-user, October 17, 2023).
End-user (Sales)	Accuracy in the system outcomes, faster policy issuance, and understanding of the system input, output, and methodology (End-user, October 17, 2023).
Agent	Understanding how the sales department came to a certain recommendation per product to explain the assigned product to their client and how the risk acceptance methodology works to inform their clients in the correct way of filling in the forms.
Client	Understanding the recommended product and the benefits of the cross-sell or up-sell.
Management	A system that fits the organization's vision. In addition, the organization should conform to corporate social responsibilities, adapt to customer needs, and ensure the business's financial stability.
LIAJ	From the given formal institution from the business act, it must be clear how the AI system is following the rules of giving sufficient explanation to elderly people and disabled people about decisions made on product recommendation level and risk acceptance level.
Government and FSA	The FSA and government have stated certain guidelines, rules, and regulations to ensure that client data is protected and may only be used with client consent. In addition, AML processes and insurance canvassing must not be disturbed by using the AI system. The guidelines, rules, and regulations state that decisions for the manual processes must be transparent and explained, with the group of elderly or disabled people mentioned explicitly.
Society	A trustworthy and ethically considered system.

Table 3.1: Stakeholders' need from the PU system.

3.4. Conclusion

This chapter analyzes the goals and restrictions for integrating predictive underwriting into insurance operations. The predictive underwriting use case is employed to establish system boundaries and address the subsequent sub-question:

"Q1: What is the goal of the predictive underwriting system?"

Answering the sub-questions resulted in understanding the primary objective of the PU system and what needs further exploration to provide AI system developers with guidelines to follow to protect stakeholders' values. This chapter highlights the need to incorporate laws and stakeholder input to establish goals and constraints for defining the scope of a specific system. The stated goals and constraints determine the choices in identifying and assembling a value framework and design guidelines. In addition, this chapter concludes that relevant stakeholder groups must be identified to know whose norms and values must be protected in developing and using the AI system. This identification can be achieved by understanding who is impacted by or has opinions about, the AI system within its established goals and constraints.

The primary objective of the PU system within the organizational context is to improve decision-making accuracy and efficiency in the sales and underwriting processes. Such improvements foster trust among insurers, customers, and agents. Moreover, the main driver for clients and agents is the risk acceptance process on a financial basis if managed accurately, transparently, and fairly. From an organizational perspective, the PU system supports the organization's pursuit of economic stability, which is achieved through precise mortality risk calculations and optimized workflow efficiency. This driver is on the condition that it considers the corporate responsibilities that must be met.

Corporate responsibilities in this context encompass adherence to existing laws, regulations, informal norms, and societal values pertinent to underwriting practices. These laws, regulations, informal norms, and societal values are constraints for the design of the AI system in its context. This study considers the following formal and informal underwriting standards established by legal and industry entities:

1. The FSA has guidelines for canvassing that directly relate to the PU system.
2. The APPI is significant due to its utilization of personal data.
3. The agent-client relationship within the informal market structure effects the information dissemination between the client and organizations.

However, the societal values regarding the PU system require further exploration as it is now described as a 'trustworthy and ethically considered system.' The absence of explicit laws and regulations regarding AI requires thoroughly examining legal and industry values, particularly concerning ethical and economic protections for society. Furthermore, the business values warrant exploration to ensure the system's impact aligns with organizational goals and financial stability.

This sub-question delineates the system boundaries based on the institutional environment. Subsequent sub-questions will delve deeper into societal, legal and industry, and business values, aiming to identify values for the safe functioning of the PU system, emphasizing accuracy, efficiency, and ethical compliance.

4

Stakeholder Values

This chapter analyzes the different stakeholder perspectives on the values that should be protected in applying AI within the PU use case. It builds upon the identified stakeholders from the previous chapter from the environment base. This chapter concluded that societal, legal, industry and business views must be considered when designing, developing, and deploying an AI system. Societal values emerge from an integrative literature review, understanding how AI interacts with Japanese society and what values must be considered to protect. Concurrently, legal and industry values are analyzed by examining different legal and insurance industries' frameworks on AI, data, and social rules, focusing on their vulnerable values. The results of the grey document synthesis serve as input for stakeholder questionnaires to gain an understanding of the business's values. Thus, this chapter addresses the following sub-question:

"Q2: Which values apply to the use case?"

Answering this sub-question provides insights into exploring each stakeholder's values when using AI in organizational settings. The results provide a foundation for the value framework and understanding the values that must be protected to integrate AI in this use case.

4.1. Values within a socio-technical system

The identification of values will have to be approached in a socio-technical context. Kroes et al. (2006) suggest that a system should include all elements and adhere to design considerations to fulfill its intended purpose. When controlling AI in engineering systems, one must recognize the incorporation of human agents, social institutions, and technical artifacts as vital components of these systems (Van de Hoven et al., 2015). Bauer and Herder (2009) illustrate these components through an intended framework based on Williamson's (2000) four-layer scheme, which outlines the interconnectedness of various social, technical, and institutional arrangements. The system exhibits top-down and bottom-up causation (Bauer & Herder, 2009). The upper levels empower and limit the lower levels, and vice versa. Figure 4.1 demonstrates the simultaneous description of technical and social subsystems.

In this research, Bauer and Herder's (2009) framework aids in understanding the focus on values from different perspectives. The primary objective of this research is to design guidelines. Therefore, a value framework comprising a set of institutions that stakeholders can utilize to protect their values in their AI design must be constructed. In socio-technical systems, as shown in Bauer and Herder's (2009) framework, there is direct and indirect interaction between social and technical subsystems. The PU system illustrates the interconnectedness of these subsystems. For instance, the PU system produces outcomes that rely on customer data analysis. This integration of AI facilitates the decision-making processes of employees regarding risk acceptance, which impacts customers' and society's financial stability. From the vantage point of the first layer of the framework depicted in Figure 4.1, the focus is on the societal values that must be protected. These values arise from the interaction between Japanese society and AI systems and how they could support or harm traditions, informal norms, and religion. The subsequent layer concentrates on the institutional setting and 'formal rules of the game' that must be protected. In the absence of 'hard rules' and because of the Japanese soft enforcement power (Miyashita, 2016), the emphasis relies on values frequently cited by industry to direct their policies. Finally, the business perspective is examined, emphasizing the values supporting the 'play of the game' and forming protocols and routines to protect the organization's financial stability.

Time scale	Social subsystem	Technical subsystem
Embeddedness Changes 10^2 to 10^3 years often non-calculative	Informal institutions, customs, traditions norms, religion	Informal conventions embedded in the technical artifacts
Institutional environment Changes 10 to 10^2 years, institutional setting	Formal rules of the game (property, polity, judiciary, ...)	Technical standards, design conventions technological paradigms
Governance Changes 1 to 10 years design of efficient government regime	Play of the game (cotracts, governance of transactions)	Protocols and routines governing operational decisions and (best available) technology
Operation and Management Continuous adjustments	Prices, quantities incentives	Operational choices

Figure 4.1: The four layers dependent on time scale and interaction with social and technical subsystems inspired by Williamson (Bauer & Herder, 2009).

By thoroughly understanding these prevailing norms and values from different stakeholder perspectives, a foundation is formulated for developing the value framework. In addition, it provides insight into what values should be protected and why.

4.2. Societal values

This section employs an integrative literature review to identify social values. The analysis initially focused on the prevailing norms regarding accepting AI in Japan. Next, we analyze the literature to identify values that may harm society while utilizing AI in organizational settings.

4.2.1. General societal norms and values

Ethics can play a role in establishing AI systems within organizations to gain acceptance from stakeholders and society. De Pagter (2023) argues that this perspective on ethics primarily regards ethical approaches to critically analyze the impact of AI systems on our wider community and culture. This perspective emphasizes the importance of ethical approaches for understanding and addressing the complex ethical implications of AI technologies. Therefore, to better understand the interaction between Japanese society and AI, we use four philosophical notions stated by McStay (2021) that are drawn from:

1. The international and national influences on modern Japanese AI ethics policy;
2. The hybridization of dominant Japanese ethical beliefs with Western discourses during and after the Meiji Restoration (1868) that initiated Japan's modernity;
3. Historical events and the mixed religious landscape.

These ethical takeaways, deeply rooted in Japanese philosophies and cultural perspectives, provide unique insights for addressing ethical challenges in adopting and using AI. McStay (2021) states that the lessons are focused on emotional AI but are widely applicable. The four moral lessons are: (1) contributing community, (2) wholeness, (3) sincerity, and (4) sensitivity.

First, the contributing community perspective emerges from a community's history of harmony and interconnectedness. The Japanese government emphasizes in the documentation about Society 5.0 the perspective where humans and machines live harmoniously (Berberich et al., 2020). This community-based perspective contributes to ethical AI design and governance, especially concerning transparency, fairness, and collective rights. Second, wholeness comes from Japanese thinking that recognizing wholes and contexts is more important than objectification and reductionism. This perspective is vital for AI ethics, highlighting the need to consider the broader context from which specific AI decisions and judgments arise (Morita, 2012). This approach aligns with the quest for transparency in AI decision-making and promotes a deeper understanding of the implications of AI on human lives. In the context of the PU system, this perspective encourages the industry to view system outcomes within their broader context. Third, sincerity

stems from religious principles emphasizing sincere and respectful conduct in human interactions. This ethical lesson underscores the significance of ethical behavior in AI development and deployment. It encourages transparency of intention, integrity, and privacy considerations in AI systems (Murata, 2019). Fourth, sensitivity stems from Japan's sensitivity to the interconnectedness of all things, as seen in religious beliefs, and extends to the insurance industry's adoption of AI. This perspective encourages insurers to approach AI systems at an operational level, acknowledging and learning about the profound impact they may have on business processes.

The philosophical notions must be considered when further identifying societal values. Using these concepts reinforces the understanding of pre-existing institutions and engenders novel insights. Participants in the research did validate these notions.

However, there are limitations to the utilization of these ethical lessons. The first limitation arises due to the choice of perspective. The four designated concepts examine a significant historical event, specifically the Meiji Restoration and Modernization in 1868. The change in human values does not include substantial events like World War II (1937) and the Japanese asset price bubble (1990). One limitation of the study is its predominant focus on emotional AI, potentially neglecting the exploration of more philosophical concepts. Nevertheless, we employ these concepts due to the constraints imposed by time and the extent of this study.

4.2.2. Societal values in organizational context

AI's increasing sophistication and integration within organizational contexts demands understanding its potential harm to society before being implemented in diverse digitalized systems. This potential harm can be identified within the AI life cycle. What should be known for the identification of the values is that the AI life cycle consists of three phases: design, development, and deployment, supported by academic literature and practical experience (De Silva & Alahakoon, 2022; Haakman et al., 2021). The design phase encompasses idea generation and exploration of potential impacts stemming from a problem. The development stage signifies the implementation of the previous stage. During the deployment phase, the primary objective is to standardize and improve the accessibility of the service and solution for all stakeholders and end-users.

According to De Silva & Alahakoon (2022), trust, explainability, robustness, usability, privacy, cybersecurity, and fairness are the parts of the AI life cycle that are the most vulnerable values from a societal perspective. Based specifically on addressing high-level risk factors across the whole life cycle, from conception to production, these findings are based on recent work on the evaluation and defense against vulnerabilities of AI applications (Falco et al., 2021; Berghoff et al., 2020; Taeihagh, 2021; Stahl & Leach, 2022). These important considerations are incomplete, as other system-level risk factors might be unique to the application domain and local practices. Variable risk factors include, for instance, those related to local insurance industry regulations. In addition, these values are purely focused on the vulnerabilities of societal perception.

Trust, explainability, and robustness stem from AI ethics frameworks. Related to this, the Japanese Social Principles of Human-Centric AI emphasize trust and transparency (Cabinet Secretariat: Council for Social Principles of Human-centric AI, 2019). The assessment of usability risk in organizational systems and processes focuses on the skills and receptiveness of employees toward introducing AI models. The assessment of privacy risks encompasses evaluating the impact of AI adoption on the personal information of accumulated systems and data (Solove, 2023). Cybersecurity risks hold equal importance as they can potentially impact the foundational and source systems utilized by AI (De Silva & Alahakoon, 2022). The fairness of AI models must be evaluated concerning the end-users or consumers and the decisions that AI influences.

The mentioned values can be evaluated for potential risks throughout the AI life cycle by reviewing and reassessing each risk factor within the context of each phase (design, development, and deployment). Developers and operators can enhance public confidence in high-risk decision-making systems by identifying and enumerating potential public- and system-safety risks and determining acceptable mitigating methods.

Depending on the local context, the assessments should consider a wide range of individuals the system might affect or use. The context dependency is particularly important, as developers and operators may overlook potential risks or impacts. The high level of safety in commercial aviation is a significant factor,

as Falco et al. (2021) demonstrate. The level of safety in commercial aviation is maintained through extensive collaboration among manufacturers, operators, employees, regulators, and researchers to identify and address potential safety issues. Additionally, these collaborations are also supportive of the problem of transparency. The criticality of transparency lies in disclosing known risks and implementing risk mitigation measures, even if the algorithms within the highly automated system lack explainability or transparency. Furthermore, the assessment functions as a mechanism for monitoring and managing potential assumptions that may become outdated or inaccurate over time for the model (referred to as model drift) when the highly automated system utilizes adaptive or learning algorithms (Falco et al., 2021; De Silva & Alahakoon, 2022; Birkstedt et al., 2023; Metcalf, 2021).

4.3. Legal entities and industry values

The assessment's values were determined by analyzing six documents from different entities. The documents are chosen based on their relevance to the specific use case. The use case highlighted in Section 3.2 the importance of considering legal documents, such as the FSA governance regulations, APPI, and LIAJ guidelines. The Japanese government's AI principles and the Life Insurance Association of Japan code of conduct documents are being considered to ensure their relevance to the respective domains.

The absence of regulations governing the use of AI asks for the adoption of best practices. Europe was selected as the focus of this study due to its ongoing development of rules and regulations, specifically the AI Act. Additionally, the insurance industry has been provided with a framework by the European Insurance and Occupational Pensions Authority (EIOPA). Furthermore, Japan and Europe engage in ongoing discussions regarding AI advancements and their implementation strategies. When considering Europe as a best practice, examining whether certain values align with Japan's formal and informal institutions is essential. This section will explain each document and its respective purpose.

4.3.1. Selected documents

The Japanese government wrote the first document outlining the Social Principles of Human-centred AI published in 2019 to provide principles for incorporating AI into society (Cabinet Secretariat: Council for Social Principles of Human-Centric AI, 2019). The three core tenets of the principles are sustainability, diversity and inclusion, and human dignity (Habuka, 2023). The Social Principles aim to use AI to accomplish these goals for every industry rather than restricting its application to protect them. The declaration aligns with the AI framework of the Organization for Economic Cooperation and Development (OECD), particularly its first principle, which uses AI to advance inclusive growth, sustainable development, and overall well-being.

The FSA, a governmental organization, released the second analysis paper (Financial Services Agency, 2021). The FSA sets regulations to guarantee that insurance businesses adhere to regulatory requirements and manage underwriting risks efficiently. The document contains a section on effective governance maintenance. The FSA monitors insurance providers to ensure they follow the law. The upkeep of technology in insurance firms is particularly covered in the governance chapter.

The Japanese government publishes the APPI. Although this law does not specifically govern AI systems, it still applies to the use of data within AI systems. Organizations that gather, use, or transfer personal data must comply with the requirements outlined in the APPI.

The Life Insurance Association of Japan (LIAJ) published the fourth document, a code of conduct, in 2018. The LIAJ is dedicated to upholding the highest reliability standards within the life insurance industry. Hence, the emphasis of this document is directed at the social corporate responsibilities of life insurance organizations, based on the fact that life insurance organizations should satisfy customer demands while also carrying out societal expectations and responsibilities in their commercial operations. The document itself was made available in both Japanese and English. The English version was outdated and lacked certain details; therefore, we translated the document as accurately as possible while keeping the Japanese version by our side. At risk is the omission of the significance of fundamental social norms. The translated copy with annotations of what was attached is included in Appendix B.

The European Union has initiated the first steps for a framework for implementing AI in the insurance sector. The document analysis incorporated this perspective as a best practice for assessing differences,

relationships, and gaps within the AI domain for insurers. The European Union is considered a best practice due to its significant engagement in AI and is a leader in creating rules and legislation (Bradford, 2020).

The European Insurance and Occupational Pensions Authority (EIOPA), a financial regulatory body, published a paper on AI governance and digital ethics in the EU insurance market (EIOPA Publishes Report on Artificial Intelligence Governance Principles, 2021). The EIOPA states that its framework's main goal is to conduct fair business. The framework is based on the coming AI European Act.

The last document is the proposal of the European AI Act ("Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS," 2021). The European Commission has proposed an AI Act to govern AI systems within the EU. The act includes standards for high-risk AI systems and focuses on safety, ethical use, and trust in AI. The rule is based on a risk-based approach and seeks to enhance the EU's involvement in establishing international standards for reliable AI. It has guidelines for reporting, assessing, tracking, and reviewing the AI framework. There will be a designation of supervisory and stakeholder authorities and the establishment of cooperative mechanisms.

4.3.2. Coding and analysis

The selected documents range from 5-500 pages and cover many AI-related topics. Content analysis was utilized to identify the existence of specific words, topics, or concepts within a given set of qualitative data to examine the content of these pages. By first scanning through the reports and looking at the word frequencies of each document, the first 92 values were identified. Because these values vary in level of detail, overarching themes were developed. An example is the value *responsibility*. Here, the different forms of responsibility were *corporate social responsibility, environmental responsibility, and financial responsibility*. Also, in the context of the use case, some values were interrelated, such as the values *freedom*, which has a similar definition, *human autonomy*. After merging similar themes, 27 values remain. Appendix C gives an overview of the coding used, and the added quotation book shows how the codes are assigned within the documents.

Once a set of values has been established, the subsequent pursuit involves examining connections, contradictions, and gaps. Therefore, a comparative analysis provides an overview of the specific words used in each document. Table 4.1 displays the values on the Y-axis and the documents examined on the X-axis. The blue boxes indicate when a value or synonym is missing from the document. For instance, there is no mention of *"accessibility"* in the LIAJ documentation.

The analysis will yield valuable insights regarding the questionnaires administered to the management members of the organization in which the use case is operational. Before starting the analysis, there are criteria set up to run this analysis:

- All the values mentioned by the FSA will be used, as they serve as the primary guidelines for the life insurance industry unless the value is inconsistent within the given use case.
- The values mentioned in all the documents will be included in the list of potential values but will be analyzed in definitions to understand their usage.
- This analysis will examine the European values absent from Japanese documentation and assess their compatibility with the Japanese life insurance industry and the expectations of legal entities.
- All values mentioned in Japanese documents will be included in the list of potential values for the questionnaire to the management.

In this regard, all the values mentioned by the FSA are retained, as they serve as the primary guidelines for the life insurance industry. Furthermore, this analysis will compare the European values mentioned in the documentation with those absent in the Japanese context. This assessment aims to identify any potential missing values that ought to be incorporated in the subsequent phase. Additionally, we will examine the values utilized in the financial documentation that are absent from the government documents.

4.3.3. Similarities in documents

Looking at the results of Table 4.1, several values appear in all documents: *accountability, accuracy, diversity, effectiveness, fairness, human autonomy, privacy, responsibility, safety, security, transparency,*

	Social principles	LIAJ	APPI	FSA	EU AI Act	EU EIOPA
Accessibility						
Accountability						
Accuracy						
Adaptability						
Diversity						
Effectiveness						
Efficiency						
Empowerment						
Continuous improvement						
Fairness						
Human autonomy						
Human oversight						
Inclusiveness						
Integrity						
Learning						
Privacy						
Profitability						
Proportionality						
Redundancy						
Reliability						
Responsibility						
Safety						
Security						
Stability						
Transparency						
Trustworthy						
Understandability						

Table 4.1: Missing values analysis of legal and industry documents. The blue-colored boxes indicate which values are not mentioned in the analyzed document.

and trustworthiness.

Many common elements can be found when comparing the Japanese and European documents. The belief in applying AI in a human-centric manner and the commitment to upholding the fundamental rights of individuals and democracy are among the key principles. The operational values such as transparency, safety, security, accountability, and effectiveness hardly seem to differ within the documents (Kozuka, 2019). The Ministry of Economy, Trade, and Industry (METI) published the 2021 AI Governance in Japan Ver. 1.1 the report describes how to implement Japan's AI Social Principles and is similar to the definitions used within the EIOPA documentation. An example of transparency mentioned in the EIOPA documentation is *"Insurance firms should strive to use explainable AI models, particularly in high-impact AI use cases..."* (EIOPA Publishes Report on Artificial Intelligence Governance Principles, 2021, p. 40). METI mentions *"...AI systems that are unacceptable, high-risk, obligated to be transparent, or of minimal or no risk"*(The Ministry of Economy, Trade, and Industry, 2021, p. 12).

The commonalities between the Japanese Social Principles and the European AI Act can be identified at the governmental level based on their shared principles and corresponding values. The social principles prioritize seven pillars: human-centricity, education, privacy, security, fair competition, fairness, accountability, transparency, and innovation. The latter lists five principles: beneficence, non-maleficence, autonomy, justice, and explicability. However, disparities exist between the definitions used within the documents. Namely, the Japanese documents prioritize policy orientation, whereas the European documents adhere to a rights-based approach.

The value of *fairness* is in all documents mentioned. There is one main difference between the Japanese and European documents. All Japanese documents show some guidance in the economic definition of

fairness, namely fair competition, whereas the European documents lack any focus on the economy. An example cited from the Social Principles document is *"A fair, competitive environment must be maintained to create new businesses and services, to maintain sustainable economic growth, and to present solutions to social challenges"* (Cabinet Secretariat: Council for Social Principles of Human-centric AI, 2019, pp. 9–10). In addition, the LIAJ mentions the value of fair competition similarly as *"In order to establish the firm trust of customers and society, life insurers shall conduct fair and impartial business activities and confirm with the norms of society, including all relevant laws and regulations"* (Code of Conduct, 2018, p. 3).

Several *fairness* principles can be connected to the underwriting case, such as (1) data collection, (2) protection of consumer rights, and (3) wealth and social distribution. First, Data collection should refrain from engaging in unfair practices associated with privacy infringements or exclusive access to data. Equal access to relevant data should be ensured for all insurance companies. Second, underwriting activities must avoid discriminatory or biased practices targeting specific groups to protect consumer rights. Maintaining fairness in evaluating risk factors and determining premiums ensures the accessibility and affordability of insurance products for all eligible customers. The final point discussed pertains to wealth and social distribution, encompassing affordable insurance coverage to a broad spectrum of policyholders. The insurer must refrain from engaging in practices that unjustly concentrate wealth or social influence in the hands of a select few stakeholders.

Another difference in definition is *privacy*. The provided documents outline a perspective that defines privacy in the Japanese context as a concept that has a focus on the economic value of data and offers narrower protection for personal information compared to the EU (Wang, 2020). Article 1 of the APPI emphasizes its objective to safeguard the rights and interests of individuals. The article explicitly states that the secondary purpose of data is its economic salience. The report asserts that utilizing personal information appropriately and effectively fosters the development of new industries and the achievement of a dynamic economic society and an enhanced quality of life.

Additionally, the APPI solely safeguards personal information, resulting in a more limited range of protection than the broader right to privacy and data protection offered by the GDPR. The scope of protected personal information is limited to data that pertains to a living individual, such as their name, date of birth, or other identifying descriptions. This scope includes information identifying a specific individual or data containing an individual identification code. The GDPR defines personal data as any information that pertains to a natural person who can be identified or is potentially identifiable: this perspective difference and regulations concerning the underwriting use case. Appendix D gives an overview of privacy laws that apply to the underwriting use case, which will later support the definition of the norms.

4.3.4. Differences in Japanese and European documents

The European documents include *human oversight, proportionality, and redundancy*, which are not found in the Japanese records. The European documentation primarily addresses AI's technical vulnerabilities affecting social values. For instance, the EU AI Act mentions, *"The robustness of high-risk AI systems may be achieved through technical redundancy solutions, which may include backup or fail-safe plans."* ("Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT," 2021, p. 52), which is redundant for creating multiple layers of protection or backup mechanisms to prevent failures, errors, or vulnerabilities from compromising the functionality and security of the AI system. Therefore, we leave it out of the analysis in the next phase.

proportionality is focused on the level of risk the AI system carries. Chapter two of the EU-initiated regulation is described as *"The proposal builds on existing legal frameworks and is proportionate to achieve its objectives since it follows a risk-based approach and imposes regulatory burdens only when an AI system is likely to pose high risks to fundamental rights and safety"*("Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS," 2021, p. 7). AI systems are classified as high-risk if they employ biometric identification and punitive measures and jeopardize the safety of individuals, as well as if utilized in critical infrastructure sectors like healthcare or energy. The use case of underwriting is categorized as high-risk under European legislation as it establishes a risk indication regarding customers for premium calculation. EIOPA emphasizes the importance of *transparency* and *explainability* in communicating the principle

of proportionality, which entails providing clear information about the choices made during the model's development and the resulting outcomes. McStay's (2021) previously stated philosophical view on the interaction between AI and society emphasizes the core value of *transparency* to mitigate potential ethical implications. Japan currently lacks a specific list of high-risk AI system applications that can be utilized. The consideration of *proportionality* in the Japanese life insurance context is highly questionable.

European documents mention *human oversight* as a requirement. Article 14(2) of the AI Law stipulates that "*Human oversight shall aim at preventing or minimizing the risks to health, safety or fundamental rights that may emerge when a high-risk AI system is used by its intended purpose or under conditions of reasonably foreseeable misuse, in particular when such risks persist notwithstanding the application of other requirements set out in this Chapter.* ("Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS," 2021, p. 51). Although the selected Japanese documents do not directly discuss this topic, there is a notable interest in it. Japan proposed a set of fundamental rules for developing AI at the G7 meeting on technology in April 2016 (Lundin & Eriksson, 2016, p. 8). During this meeting, Japan proposed a rule where humans must control AI systems. The Japanese documents do not employ a specific definition for human oversight; rather, they express it as "proper control," indicating the condition to mitigate or eliminate the risk associated with AI systems before their utilization.

The Japanese documentation repeatedly references *continuous improvement*, which is not mentioned in the European documentation. Bhuiyan and Baghel (2005) define *"continuous improvement"* as the organizational culture of ongoing improvement to eliminate waste in all systems and processes. Individuals work together collaboratively to make improvements without requiring substantial capital investments. Japan developed this working method by incorporating quality control as a management tool for ongoing modification based on their ideas. The term kaizen has expanded beyond its original application in manufacturing to encompass a wider scope, involving all members of an organization (Imai, 2007). The LIAJ emphasizes the importance of continuous improvement in risk management measures for life insurers. They assert that *"Life insurers shall strengthen risk management measures under the leadership of executives with proper operation and continuous improvement to meet obligations to customers and establish trustworthiness"* (Code of Conduct, 2018, p. 5). This is crucial to fulfilling customer obligations and establishing a reputation for trustworthiness. In addition, the FSA mentions 'continuous improvement' also for minimizing risks in general, such as *"Whether the insurance company is making continuous efforts to identify the risks inherent in the current system..."* (Financial Services Agency, 2021, p. 134) but also within cybersecurity written as "*In addition, whether it is making continuous efforts to improve its information security control environment through the PDCA cycle, taking notice of illegal incidents or cases of problematic conduct at other companies*" (Financial Services Agency, 2021, p. 134). We can now revisit the principle of proportionality, as mentioned in European documentation, as continuous improvement focuses on refining risk management by identifying risks in small steps.

4.3.5. Differences in financial and legal documents

The FSA, LIAJ, and EIOPA documentation repeatedly references *profitability*. Profitability is consistently linked to asset management. An example from the LIAJ is, "*Life insurers shall engage in asset management seeking to ensure safety, profitability, and liquidity, taking into consideration its social and public nature*" (Code of Conduct, 2018, p. 4). The EIOPA places it under the definition of corporate social responsibility as *"As suggested by the AI HLEG, insurance firms should, as part of their commitment to Corporate Social Responsibility (CSR) and bearing in mind the nature of free competition in the markets, find a balance between the various and changing interests of different stakeholders when considering the ethical challenges of AI and digitalization. … there are the interests of the insurance firm and its shareholders in sustaining a profitable business in competitive markets"* (EIOPA Publishes Report on Artificial Intelligence Governance Principles, 2021, p. 24). Profitability in this context pertains to insurance companies' economic health and regulatory supervision, specifically emphasizing the PU process. These statements do not directly pertain to the PU use case and will be eliminated for the next phase.

4.3.6. Conclusion of the coding and analysis

After conducting the comparison analysis and gripping about the definitions of the values mentioned in the documentation, it was decided to eliminate four definitions from the list:

- **Empowerment:** The value is exclusively referenced in the AI Social Principles established by the Japanese government, wherein the provided definition does not align with the specific use case.
- **Proportionality:** Proportionality is exclusively referred to in European documentation and does not align with the Japanese control approach.
- **Redundancy:** The mentioned control mechanism is exclusively referenced in European documentation and is deemed insufficiently advanced to effectively operate within the AI value framework, as this value is too low-level focused on data.
- **Profitability:** The documents lack focus on the use case within the given context. Although intriguing for the business, profitability is omitted as it does not align directly with the purpose of the AI value framework.

One notable distinction between the European and Japanese approaches lies in Japan's explicit emphasis on a risk approach rooted in the value of continuous improvement. Therefore, this value is included in the value framework.

4.4. Business values

After analyzing the government and industry documents, the practical application of values within the Japanese life insurance industry must be juxtaposed with the day-to-day operational realities of businesses. This section delves into the organizational values experts deem essential to protect when implementing and using the PU use case. By employing a multi-criteria decision-making analysis (MCDA), this section extracts insights from the viewpoint of the experts at the forefront. These insights will build upon the identified societal values from Chapter 4 and the previous section about the legal and industry views.

The MCDA was chosen for its capacity to handle multiple, often conflicting criteria, ensuring that all relevant criteria gathered from legal and industry documents are considered (Baran-Kooiker et al., 2018). The MCDA template used is shown in Appendix E. The methodology was applied in the following steps:

1. **Criteria definition**: 21 values derived from the legal and industry documents served as the criteria for the MCDA.
2. **Perspective of the value framework**: The participants are asked what perspective the value framework should cover to understand why certain decisions are made.
3. **Stakeholder identification**: Participants working at a strategic and operational level, directly or indirectly involved in the PU case, were identified as the primary stakeholders for this questionnaire. The perspectives of legal, human resource, risk, technology, and underwriting are chosen to cover the various views within the value framework. These stakeholders, mentioned in Chapter 3.3.3, function as intermediaries between legal entities, the industry, customers, agents, and AI developers.
4. **Prioritization assignment**: The PROMOTHEE method is used within the questionnaire for the stakeholders. Each participant was asked to prioritize values from the 21 values related to their expert role. The qualitative outputs were used to understand what is understood under the different values and how to apply it in the value framework. Three stars were defined as 'core value.' An explanation must be given for the selected values.
5. **Analysis**: The weights and qualitative answers were analyzed to identify commonalities and discrepancies among the stakeholders.

The prioritization of values by participants highlights accountability, understandability, fairness, reliability, trustworthiness, accessibility, accuracy, responsibility, transparency, inclusiveness, integrity, privacy, security, and effectiveness as paramount. The study's value selection incorporated at least three values that diverged from social values to encapsulate the business perspective, with the number "three" being chosen without specific rationale. Appendix F shows aggregated results of the prioritization of the participants, synthesizing participants' comments and legislative references that underscore this connection.

The participants prioritize three values that do not directly adhere to social values. The gathered empirical findings from the questionnaires indicate the following results: accountability, responsibility, and effectiveness. Accountability was universally endorsed as an important value by all stakeholders. Reflective remarks, such as the assertion that "*clear accountability for a specific system is necessary to ensure proper governance and control*" and the recognition that "*defending our decision-making processes to varied*

stakeholders is crucial," further underscored this consensus. Furthermore, the pursuit of responsibility within the value framework was affirmed, with the LIAJ acknowledging responsibility as an essential value in its Code of Conduct (2018, p. 6).

One disagreement among participants pertains to the effectiveness of the AI system. As mentioned earlier, the difference can be attributed to the divergent approaches towards the framework. While some participants perceive the value framework as a means to identify ethical and business risks, others view it solely as an ethical framework. This diverging perspective leads to a divisive debate regarding including control mechanisms within the framework. However, economic incentives drive the use case. Therefore, the decision was made to retain it to identify conflict areas with this value during the identification of the norms phase at a later stage.

Last, there was a difference in prioritization between stakeholders about societal values. For instance, the variations in participants' viewpoints regarding transparency and explainability in AI can be attributed to their unique roles. The risk department emphasizes the need of transparency concerning financial risks, ensuring accuracy, stability, and clear outcomes for stakeholders. In contrast, the technology department prioritizes transparency differently. It recognizes the complexity of AI systems, which can present obstacles to achieving complete transparency, but we must be able to explain what the impact is. However, as this value must be attained from a societal perspective, the differences between the business stakeholders' value perceptions lie outside this research's scope but are worth further research.

4.5. Conclusion

This chapter analyzes the values that must be protected while integrating and using the PU system within an organizational setting. The values are identified from a societal, legal, industry, and business perspective. This results in the following values: accountability, responsibility, effectiveness, continuous improvement, fairness, usability, privacy, security, trust, robustness, transparency, understandability and explainability. The identification of the values answered the following sub-question:

"Q2: Which values apply to the use case?"

Examining the sub-question results in various values that may be at risk through designing, developing, and deploying AI systems. This chapter highlights the need to incorporate different methods and information sources to understand stakeholders' perspectives of values. Therefore, the framework of Bauer and Herder (2009), inspired by Williamson's (2000) four-layer scheme, provided the focus of every stakeholder. The chapter shows that the AI application requires different information sources, such as legal requirements, industry standards, business input, and academic literature. A concise collection of overarching values relevant to the PU use case context has been distilled by integrating multiple viewpoints. However, these values are still broad and not actionable. Therefore, these values are the backbone of the next sub-question. Table 4.2 shows an overview of the values from each stakeholder perspective, which will support the identification of the norms.

In the PU systems use case, it takes different information sources and methods to identify the relevant stakeholders' values. First, societal values are identified through an integrative literature review, focusing on the social norms of AI within Japan in an organizational context. The identified social norms are linked to the values that raise concerns in organizational settings. Therefore, seven relevant values were identified: fairness, explainability, usability, privacy, security, transparency, trustworthiness, understandability, and robustness. In addition, four social norms must be considered in developing the value framework and guidelines: contributing to community, wholeness, sincerity, and sensitivity.

The authority and industry values are identified through a content and comparison analysis of relevant grey documents. A comparison was made using European documentation surrounding AI to understand what values must be protected from the Japanese government's and industry's points of view. The comparison showed how most values were related to technical operations and had the same definition. One major difference is the value of continuous improvement as a strategy for policies and risk-based approaches. This value is often mentioned in Japanese documentation, which was not mentioned in European documentation. The business values are identified by questioning the management about what values must be protected within the AI lifecycle. The management participants performed a questionnaire using a multi-actor decision-making template to prioritize their choices and space to substantiate them. The questionnaires resulted in mostly the same values that must be protected as stated by society, authorities,

	Society view	Legal & industry view	Business view
Accountability			
Responsibility			
Effectiveness			
Continuous improvement			
Fairness			
Explainability			
Usability			
Privacy			
Security			
Transparency			
Trustworthy			
Understandability			
Robustness			

Table 4.2: Values per stakeholder. The blue boxes show which values the stakeholder group mentions as important to protect.

and industry. However, three values have been added to protect organizations' operations from risks. These values are accountability, responsibility, and effectiveness. In addition, trade-offs have been made as every expert filled in the values from their perspective. The trade-off resulted in adding effectiveness as a value.

The practical contribution is shown in the identified values to draw an understanding for AI system developers of what to consider while developing and deploying AI systems. Identification is also a first step in creating the value framework. Also, different methods were utilized to gather information about stakeholders's values. In this research, methods that are accessible in the environment are used. Other methods could be applicable for the identification of stakeholders's values.

5

Norms

This chapter is within the third research phase and uses the knowledge base and practical environment to produce results. This chapter operationalizes the values delineated in the preceding sections, translating them into actionable norms from a safety perspective. The chapter answers the following research sub-question:

"Q3: Which norms emerge from the use case?"

Answering the sub-question informs AI system developers what constraints are attached to safeguarding values in the use case's context. Furthermore, it offers guidance to the knowledge base regarding applying various concepts to transforming values into norms within an organizational setting.

5.1. Empirical set up

An empirical study was systematically orchestrated to explain the norms that emerge from the predictive underwriting system within the Japanese life insurance domain. Eight participants were engaged in semi-structured interviews and workshops, representing each key role in predictive underwriting. From front-line end-users to back-end data scientists, these individuals were selected to ensure a holistic representation of the predictive underwriting ecosystem. The participants, including the operators, examiners, and developers, were selected based on their operational work in developing and implementing the PU system. The details of the interviewees, including their interview ID and corresponding role within the AI system, are cataloged in Table 5.1 for reference. Appendix G shows examples of the informed consent templates signed by the participants.

The participants were invited to fill out the identical MCDA questionnaire as the management members. This MCDA assists participants in responding to questions posed during the workshop. The questionnaire aids in providing the developers and examiners with a prioritized list of values. The template is documented in E

Examiners and developers may utilize these values during the workshop. Throughout the session, developers and examiners review each phase of the PU system's life cycle, highlighting the points at which its value must be met. Next, the participants are asked to enumerate any risks related to the significance of that specific process step. The last question concerns the available risk control methods at the moment.

The last step is to identify norms per value and use the information from the workshop and questionnaires to conduct semi-structured interviews. For every participant, there have been sets of questions. The questions with quotes from the interviews have been listed in Appendix J. The questions are meant to get more in-depth information about certain topics. For example, fairness is one of the main values a data scientist prioritizes. During the workshop, the participant appointed fairness as a value to protect in the process step '*model outcomes with a decision whether or not to accept the risk per customer*.' The risk mentioned is '*outcomes not used fairly*.' Because the spot was left open on the question about control mechanisms, a question is asked during the semi-structured interviews if there are any control mechanisms for fairness regarding that process step. Questions like '*Do we measure fairness in the current state?*' and '*Do we have tools to measure fairness?*' are questioned in the semi-structured interviews.

To preserve the integrity of the gathered data, every interview was transcribed structurally. A qualitative analysis was performed using the Copilot software to examine and interpret the data methodically. This

procedure utilized a classification methodology based on the values that encapsulated the stakeholders' insights.

Participant function	Stakeholder role	Participant ID
End-user (underwriter)	Operator	1
Information Risk Manager	Examiner	2
Operation Risk Manager	Examiner	3
Operational Risk Analyst	Examiner	4
Data scientist	Developer	5
Data scientist manager	Developer	6
Data & AI manager	Developer	7
Financial Risk manager	Examiner	8

Table 5.1: Participants of the questionnaire, workshops, and semi-structured interviews. The participants are provided with their role and ID.

5.2. Specification of the use case

Predictive underwriting encompasses diverse stakeholders, each with vested interests ranging from the development and maintenance to the operation, audit, and regulation of predictive underwriting models. The need for predictive underwriting systems has been explained in Chapter 3, with the operational process and development of these systems delineated in Sections 3.1 and 3.3.1. Given the extensive nature of stakeholder involvement, this research has pinpointed the most pertinent stakeholders. These key stakeholders are enumerated in 5.1, relating their roles to the AI system functions as characterized by Zednik (2019).

The subsequent subsections address the use case's problem statement, each participant's roles in the use case, the control mechanisms employed in the current process, and the interconnection between these aspects to provide an understanding. Additionally, the factors influencing the data collection process are examined.

5.2.1. Problem definition

The problem definition was previously stated but will be reiterated here for convenience. The identified problem concerns the company's desire to incorporate AI into its underwriting systems. This objective aims to enhance decision-making accuracy and efficient work processes. Nevertheless, the company lacks a standard process to protect stakeholders' values. Failure to effectively manage potential risks can significantly damage an organization's reputation, potentially leading to its demise. To tackle this problem, AI system developers must understand what values must be protected and how to protect them.

5.2.2. Stakeholders involvement

Table 5.1 demonstrates the presence of three distinct stakeholder roles in the PU use case, namely operators, examiners, and developers. All participants were asked to complete the identical MCDA questionnaire during a focus group, previously administered and analyzed in Chapter 4 as part of the organization's strategic layer. The question asked during the focus group was, '*what value do you want to attain within your expert role perspective?*'. Incorporating this analysis makes it evident which individuals within the executive hierarchy should be approached for specific inquiries.

End-users (operators): As the end user, the underwriter is responsible for accurately evaluating risk for underwriting purposes on a customer-specific level (End-user, October 17, 2023). The end user bears responsibility for the agent and the customer, as they receive and are impacted by the assessment. The PU system is valuable for enhancing decision-making accuracy and streamlining work processes. Furthermore, the underwriter must possess the ability to articulate the rationale behind its decision-making process and the primary driver. The agent and the customer can ask for an explanation during potential rejection. Both the agent and the customer require a fair and reliable assessment. The underwriter utilizes business rules to regulate these requirements in the manual process. The main drivers identified by the end-user encompass a range of values, including accountability, adaptability, effectiveness, continuous improvement, privacy, reliability, responsibility, security, transparency, and understandability.

Risk managers (examiners): The risk managers are responsible for the development and operational

activities implemented to mitigate risk scenarios (IRM, October 17, 2023; ORM, October 17, 2023). Each risk manager is accountable for managing a team that collaboratively works towards effectively and efficiently mitigating risks. There are three distinct risk management teams, namely operational risk management (ORM), information risk management (IRM), and financial risk management (FERM). Every team possesses a distinct interest, whether direct or indirect, in the efficient workflow and accurate decision-making processes of the PU system.

First, the ORM department. The operational risk management (ORM) department participants are responsible for mitigating risks within the operations and implementations of business processes. ORM focuses on IT governance, using predefined procedures to support the first-line divisions in controlling risks. The primary objective of ORM in the PU use case is to guarantee the seamless functioning of business processes, minimizing any operational disruptions that could impact the organization negatively. Therefore, it is indicated that privacy and security are the main areas ORM focuses on and values. These values should focus on accessibility, integrity, confidentiality, and understandability at the technical system level. Participants also added missing values within the MCDA schema, namely availability and confidentiality, which are also extremely important. In the current situation, ORM uses standard procedures such as IT and data governance through Information Security Management Systems (ISMS) and Confidentiality, Integrity, and Availability (CIA) assessments to control the risks at the technical system level.

Second, the FERM department. The Financial and Enterprise Risk Management (FERM) department holds responsibilities encompassing a spectrum of duties that sustain the insurance entity's operational integrity and strategic risk positioning. A relevant example is model governance, governing the underwriting models' lifecycle, from validation to ongoing performance evaluation, and upholding ethical modeling standards (Eggert, 2014). Other direct relations with the PU system are regulatory compliance, such as solvency and precision. The last important thing is portfolio surveillance, where FERM continuously scrutinizes the insurance portfolio, adapting underwriting decisions to the evolving risk landscape (Van Der Heide, 2023).

Consequently, the FERM's main priorities are transparent and understandable system design and development. It must be clear what methodology is used and how it is effective for the business. Other values mentioned are accountability, accuracy, integrity, reliability, safety, and stability. Currently, FERM does not control mechanisms for effectiveness and transparency.

The last one is the IRM department, which handles the information flow and mitigates the risks within business processes. The focus relies on data governance and uses predefined procedures to support first-line divisions in controlling information risks (IRM, October 17, 2023). To be more specific, IRM is responsible for identifying potential information security threats and vulnerabilities in AI systems. The interest of IRM in an accurate and efficient PU system is that the system complies with rules and regulations, as they are accountable for that part. Therefore, a data protection impact assessment is used as a control mechanism to measure privacy risks. The security risks are controlled through a business impact assessment, where each project must be documented, including the CIA, data quality assessment, and allocating responsibilities and accountability on the data governance level. Therefore, the main value assigned by IRM is privacy and security, controlled by integrity, stability, transparency, understandability, effectiveness, accountability, and responsibility for the data used within the PU system.

Data scientists (developers): Data scientists are responsible for developing AI systems that assist end users in their business process activities, enabling them to make more accurate decisions and optimize their workflows. To attain this design objective, risk owners assign data scientists to ensure compliance with industry regulations and address integrity risks concerning their portfolios. Data scientists are pivotal in the PU process as they are responsible for developing and maintaining PU systems. Consequently, they bear a significant responsibility in comprehending the impact of these systems on stakeholders. Data scientists have a compelling rationale for integrating a value framework via a risk-based approach, primarily due to their role in facilitating and developing AI systems for business stakeholders. A risk-based approach provides data scientists with a framework to design, develop, and deploy AI systems effectively and safely. Hence, their emphasis within the system manifests in diverse values throughout each life cycle stage. The primary focus lies on trust, which can be categorized into various aspects such as effectiveness, fairness, privacy, security, transparency, understandability, reliability, safety, stability, and accountability at the system level.

5.2.3. Technical specifications

When identifying risks, it is important to consider the technical details of the PU system:

- The system runs on an online running server.
- The system uses financial and medical data.
- The system does use external data.
- The model does not use external data.
- The system uses a combination of models to arrive at its results.
- The outcomes of the system are visible to the sales and underwriting teams through an online server.
- The system is updated once every six months.

5.3. Method: concepts of system safety

Systems theory originates from our aim of developing increasingly complex systems. Traditional science involves dissecting systems into isolated physical components and reducing behavior to discrete events over time (Smith & Johnson, 2019). Modern systems theory aims to address organized complexity, which refers to systems that are too complex for in-depth analysis and too organized for statistical methods (Leveson, 2012). The focus shifts from isolated parts to holistic systems, acknowledging that an understanding of certain system properties requires considering all elements, including the interplay between social and technical aspects.

Within PU, systems safety theory transcends traditional analysis by advocating a holistic examination of systems for two reasons. First, the holistic approach is predicated on the understanding that the characteristics of the PU system are not solely derived from its elements but rather from the symbiotic relationships and interactions among them (Leveson, 2012). As a result, safety in PU is seen as an emergent property that cannot be found by looking at discrete data inputs or model outputs separately.

Second, understanding system safety theory presents significant prospects for the Japanese insurance industry, given their commitment to protecting AI-driven stakeholders' values. The holistic approach to protecting stakeholders' values also aligns with the Japanese cultural idea of "wholeness," which is explained in Chapter 4.2.1. The complexity of the PU system relies on the different stakeholders involved and the impact of the system on other subsystems. Therefore, there is a need for organized interactions, which highlight the significance of guaranteeing both efficiency and ethical conduct (Ashok et al., 2022). By considering all the interactions within the AI system, insurers can better anticipate and control the multifaceted risks associated with AI underwriting, thereby establishing a safer and more reliable application of the PU use case.

This multifaceted interdependency implies that what is a safe underwriting practice under certain conditions might, under alternative circumstances, introduce risks. Hence, applying system safety theory to AI-driven underwriting is instrumental. It shows how important it is to have a safety design that incorporates safety into the system architecture ahead of time. It mitigates potential risks before they lead to widespread failures (Leveson, 2012). The approach aids in protecting stakeholder values by identifying the situations in which these values are at risk.

5.3.1. Systems safety theory and design for values

The relationship between system safety and design for values is linked through the concepts of norms and identifying safety constraints. Van de Poel (2013) defines norms as guidelines to achieve a goal. According to Leveson (2012), safety constraints represent acceptable ways the system or organization can achieve its mission goals. In addition, Leveson (2012, p. 12) describes a capability of system safety as "*Some systems safety is part of the mission or reason for existence, such as air traffic control or healthcare, in others safety is not the mission but instead is a safety constraint on how the mission can be achieved.*"

Take PU as an example: efficiency is a goal, and privacy is a safety constraint in the PU process. In certain systems, privacy is not the primary mission, but instead, safety constrains how the mission objectives can be realized. Rigorously adhering to privacy constraints may require refraining from building or operating the system altogether. The most robust safeguard against data breaches and privacy violations within the PU process is to refrain from collecting certain sensitive customer data. Acknowledging the inevitability of

compromise, complete abstention from data collection may not always be feasible. Apart from abstaining from certain data collection, the most robust design protections for privacy also decrease the likelihood of privacy breaches when data utilization is deemed for achieving efficiency objectives. The design and operation of the PU process involve a complex undertaking of balancing efficiency goals with the safety constraints imposed by privacy.

Leveson (2012) argues that systems safety theory adopts a systems theory perspective on causality to identify potential accidents within complex sociotechnical systems. This study focuses on a value framework with a risk-based approach that emphasizes the achievement of emergent properties through accident prevention. Systems theory considers hierarchical structures, wherein each level imposes safety constraints on the activity of the level beneath it, thereby enabling or regulating behavior at lower levels based on the presence or absence of safety constraints at higher levels. This construction resembles Van de Poel's (2013) hierarchy of values. By integrating the design for values and the control structure for system safety, experts from different domains can contextualize and identify the safety-constraining norms to achieve the desired values step by step. This structured approach allows for specific questions at each phase of the AI life cycle. The subsequent sections explain the functioning of this process.

5.3.2. Safety control structures in predictive underwriting

Within the PU system, the foundational element of system safety resides in the collective understanding and delineation of accident types and potential losses. This understanding is cultivated through multi-stakeholder consensus input from customers, governmental bodies, and regulatory entities, all of whom contribute to a paradigm where safety is prioritized in the advent of AI applications (Leveson, 2012). Safety prioritization' is not merely a technical effort but an institutional commitment reflected in the structural design of systems, often termed the safety control structure.

The safety control structure is the institutional backbone for managing hazards, transcending beyond the remit of system developers, who may lack control over all accident-related conditions (Leveson, 2012). In this broader framework, hazard identification and control responsibilities are appropriately allocated. For instance, in PU, the system boundary must encapsulate the technological aspects, data provenance, and application contexts. The boundary setting ensures that institutional biases are factored into the safety analysis, averting a superficial deployment of predictive tools (Leveson, 2012).

Figure 5.1 shows the socio-technical hierarchical safety control structure in the PU system. This model is adapted from Leveson's (2012) example but adjusted to the applicability of the PU system. It shows two structures, namely the development and operational sides, based on several layers, with the parties communicating in between. For instance, the development phase in PU involves designing systems that accurately predict risk. In contrast, the operational phase involves applying these systems to customer data. Leveson (2012) illustrates with this structure that effective safety relies on the interplay between the different layers, calling for open communication channels for sharing safety assumptions and operational feedback.

Notable parallels emerge when integrating van de Poel's (2013) concept of a value hierarchy with Leveson's (2012) safety control structure. The value hierarchy framework centers around general values, where the conversion to design requirements is contingent on the specific context. Leveson's safety control structure illustrates how, from a higher hierarchical level, such as legislative institutions, control mechanisms are distributed through the system's development. These mechanisms traverse various layers, adapting to fit the contextual requirements of each level.

The utility of the system control structure in this research is particularly evident in its capacity to translate high-level values into actionable norms and design requirements. This translation process involves identifying potential risks within the development and operational phases where these values might be compromised. As shown in figure 5.1, values at each level can be systematically mapped. Starting from the national legislation level, where a specific set of values is reported in legal documents, these values are then adapted and expanded by the industry to align with industrial applications and further communicated down to the operational layer of an organization. The organizational, operational layer, depicted in the blue box of figure 5.1, needs each process step to protect values sourced from legal, industrial, and organizational levels. For this research, there needs to be a review with experts of potential risks that could undermine these values at each process step. Subsequently, this information is given

by operational management experts through workshops and semi-structured interviews. This structured approach underpins the safety of values throughout the AI system's operational life cycle.

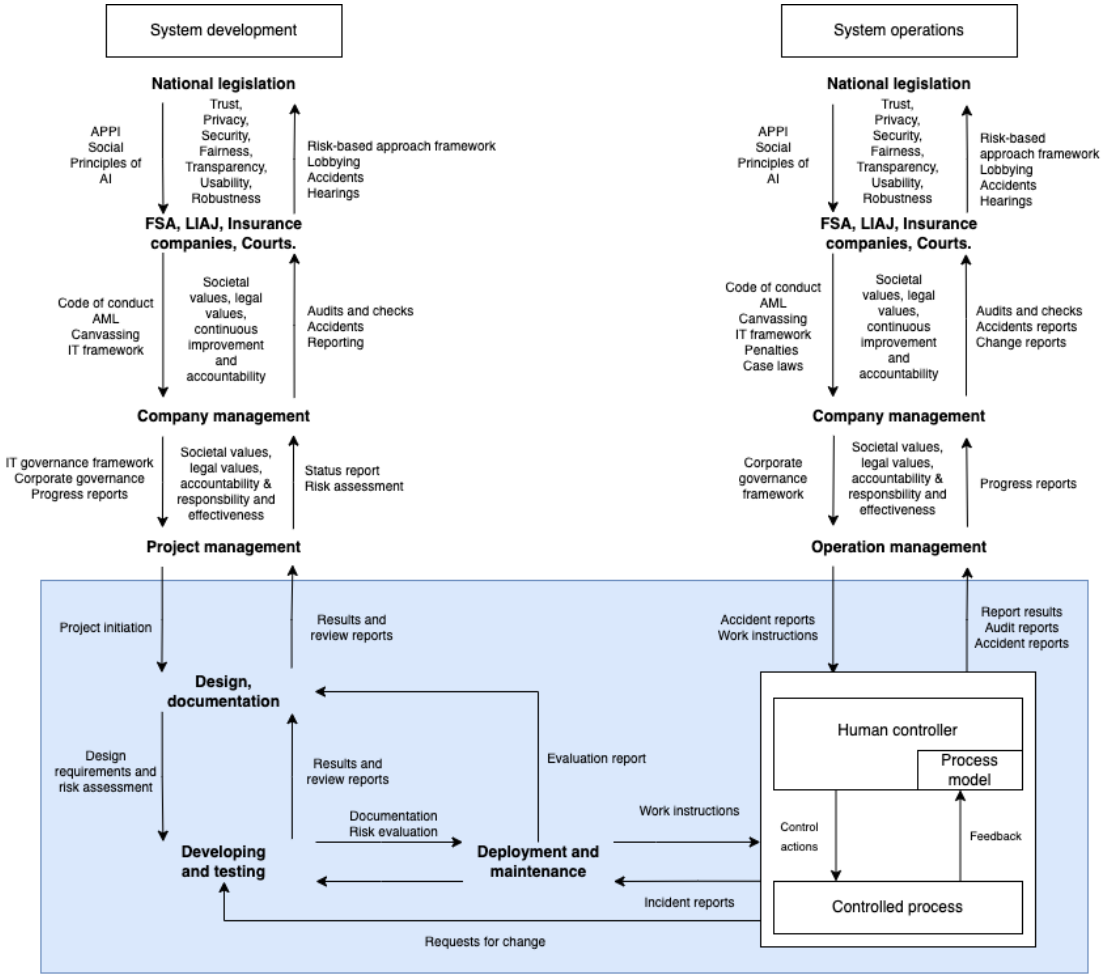


Figure 5.1: Hierarchical Safety Control Structure of the PU system, inspired by Leveson (2012). The identified values that must be protected are mapped per layer. The blue box shows the current operational layer that must protect the values of the higher layers and achieve the goal of the PU system.

Dobbe (2022) complements Leveson's (2012) view by advocating for delineating system boundaries to translate hazards into concrete requirements, acknowledging that while designers may not control all accident-related conditions, those within their purview must be effectively managed. Dobbe (2022) and Leveson (2012) underscore the need for communication between the development and operational phases to ensure system safety, with institutional context playing a pivotal role in responsibility allocation. In addition, Dobbe (2022) mentioned that the control structure must consider the institutional context when identifying risks. This statement aligns with Leveson's safety control structure, which states that the safety control structure is the institutional backbone. Figure 5.2 is a translation of these statements, where the risk analysis is the bridge between the institutional and technical constraints. For this research, the focus relies on institutional constraints because of the limited time and relevance. This research prioritizes the analysis of institutional constraints and risks to identify and protect values.

5.3.3. Method to translate values into norms

The preceding subsections explain system safety and justify its suitability for addressing the problem. This subsection explains the utilization of relevant concepts in gathering empirical data and transforming it into norms. A template has been created to assist in workshops with experts to identify risks and establish safety constraints. The template is depicted in Figure 5.3, and a more detailed version can be found in Appendix H. The workshop results will inform semi-structured interviews, which will be used to develop

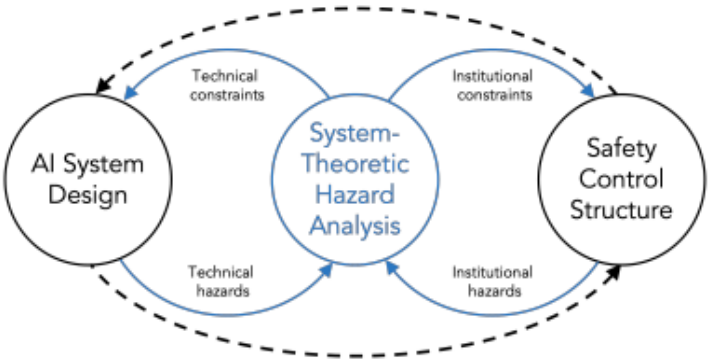


Figure 5.2: Hazard analysis informs AI system design and institutional safety control structure design. Adapted from (Dobbe, 2022, p. 4).

norms and finalize the value framework. The outcome offers AI developers a set of guidelines that can be applied throughout the AI life cycle to protect stakeholder values. These norms are currently impractical because they address all potential risks instead of solely prioritizing stakeholder interests in the design and use of the PU system.

The template was created by analyzing the operational flow of the safety control system, shown in the blue box of Figure 5.1. This template shows the role of communication as the input in the operational process and its destination as the output. The template was structured based on the AI lifecycle of the PU system to provide a more organized and concrete framework for the experts participating in the workshop and interviews. The AI lifecycle was established by reviewing relevant literature, conducting interviews, and analyzing relevant documents. The lifecycle is divided into three phases: design, development, and deployment (Alahakoon, 2022; Haakman et al., 2022; Data scientist, October 17, 2023). Appendix I provides a detailed description of the three phases.

The following questions were used during the workshops to acquire empirical data to gather safety constraints and risks:

- *Values at stake:* The discussion's central focus was identifying values potentially affecting each process stage. These values are the ethical and operational foundation for the system's integrity. By precisely identifying the specific value in question, such as transparency, fairness, or privacy, individuals can effectively link their concerns about safety to concrete and outcome-oriented criteria.
- *Risk Potential:* The second aspect of the investigation focused on the risk landscape, assessing the vulnerability of each process step to particular risks. Stakeholders were prompted to identify and express potential pitfalls that may jeopardize the system's safety or the ethical principles upheld by the organization. The forward-looking perspective played a role in anticipating the manifestation of risks and influencing the development of preventive measures.
- *Prohibited System Actions:* The inquiry concluded with a focus on actions that the system must refrain from to uphold the identified values and effectively mitigate risks. The prescriptive approach enabled the establishment of explicit boundaries and operational guidelines, ensuring the AI system's adherence to safety measures and integrating them into its architecture.

The workshop yielded a set of safety constraints developed through the participants' collaborative efforts and extensive knowledge. The constraints were applied to the AI lifecycle, establishing a safety-conscious framework that ensures alignment with ethical standards and organizational risk appetite. By using this approach, exploring the values, risks, and potential control mechanisms at each stage of the AI lifecycle ensures an understanding of the system's safety and what could endanger the stakeholders' values. However, one must mention that these norms are established by considering one use case. Therefore, iterations of different use cases are needed to gather more expert knowledge about risks that could arise from the use and development of AI systems.

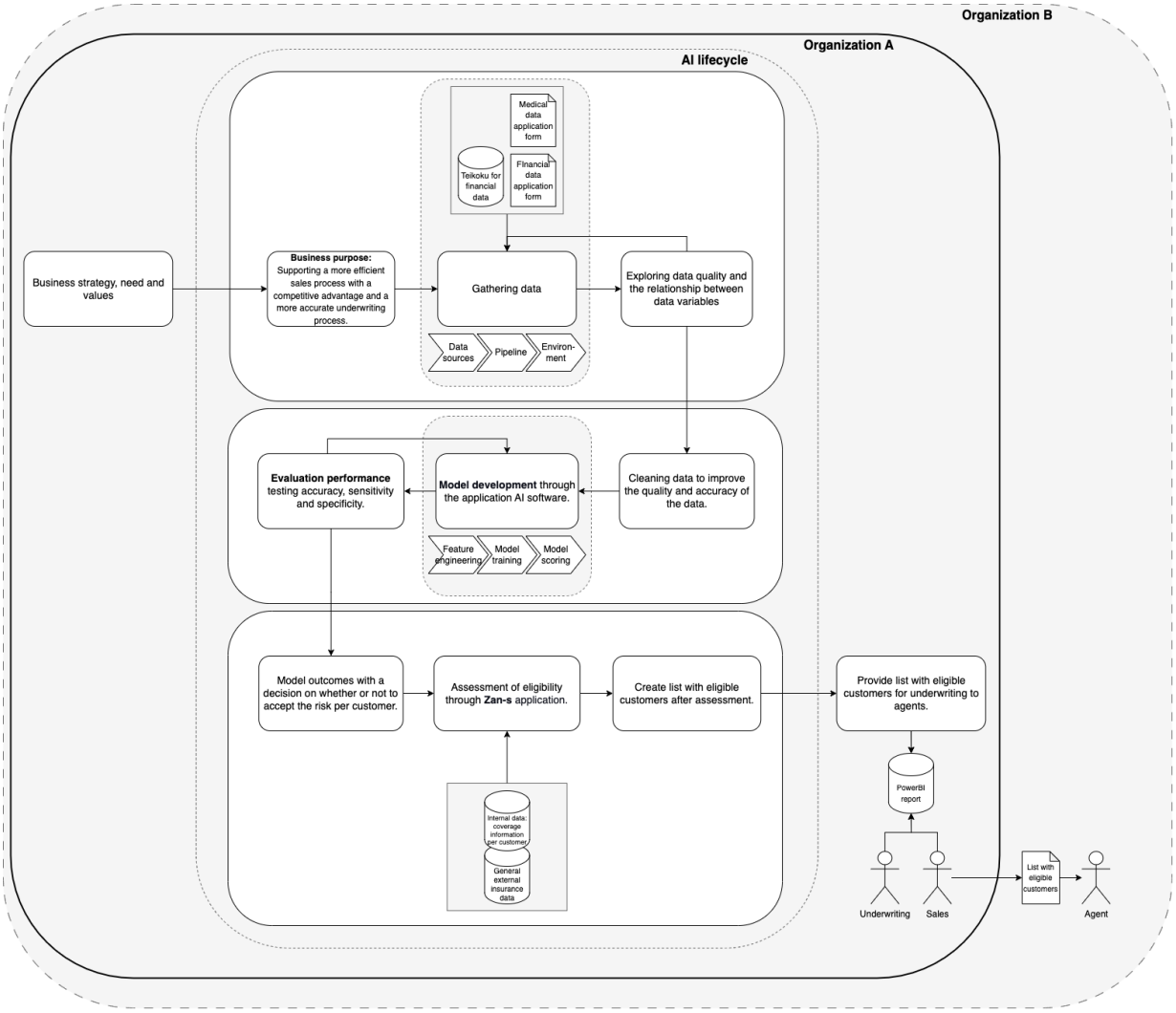


Figure 5.3: AI life cycle template used during the workshop with participants to identify risks within the PU system assuming no controls.

5.4. The design of the value framework

The value framework emerged from an amalgamation of expert input, legislative mandates, and industry standards. The following sections will explore creating a value framework, focusing on converting values into norms and integrating them into the operational context. First, using system safety concepts during workshops and interviews resulted in 84 norms. These norms were translated to the same abstraction level, which resulted in 54 norms. A method utilized by Garst et al. (2022) for reporting standards in an organizational context is followed to depict the relevant deriving of the 84 norms and the methodology employed to consolidate them into a concise set of 54 norms that could be differentiated into two categories, which results in 31 assessment norms and 23 process norms. The following chapters explain the method to remove redundancy in the value framework to make it operational.

5.4.1. Results of the empirical cycle

Table 5.2 presents an example overview of the alignment between stakeholders and the values they want to protect. The clarification process involved the utilization of value worksheets to ascertain the specific topics within the organization that each expert is engaged with. The value worksheets facilitated the identification of supplementary underlying themes, eliciting focused questions for the semi-structured interviews. For example, The underwriter underscored the significance of the reliability of the PU system. Therefore, the question was asked, ”*How is the reliability of the manual underwriting process measured?*”.

The interview questions for each participant are outlined in Appendix J.

Interviews	Stakeholder role	values
End-user (underwriter)	Operator	Accountability, effectiveness, continuous improvement, privacy, responsibility, security, transparency, trustworthiness and understandability
Information Risk Manager	Examiner	Effectiveness, privacy, responsibility, security, transparency, trustworthy, understandability
Operation Risk Manager	Examiner	Privacy, security and understandability
Data scientist & manager	Developers	Accountability, effectiveness, fairness, privacy, responsibility, transparency, trustworthy and understandability
Financial Risk manager	Examiner	Accountability, effectiveness, transparency, trustworthy and understandability

Table 5.2: Results of questionnaires shown per participant. The table shows the participant, their role, and the values they consider as a risk within the PU system.

Table 5.3 shows three norms established through interviews and rules and regulations. The selection of these three norms is arbitrary, exemplifying how they were created. Table 5.3 delineates three parts. The first column refers to the source of the norms from expert knowledge in the operational domain. The second column refers to the authorities and industry's laws and guidelines. The third column examines the presence of a pre-existing control standard mandated by law. This process resulted in 54 norms, shown in Appendix K. The process of how these norms were established with the three norms from table 5.3 is explained in the following paragraphs.

	Interviews	Laws & guidelines	Standard control
Robustness: *Implement ongoing assessments and monitoring of the AI system's data quality, integrity, availability, accessibility, and confidentiality.*	ORM	FSA & APPI	YES
Fairness : *Ensure AI systems promote fairness and avoid excessive bias towards specific stakeholders in wealth and society.*		Social Principles for AI	NO
Privacy: *Activities related to customer and organization data should be preempted with impact assessments to mitigate risks and safeguard customer interests.*	IRM	APPI	YES

Table 5.3: Example of norms from the value framework derived from regulations and expert knowledge. The interview column indicates the individuals who mentioned the norm. The laws and guidelines reference the relevant legal and industry documents. The final column about standard controls indicates whether the legal or industry documents offer standardized controls for organizational utilization. Existing standardized controls are assigned with a YES, missing controls are assigned with NO.

Robustness The first norm is regarding the value robustness. Achieving robustness in an AI system relies on its ability to manage and maintain data quality, integrity, availability, accessibility, and confidentiality through continuous monitoring and refinement. The concept can be delineated into two distinct parts. The first part arises from the organization's internal structure, which conforms to established protocols governing these aspects. Adherence to data quality and integrity is important in predictive underwriting, as it directly impacts the accuracy and reliability of risk assessments (End-user, October 17, 2023). Individuals within the organization are entrusted with scrutinizing system adherence to these regulations, ensuring compliance. The organizational framework includes a range of control systems that incorporate predefined thresholds for confidentiality, integrity, and availability, in addition to compliance with the Information Security Management System (ISMS) standards, which are obligatory for high-risk systems (IRM, October 17, 2023; ORM, October 17, 2023). Stringent controls are essential in predictive underwriting to protect

sensitive financial data from breaches and misuse, thereby upholding the credibility and efficacy of the system. The organizational controls stem from the guidelines established by the Financial Service Agency (FSA). The FSA's latest documentation refers to these controls as "*whether the company is managing information security by designating individuals responsible for it and clarifying their roles and responsibilities to maintain the confidentiality, integrity, and availability of information*"(Financial Service Agency, 2023, p. 92).

The second aspect of the norm robustness pertains to the need for ongoing assessments, as indicated in the FSA documentation as *"Furthermore, it involves the consistent endeavor to enhance the information security control environment through the PDCA cycle..."* (FSA, 2023, p. 91). Continuous improvement is compulsory within the context of predictive underwriting. The adaptability and proactive evolution of financial systems are needed due to the dynamic nature of markets and risk factors, enabling anticipation of future trends and vulnerabilities. The organizational standards embody a holistic commitment to enhancing information security practices, encompassing the pursuit of continuous improvement.

Fairness The second norm is regarding fairness. The government mentions fairness in the social principles for human-centered AI. The Cabinet Secretariat says in this document, "*The use of AI should not generate a situation where wealth and social influence are unfairly biased towards certain stakeholders"* (Cabinet Secretariat: Council for Social Principles of Human-centric AI, 2019, p. 10). The quote highlights the significance of fairness within AI-facilitated decision-making procedures. In predictive underwriting, algorithms must avoid unjust discrimination against specific groups or stakeholders. The permissibility of avoiding unjust discrimination is contingent solely upon objective criteria duly substantiated by experts such as medical professionals or actuaries (End-user, October 17, 2023).

Table 5.3 shows no concrete control mechanism or guidance for fairness provided by authorities or industry. The organization possesses tools to ensure fairness (Data Scientist and Manager, October 17, 2023). The following chapter on design requirements will guide designing controls within their context. Moreover, it is important to highlight that fairness is a realm that is fundamentally regulated by expertise, and the organization presently confronts a scarcity of knowledge in this particular sphere. Hence, a compelling mandate arises, demanding a collective endeavor to retrain the labor force. The reskilling initiative is crucial for developing a nuanced understanding of interpreting each use case within the complex realm of fairness metrics (D&AI manager, October 17, 2023).

Privacy The last example from table 5.3 concerns privacy, demanding safety constraints. Designing AI systems using personal or organizational data must always be assessed through an impact assessment. The comprehension and utilization of customers' data require an assessment, enabling the implementation of precautions in response to elevated risks. The PU system uses personal and sensitive information, elevating privacy as a paramount concern (IRM, October 17, 2023). The origin of this norm can be traced to the urgent need expressed by experts for a clear explanation of the influence of data in high-risk scenarios and the ethical handling of personal information (IRM, October 17, 2023).

The norm can be deconstructed into three discrete parts. The first part comprises personal data stipulated by the APPI regulation, which includes information such as a name, date of birth, or other identifying factors (Ministry of Justice, 2003, p. 2). The organization already has an extended control for this norm, namely a Privacy Officer who takes care of impact assessments. The APPI does not explicitly articulate the privacy officer's responsibilities as stipulated in the legislation. The organizational sphere has intentionally designated the Privacy Officer with a distinct responsibility. The responsibility involves offering guidance and support in completing the data protection impact assessment (IRM, October 17, 2023). The subsequent element of the norm is the data impact assessment. The condition requires that this assessment is required solely in cases about utilizing novel customer data, where explicit consent from the customer has been obtained.

5.4.2. Converging process of norms

This subsection analyzes the convergence process of the framework's content, which includes 54 norms that address all potential risks within our organizational sphere. These complex and extensive norms require a strategic approach to identify a subset of utmost significance for focused control. The methodology aligns with Garst et al.'s (2022) advanced six-step process, which presents an approach for selecting pivotal subjects in sustainability evaluation:

1. **Identification of the process perspective:** This step focuses on what must be considered by the firm when designing and using AI about the values of stakeholders.
2. **Specifying the chosen topics:** Garst et al. (2022) describe the second step as an approach to subject specification, emphasizing relevance and contextualization. Given that the process steps focus on norms and the topics have already been selected, specifically the values of the ten identified stakeholders, it has been determined to approach this step from a different perspective. This step involves dividing it into two parts: defining and abstracting norms to specify them.
3. **Determining the information sources:** Choose resources encompassing stakeholders outside the company's value chain and network. This step establishes stakeholder criteria.
4. **Assigning scores:** This step involves collecting data to assign scores to topics and considering stakeholders' perspectives. This step is not included as it does not fit the Japanese social norm of 'wholeness.'
5. **Selecting the topics:** Select the topics that apply to the different stakeholder's criteria.
6. **Deciding the cycle:** the control and record of changes about the topics chosen. This step is not included as it is not yet relevant for refining norms but will be employed during the design requirements and guidelines.

We have adjusted these steps for our research as this process aims to converge the norms to make them relevant to the PU system. The first and second steps have already been done in the previous chapters. In addition, we add a new step as this is relevant to our process. The adjusted and applicable general steps for the converging process are:

1. **Define:** This step defines the norms from the empirical data collection and uses system safety concepts to identify potential risks. 54 norms were identified through this process.
2. **Categorization:** The norms underwent a two-part categorization process. First, they were divided into 23 process and 31 assessment norms. Then, the assessment norms were divided into 18 data and 13 AI assessment norms using the AI governance framework as a guide.
3. **Determining the information sources:** This step identifies the relevant stakeholders and their criteria for the AI system that could affect the system's goal.
4. **Selecting relevant topics:** This step narrows the assessment norms to the relevance of the PU system and stakeholder criteria.

These steps are taken to identify and adapt norms to the specific context, ensuring their practicality and effectiveness. The explanation for steps one and two can be found in Subsection 5.4.1. The following paragraphs discuss categorizing, identifying information sources, and selecting pertinent topics. Determining the cycle is a step that will be explained in developing design guidelines. By following these steps, developers of AI systems obtain a set of norms that pertain to the stakeholders' values. These norms are crucial for safeguarding the stakeholders' interests within an organizational setting. The findings inform the development criteria and guidelines for design. The subsequent paragraphs will explain these steps.

Categorization of topics

A step was added in this process, as we categorized the norms into process and assessment norms, resulting in 31 assessment norms and 23 process norms. Process norms are identified due to their representation of an ongoing activity characterized by active engagements. Conversely, assessment norms indicate a more static evaluation or review at specific intervals. Next, the concept of a safety control structure from system safety employed during the workshops maps the values identified in our value framework. The proposed framework enables stakeholders to better understand and identify instances within the AI lifecycle where their values are not sufficiently safeguarded without a control mechanism. Figure 5.4 illustrates the application of this safety control structure, encapsulating the insights gleaned from workshops and semi-structured interviews. This approach provides a structured overview, enabling an understanding of the interconnectedness between stakeholders' perceived risks and the corresponding values within the system.

Based on the mapped values from Figure 5.4, participants identified data-related concerns as a risk during the design process. The insights from expert interviews and workshops align with Mäntymäki

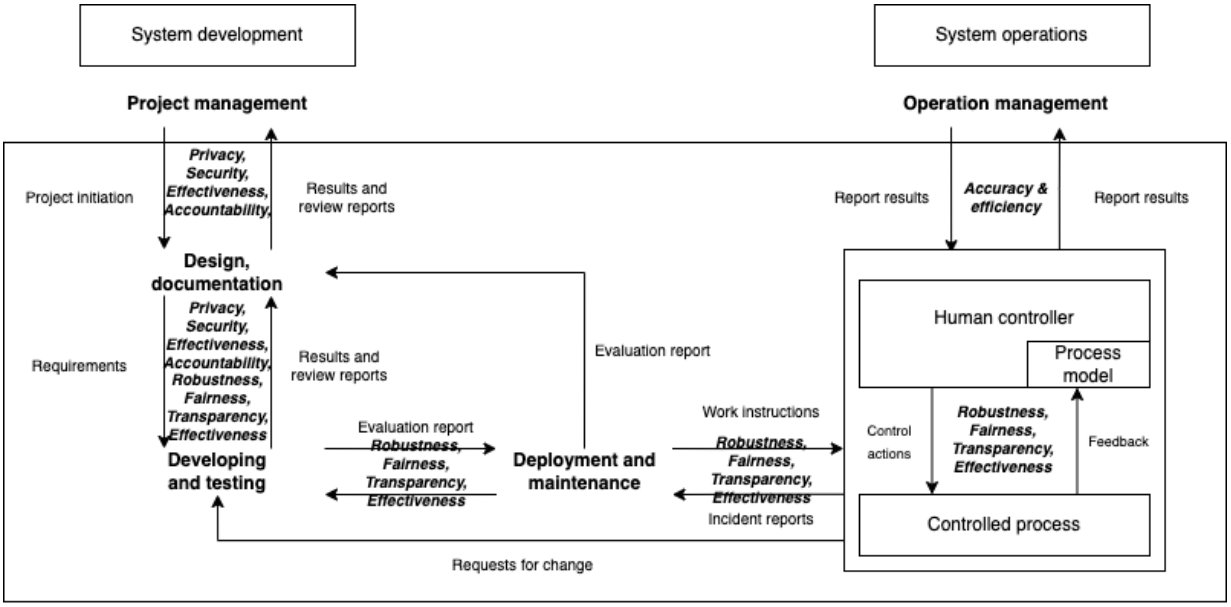


Figure 5.4: Safey control structure with the mapped values from stakeholders's input through interviews and workshops.

et al.'s (2022) governance framework, which identifies intersections between AI and Data governance. Mäntymäki et al. (2022) reviewed several academic literature and composed a definition for AI governance from a governance perspective. Furthermore, the definition connects the ties with the three other areas of governance: corporate governance, IT governance, and data governance. Mäntymäki et al. (2022) depict the relationships between an organization's governance areas. The framework shows that AI governance is a subset of corporate and IT governance, partially overlapping with data governance. Corporate governance serves as the overarching structure within an organization. AI systems, being a form of IT systems, fall under the purview of IT governance. While data is integral to AI operations, AI governance extends beyond traditional data governance, encompassing the broader aspects of AI system management. The definition and framework are further explained in Appendix L. Assessment norms can also be further categorized into data and AI assessment norms. The subdivision is done to make it suitable and organized for an organizational context. This step provides 12 data and 10 AI assessment norms. The filled-in assessments for the PU system are documented in Appendix M and Appendix N.

Upon revisiting the safety control structure, it can be observed that values related to data governance, such as privacy, security, robustness, and accountability, are prominently addressed. Privacy concerns focus on protecting customer data with norms such as "*Where feasible, personal data should be anonymized to protect customer privacy*." Security encompasses safeguarding customer data and information systems with norms such as "*Obtain the required certification to gain access to the necessary data*." Robustness is focused on data robustness, mostly on data quality, availability, confidentiality, etc. Accountability pertains to clearly delineating responsibilities in utilizing customer data, reflecting the governance framework's emphasis on managing and protecting data within AI systems, with norms such as "*Uphold data governance practices, requiring developers to obtain appropriate approvals for the use of customer data, ensuring adherence to privacy standards and security protocols*." In the development phase, key values pertinent to AI governance, specifically transparency, robustness, and fairness, come to the fore. These values are characterized as emergent properties, subject to variation with each step in the development process. This stage allows for determining the degree of transparency, assessing the model's performance robustness, and evaluating the model's adherence to fairness metrics. In the deployment phase, the emphasis shifts to the end-user, examining what aspects of the system need monitoring for effective user response. Key values such as robustness, transparency, fairness, security, and effectiveness become focal points, particularly as they are interrelated and susceptible to changes due to modifications in the system. Any alterations in technical functionalities have the potential to impact these values, so carefully consider how these changes influence the overall system outcomes and user experience. From this observation,

the data and AI assessment norms should be conducted at different stages, with a conclusion drawn from the workshop and semi-structured interviews. In the design phase, the priority should be to assess risks associated with data governance, while the development and deployment phases should center on evaluating risks related to AI governance.

Table 5.4 gives an example of three norms showing the categorization of norms, their information source, and the presence of established controls. The first norm in this table relates to the assessment of AI governance. This norm is linked to the results derived from the values integrated into the safety control framework. It emphasizes the importance of transparency and clarity during the developmental phase, particularly in explaining the construction of the model based on its technical capabilities. Further, the second norm highlighted in the table, identified as 'robustness,' falls under the category of data governance assessment. This classification is grounded in its focus on data rather than examining the technical competencies of the AI system. Nonetheless, this norm significantly influences the technical capabilities of the AI system. The concept of robustness is strategically positioned within the safety control structure, bridging the design and development stages. The last norm is a process norm, focusing on transparency. It clearly shows that a task must be done during the AI project to control the risks within the AI lifecycle. These process norms are input to shape the design guidelines.

Norm	Norm categorization	Source	Standardized control
Transparency: Verify that the AI development process and its results are transparent and understandable for all stakeholders, ensuring effective integration into user workflows.	AI governance assessment	End-user (interview)	NO
Transparency: Establish and communicate clear accountability and responsibility for the AI's development and operational processes within the organization.	Process	FSA (LAW) and mentioned by all participants (interviews)	YES
Robustness: Use assessments and monitor the AI system's data quality, integrity, availability, accessibility, and confidentiality.	Data governance assessment	FSA (LAW) and APPI (LAW), IRM (interview)	YES

Table 5.4: Overview of the three chosen arbitrary norms. The table displays the classification of norms based on their category, information source from law or interview, and whether standardized controls are provided by law. YES means that the law provides a standardized control, and NO indicates missing controls.

Determining the information sources

Information sources must be identified from which stakeholder criteria can be established to identify relevant topics of concern for stakeholders regarding AI systems. The PU system impacts four stakeholders, which can be divided into three groups: agent and client, society, and business. These groups are distinct from the ones identified for values, as the norms have been established through operationalizing the values. In this step, criteria are established from the perspective of stakeholders who may experience direct impacts.

The stakeholders in this thesis, the agent and client, obtain information from the underwriter's experience. The underwriter maintains direct communication with this group. According to the source, the agent and the client share a common interest in establishing trust in the system (End-user, October 17, 2023). Trust is established by ensuring the PU system consistently delivers fair, transparent, and precise decisions. The stakeholder perspective emerged when some individuals hesitated to embrace the PU system fully due to concerns about its accuracy and quality (End-user, October 17, 2023). There was a consistent emphasis on the importance of a transparent model that can effectively explain its outcomes, fostering understanding and accountability for customers. Hence, the subsequent criteria were established:

- **Stakeholder:** Ensure a positive impact on customer trust by avoiding unclear, unfair, and inaccurate decisions.

The societal information is derived from academic literature and government documents about AI in Japan. The societal criteria placed importance on upholding fundamental rights. The criteria were established for this purpose:

- **Society:** Ensure no violation of social norms

The organization's criteria were determined by comparing them with societal and stakeholder criteria, with a specific emphasis on the reliability and accuracy of the model (Data scientist, October 17, 2023). These findings are also a result of the objectives and limitations discussed in Chapter 3. Therefore, the following criteria are noted:

- **Business:** Ensure a positive impact on the accuracy and reliability of the results

Setting these criteria leads to selecting appropriate assessment questions for constructing the PU system. This step is needed for controlling the number of questions in an organizational context. It ensures that attention is directed towards the relevant components that require focus and allows stakeholders to safeguard their values.

Selecting topics
After categorizing, 31 assessment questions remained, with 19 about data and 12 on AI. The stakeholder criteria from the previous step can be utilized to identify the pertinent questions for the PU system filtration process. After applying stakeholder criteria, it is determined that 11 remaining data assessment norms and 10 AI assessment norms must be tested. These norms can be satisfied if suitable measures safeguard stakeholders' interests. The completed assessment templates for the PU system can be located in Appendices M and N. The applicable assessment norms can be used in setting design requirements; the process will be explained in the next chapter.

Let us take an example by considering the value and fairness of the use case and different criteria. The filtering process reduced five norms to three. Table 5.5 provides an illustrative instance related to the first three norms with an explanation. The blue boxes in table 5.5 signify a direct relation between the criteria and the norm.

	Business criteria: Ensure positive impact on accuracy and reliability of the results	**Customer and agent criteria:** Ensure positive impact on customer trust through avoiding unclear, unfair, and inaccurate decisions	**Society criteria:** Ensure no violation of social norms
Develop AI systems that avoid excessive bias towards specific stakeholders in wealth and society.	NO	NO	YES
Prevent unfair competition and excessive data collection by dominant companies in AI.	NO	NO	NO
Incorporate only those biases into AI system outcomes based on transparent and verifiable objective criteria, subject to regular review and consensus among stakeholders.	YES	YES	YES

Table 5.5: Example of selecting topics. The columns represent each stakeholder's criteria, and the rows represent arbitrarily chosen norms. An affirmative 'YES' signifies that the stakeholder criteria apply to the norm, whereas a negative 'NO' denotes the criteria's irrelevance to the norm.

The first norm is an illustrative exemplar: "*Develop AI systems that avoid excessive bias towards specific stakeholders in wealth and society*." The societal criteria in this context directly relate to the norm, originating

from the core values expounded in Chapter 4.2.1. This relation stems from the societal interaction with AI, where societal sincerity and the commitment to respectful conduct in human interactions are unmistakable in contemporary society (McStay, 2021). Therefore, we need to include this norm in our design.

Another example is the norm: "*prevent unfair competition and excessive data collection by dominant companies in AI.*" The manual underwriting process served as the basis for the data input of the PU system. The manual process employed an external database to acquire supplementary financial data about customers, as required. The process at the FSA is accepted and validated. The predetermined utilization of data is established. Hence, the applicability of this norm to the given use case is negligible.

The last example is the norm: "*Incorporate only those biases into AI system outcomes based on transparent and verifiable objective criteria, subject to regular review and stakeholder consensus.*" The application of norms in business is imperative due to the potential compromise in the accuracy and reliability of AI outcomes resulting from bias. Schwartz et al. (2022) highlight the significance of computational and statistical biases, emphasizing their partial contribution to the overall equation. Accurate and reliable AI systems require an understanding encompassing human and systemic biases.

Establishing customer trust hinges on AI systems' fair and transparent functioning (End-user, October 17, 2023). The significance of this statement is emphasized by Schwartz et al. (2022), who assert that AI's operation is not isolated and that public/customer trust asks for a consideration of all factors, including those beyond the technology itself. Using biased information by AI systems can result in inequitable outcomes for individuals predicated on extraneous personal attributes like genetics, zip code, or race (Rothstein & Joly, 2009; Signorello et al., 2014). These attributes are irrelevant to risk acceptance in the use case because they are not directly related to mortality or morbidity. In addition, using these attributes is prohibited by law (End-user, October 17, 2023). The undermining of system accuracy and the erosion of customer trust are concurrent consequences. A system that relies on biased or capricious criteria rather than impartial, transparent, and objective analysis undermines customer trust. Ensure no violation of social norms is similar to the stakeholder view in this context.

5.5. Conclusion

In this chapter, an empirical study was conducted within a Japanese life insurance organization focusing on the PU system. The empirical study was conducted to gather insights for the value framework. The values were translated into norms using system safety concepts: safety control structure and constraints. This translation involved a refinement process utilizing reporting standards and an AI governance framework, resulting in an initial value framework of 54 norms addressing the sub-question:

"Q3: Which norms emerge from the use case?"

The study highlights the lack of a standardized process in the scientific literature for converting identified values into practical organizational guidelines. Through empirical research, this chapter examines a combination of methods, including system safety concepts from Leveson (2012), reporting standards by Garst et al. (2022), and the AI governance framework by Mäntymäki et al. (2022). Combining these methods results in a set of norms that define each value's meaning and outline necessary protections for stakeholders' values in an organizational setting. This contributes scientifically by providing an initial method to translate high-level values into operational norms from a safety perspective. The chapter shows how design for values can be operationalized in an organizational context through empirical evidence. In addition, the empirical study shows how the different concepts of theory can be used in different contexts.

The practical outcome establishes a foundation for the industry to build upon, as it has only been tested in the context of the PU use case. Furthermore, developers of AI systems now have access to norms that outline the specific values they should adhere to and comprehend. The framework guides AI system developers and the life insurance industry regarding implementing the required controls. In the next chapters, these findings are the basis for establishing controls and design guidelines.

For the PU system, an empirical study was conducted with eight experts participating in questionnaires, workshops, and semi-structured interviews. These activities were guided by concepts such as the safety control structure and safety constraints. The experts defined 54 norms encompassing all known risks regarding the PU system that could endanger the identified values from sub-question 2.

These 54 norms needed to be translated into practical and applicable norms within the organizational setting. The steps outlined by Garst et al. (2022) for selecting relevant topics to report in organizational settings

were modified to apply to AI systems. The steps involved in this process are categorizing, identifying information sources, and selecting relevant topics.

The categorization process involved two steps. The initial step entails the division of process and assessment norms. The second approach entails categorizing the assessment norms into data and AI assessment norms. The process norms guide the development of guidelines in Chapter 7. The assessment norms serve as an input for determining design requirements in the next phase of this research. These steps are needed to structure the identified norms in an organizational setting to make it manageable.

The assessment norms were established based on the safety control structure, which also provided oversight of the relevant values identified by experts. The experts' choices were found to be associated with Mäntymäki et al.'s (2022) AI governance framework. The aid of this framework resulted in 18 data and 13 AI assessment norms. Likewise, these steps are needed to align the identified norms with the IT infrastructure in an organizational setting.

Determining the sources of information helps identify the stakeholder perspectives that should be considered when designing, developing, and deploying AI systems. The information sources can provide pertinent stakeholder criteria in designing and implementing AI systems, considering the perspectives of stakeholders whom the system may impact with a clear focus.

The next step is selecting topics that focus on delimiting the number of assessment norms and using only the relevant ones to make them manageable for the organizational context. The PU system adopted 11 data and 9 AI assessment norms to safeguard stakeholders' values. This step supports AI system developers in understanding what controls must be built to protect stakeholders' needs and values. However, controls must be implemented to ensure their protection.

6

Specification of design requirements

This chapter explains the systematic translation of norms into design requirements. This step addresses how risks can be controlled while assessing the predictive underwriting system. Therefore, this chapter answers the sub-question:

"Q4: Which design requirements can be derived from the norms?"

This sub-question helps AI system developers design the controls to protect stakeholders's values. It equips AI system developers with tools to safeguard the core principles of the key stakeholders within the particular AI system's framework. This sub-question uses an example of the value of fairness with the appending norms with the PU system as a baseline to illustrate the necessary steps.

6.1. Approach

The specification process, translating norms into design requirements, is, according to Richardson's (1994) conceptual framework, a transformative step that situates general norms within the practical realm of AI-driven systems. According to van de Poel (2013), the refinement involves three key aspects:

- Delineating the scope, which confines the norm's application to precise operational settings;
- Delineating objectives, which define the specific outcomes the system is expected to achieve;
- Delineating methods outline the particular actions or technological features the system must incorporate to realize the objectives.

Van de Poel (2013) states that the description of the action or objective includes the elements of location, time, purpose, method, agent, or recipient involved in its execution or pursuit. For instance, the overarching aim of enhancing operational safety in a predictive underwriting system can be further specified as the reduction of errors or data breaches. The adequacy of this refined specification hinges on its method, demonstrated by its ability to effectively mitigate risks and achieve the overarching objective of operational safety. This type of translation is needed to make the norms actionable in its context.

To illustrate this approach in our use case, we will show the three norms of fairness in the following sections. These norms exemplify the criteria applicable (for fairness) after completing the converging process discussed in Chapter 5.4.2. In addition, as mentioned in Chapter 4.3, this focus will rely on group fairness as that is one of the social norms identified. The fairness norms applicable to the PU system are:

- Develop AI systems that avoid excessive bias toward specific stakeholders in wealth and society.
- Incorporate only those biases into AI system outcomes based on transparent and verifiable objective criteria, subject to regular review and consensus among stakeholders.
- Adopt a product selection that is transparently designed and executed based on verifiable medical data and expert consensus, avoiding reliance on subjective judgment.

6.2. Avoid excessive bias

The norm, "*develop AI systems that avoid excessive bias towards specific stakeholders in wealth and society*," derived from Japan's Cabinet Secretariat's Social Principles for Human-Centered AI, can be

interpreted in two ways. In a broad sense, it can focus on customer selection; the second is the impact of the system outcomes. Because the PU system focuses on the pool of existing customers who are part of SMEs, we will only set design requirements to identify the risks within the outcomes of the system.

The norm itself can be divided into two parts. The first part is the definition of 'avoid excessive bias' in the context of the PU system, and the second part is an understanding of the meaning of 'specific stakeholders in wealth and society' within the PU system.

The determination to "*avoid excessive bias*" calls for contextual interpretation within the framework of the PU system to understand the influence of the PU system on wealth distribution and societal dynamics. The system's outputs have two functions: product allocation and risk acceptance. First, product allocation relates to the organization's product portfolio. Moreover, the system is specifically engineered to guide current clientele regarding the strategies of up-selling, cross-selling, or maintaining their current state. Decisions are based on examining medical and financial data, verified through a centralized insurance database, to ascertain the customer's financial capability for product modifications or expansions.

Furthermore, the sales personnel critically evaluate the model's recommendations, leveraging their expertise to assess the product's suitability. Following the customer's agreement to accept the proposal for up-selling or cross-selling, the application is promptly forwarded to the underwriting team. Bypassing the financial screening step is permissible for existing customers who have already undergone the process. The medical screening is excluded due to prior verification during the initial application. Second, the following phase involves the evaluation of risk acceptance for the model, utilizing the existing medical and financial data of the customer. Here, high risk is classified as 1, and standard risk is classified as 0. The classification as a standard risk has the benefit of omitting subsequent procedures. Identification as high-risk requires manual review by the underwriting team, thus extending the customer's process duration.

The stakeholder group is a criterion for fostering customer trust by avoiding decision-making characterized by unclearness, unfairness, and inaccuracy (End-user, October 17, 2023). Monitoring the predictive parity (PPV) statistic over time gives insight into the equal accuracy and reliability of the system across all customer groups (Loi & Christen, 2021). The PPV metric quantifies the anticipated precision among distinct sensitive groups, denoted as $\frac{TruePositive(TP)}{TruePositive(TP)+FalsePositive(FP)}$.

The potential impact on clients is punitive, whereby individuals identified as high-risk are subjected to an extended procedural duration. The utilization of the False Positive Rate (FPR) enables, in this case, the quantification of the proportion of individuals without risk who are erroneously classified as high risk, denoted as $\frac{FalsePositive(FP)}{FalsePositive(FP)+TrueNegative(TN)}$ (Quang, 2017). Although the impact of the system has no direct financial implications, the organization may ask for customer feedback to get an idea of whether customers perceive it as unfair to have to wait longer if not necessary. As mentioned in chapter 4, the legislative framework in Japan shows a policy-oriented approach strategically designed to foster economic growth. The PU system's influence is negligible in this context, as it pertains to a pre-established customer base and is accompanied by numerous human oversight mechanisms.

Narayanan (2018) states that it is impossible to satisfy all fairness definitions. The same applies to FPR and PPV metrics within the predictive underwriting system, where achieving perfect results remains unattainable due to their inherent interdependence. Developers and domain experts should monitor the metric's influence on system functionality and stakeholder outcomes over time. The diligent monitoring guarantees the potential for optimization of the balance between these measures. In addition, the measures must be compared to the ground truth of the manual process to understand the system's performance. The ground truth in the PU case is measured through mortality and morbidity.

Second, specific stakeholders in wealth and society. As mentioned in Chapter 3, The FSA said in their guidelines that there might not be discrimination against people above +65 and people with disabilities (Financial Service Agency, 2021, p. 304). However, we need to understand on a high level how the healthcare system in Japan works concerning life insurance. Japan's Long-Term Care Insurance System supports lower socioeconomic individuals and those aged 65 and above or between 40 and 64 with specific age-related conditions (Tamiya et al., 2011; End-user, October 17, 2023). The system is designed to take users' financial capabilities into account. Individuals with lower incomes may be eligible for amplified subsidies and diminished co-payments for the healthcare services they avail of.

The background information yields three primary findings: the identified metrics for standard measurement, the specified groups for measurement, and the designated individuals for standard monitoring. Three design requirements emerged from these findings.

DR1:*Developers must measure the middle and high socioeconomic classes, the age group between 40 and 65, and individuals without disabilities using the predictive parity rate.*

DR2: *Developers must measure the middle and high socioeconomic classes, the age group between 40 and 65, and individuals without disabilities using the false positive rate.*

DR3: *Developers must monitor both metrics over time and establish relevant thresholds with experts' and stakeholders' feedback.*

6.3. Personal attributes

De norm '*Incorporate only those biases into AI system outcomes based on transparent and verifiable objective ground, subject to regular stakeholder review and consensus.*' stems from the FSA and internal organizational policy (End-user, October 17, 2023). Foremost, we split up the norm into three parts:

- What is seen as bias based on transparent and verifiable objective grounds?
- Who must regularly review the model?
- Which stakeholders must agree on the biases in the system?

The first is the statement is *"Incorporate only those biases into AI system outcomes based on the transparent and verifiable objective ground."* Sensitive information, including race, creed, social status, medical history, criminal record, and other personal identifiers, is addressed in the APPI's definition of social biases (Cabinet Secretariat, 2017, p. 2). These are considered sensitive because they may give rise to unfair prejudice or discrimination. Notably, given its direct association with important insurance measures like mortality and morbidity, the PU method employs certain sensitive data, such as age. This use is supported objectively and verified by expert validation and statistical analysis. The FSA has forbidden filtering based on past genetics and ethnicity because these factors have no bearing on insurance metrics (End-user, October 17, 2023). In line with the consensus of medical specialists, local policies prohibit filtering based on zip code (End-user, October 17, 2023).

Ethnicity, genetics, and zip code are the three main attributes inappropriate for filtering. The fact that sensitive attributes cannot be completely addressed by deleting or ignoring them is a major challenge in high-dimensional algorithmic models. If non-protected traits are connected, protected information can be deduced from them. Usually happening during the fitting process of complicated models, this behavior results in proxy or indirect discrimination (Lindholm et al., 2023).

To identify and mitigate such proxies, we examine two phases: pre-processing and post-processing. In the pre-processing phase, which involves data and feature selection, it is vital to note that this phase focuses simply on the input data (X), applying protected attributes (A) without considering the target variable's response (Y). Various statistical methods are available for checking biases at this step. For instance, we can examine entropy to evaluate the uncertainty or variability in (X) and (A) (Deldjoo et al., 2019; Kallus et al., 2022). A high entropy suggests a significant degree of variance in the data, implying an absence of evident patterns or biases. Comparing entropy with and without considering protected attributes might reveal the level of bias. If entropy dramatically changes while accounting for protected attributes, it may suggest the existence of bias.

During the post-processing phase, our primary objective is to discern any biases in the outcomes. The variables of zip code, ethnicity, and genetic information are assessed separately from the model's outcomes to identify potential biases. This evaluation utilizes metrics such as PPV and entropy to analyze the relationship between $Y = F(X), A$. One can utilize bias detection tools like IBM AIF360 or internal organizational software designed specifically for this task to detect bias.

Both stages have the objective of quantifying biases using objective metrics. Nevertheless, consistently assessing and apprehending the circumstances that give rise to these metric outcomes is paramount. For example, suppose the PPV is 0.9 for one ethnic group and 0.5 for another. In that case, an examination over time is needed to understand how these metrics change with varying data inputs.

In the present PU system, the metrics' maintenance and measurement involve actuaries calculating models and their validation by medical professionals (End-user, October 17, 2023). Therefore, the PU system developed by developers should be reviewed by actuaries against the statistics mentioned earlier and accepted by underwriters and medical professionals. This periodicity should be observed during each iteration of data updates, which commonly takes place every six months.

DR4: *Utilize statistical methods such as entropy to identify indirect relationships with unfounded sensitive variables in the pre-processing phase.*

DR5: *Employ statistical methods like entropy and PPV in the post-processing phase to identify indirect relationships with unfounded sensitive variables.*

DR6: *Evaluate the outcomes of fairness metrics with actuaries to understand the results' context.*

DR7: *Validate the results with medical professionals to ensure objective continuity in the system's operation.*

6.4. Product selection

The norm, '*Adopt a product selection that is transparently designed and executed based on verifiable medical data and expert consensus,*' is an outlined internal business rule developed in partnership with the FSA to ensure fairness (End-user, October 17, 2023). This norm can be analyzed from three separate viewpoints.

First, the part about '*adopt a product selection that is transparently designed.*' Transparency in this particular situation refers to the clear and open communication of the reasoning behind the product's functionality, its intended use, and any inherent limitations or potential risks. Integrating model-agnostic, additive feature attribution methods is required within the PU system to improve comprehension of the system's output. Currently, the system utilizes elastic net regression, complicating the understanding of decision-making processes. Researchers can employ statistical techniques like Shapley Additive explanations (SHAP) or LIME (Local Interpretable Model-agnostic Explanations) to conform to this standard and clarify the results of the model (Maier et al., 2020). The explainability tools accurately measure the impact of different input features, such as medical and financial data, on the model's predictions. The approach used to measure the contributions of features meets the need for transparent decision-making by replacing subjective evaluations with insights derived from data. Furthermore, an explanation tool encompasses local and global explainability, thereby providing profound insights into the decision-making process driven by the model.

The last component of the norm requires that the choice of products be based on reliable data and agreement among medical and financial professionals. In pragmatic terms, a panel of experts must approve the system's decisions. These experts evaluate the recommendations made by the model, using their professional expertise to assess the appropriateness of products for customers. This approach guarantees that the model and its data do not solely influence decision-making but also incorporate valuable human insight and expertise. This approach enhances the integrity and robustness of selecting products by ensuring a fair and informed decision-making process.

This norm emphasizes the need for clear and open design, supported by quantitative explainability techniques, and emphasizes the significance of expert agreement in validating and directing the decision-making procedure. This approach guarantees that selecting products is based on reliable technical analysis and data and incorporates strong ethical considerations and expert perspectives. As a result, it promotes a responsible and reliable implementation of artificial intelligence in the life insurance industry.

D8: *Implementing explainability models for decision support, enhancing transparency and understanding of model outputs.*

D9: *Development of an interface to display the contributions of various data elements to model predictions visually and understandably for end-users.*

D10: *Establish an expert panel to review and assess model recommendations and outputs regularly, ensuring product selection aligns with expert consensus and real-world applicability.*

D11: *Implementing a feedback system for continuous model improvement, enabling experts to provide input that shapes the system's evolution.*

6.5. Conflicts, relations, and trade-offs

Conflicts may arise among values if their interaction is not considered and suitable solutions are not developed. Therefore, trade-offs must be made, and design requirements must be refined to achieve balance (van de Poel, 2013). For instance, in the PU system, prioritizing fairness could compromise the system's robustness and vice versa. Striking a balance between these values is needed to consider all stakeholders' views. It must be considered when translating them into specific design requirements to make them controllable in an organizational setting. The design requirements formulated in this chapter are centered primarily on fairness. Conflicts may arise if there is an imbalance in the implementation of design requirements.

Aizenberg and van den Hoven (2020) suggest the importance of understanding the harm-benefit trade-offs in AI systems, which can uncover stakeholder considerations that might not be immediately apparent. One such approach for the PU system could involve implementing a nuanced risk assessment framework that recognizes the variety of employees insured within SMEs, which entails designing a system capable of adapting and learning from diverse data sets to minimize biases. Concurrently, it is essential to integrate tools for ongoing bias detection and mitigation while maintaining overall performance and reliability. Nonetheless, this process involves inevitable trade-offs, continuous refinement, and expert evaluation to balance fairness and robustness.

Table 6.1 represents the conflicts and relations between design requirements outlined in this chapter. The table displays conflicts in red and relationships in blue. The design requirements' relationships suggest that they can be established as controls to safeguard the stakeholders' value and ensure fairness. The conflicts highlight the need to find a balance that upholds the importance of fairness. The conflicts that have been identified are:

- **DR1 & DR2:** Narayanan (2018) states that it is impossible to satisfy all fairness definitions. While both focus on similar demographics, using different rates (predictive parity rate vs. false positive rate) for assessment could lead to conflicts in prioritizing which metric to emphasize.
- **DR4 & DR5:** Using the same statistical methods in different phases (pre-processing vs. post-processing) might lead to redundancy or conflicts in the interpretation of results (Wan et al., 2021).
- **DR3 & DR6-DR8:** DR3 emphasizes ongoing monitoring, while DR6-DR8 focuses on evaluating outcomes and enhancing transparency and understanding through the input of different experts. If the methods for monitoring and evaluation are not aligned, this could create conflicts in how data is interpreted and used.

Furthermore, interconnections exist between the various design requirements:

- **DR3 & DR1-DR2:** DR3 complements DR1 and DR2 by emphasizing monitoring these metrics over time, which ensures an ongoing evaluation of the predictive parity rate and false positive rate by experts and stakeholder feedback (Castelnovo et al., 2022; López-Paz, 2022). In addition, this relationship could also support the conflict between DR1 and DR2, as the continuous evaluation of experts could find the balance between the two metrics.
- **DR3 & DR4-DR5:** This relation has the same argumentation as the previous relation between DR1-DR2. The relationship supports the conflict between DR4 and DR5, as the continuous evaluation of experts could find the balance between the two statistical results.
- **DR6 & DR7 & DR8:** These requirements are interconnected in ensuring the fairness and understanding of the model. The requirements focus on evaluating outcomes (with actuaries), validating results (with medical professionals), and implementing explainability models to strengthen the evaluation and validation (Chen et al., 2023).
- **DR9 & DR1-DR2, DR4-DR5:** DR9 is focused on a user interface, while DR1-DR2 and DR4-DR5 focus on different calculation techniques. These calculations could all be gathered in the same user environment to oversee the model's performance.
- **DR9 & DR 11:** Both focus on interactions with stakeholders and experts. DR9 deals with developing a user interface, while DR11 emphasizes a feedback system for continuous improvement. Combining the two design requirements would optimize the communication channels and systems' effectiveness (Følstad, 2017).

	DR1.	DR.2	DR.3	DR.4	DR.5	DR.6	DR.7	DR.8	DR.9	DR.10	DR.11
D1.											
D2.											
D3.											
D4.											
D5.											
D6.											
D7.											
D8.											
D9.											
D10.											
D11.											

Table 6.1: Conflict-relation diagram of design requirements. The red boxes represent conflicts, while the blue boxes represent a reinforcing relationship between the requirements.

In summary, the list of conflicts acknowledges that these possible conflicts are not inherent deficiencies but deliberate compromises and trade-offs. Academic literature often mentions the challenge of reconciling fairness and accuracy in AI ethics (Narayanan, 2018). Achieving this balance must consider its context and the organization's values, regulatory obligations, and impact on the stakeholders. Therefore, founded relations can support resolving conflicts and find effective controls for AI system developers to protect stakeholders' values.

6.6. Conclusion

This chapter's conclusion addresses transforming norms into design requirements in the practical environment, namely the PU system. Therefore, this chapter provides an example of the value of fairness. Three norms were examined in the predictive underwriting practices with experts. These norms' examples demonstrate how to address potential relationships and trade-offs among the identified design requirements. Following these steps provides an answer to the subsequent sub-question:

"Q4: Which design requirements can be derived from the norms?"

Answering this sub-question outlines the steps and resources that AI system developers should utilize to convert the norms identified in sub-question 3 into technical requirements for the AI system. Therefore, this chapter highlights the need for expert input to understand the contextual application and validation of design requirements while defining the scope, methodology, and objectives. In addition, the contextual application must be emphasized in the organizational, societal, and institutional aspects to make it controllable to comply with social norms. The academic literature provides techniques for the expert domain of ethical AI and how to make norms measurable and actionable. The scientific contribution of this chapter is to provide steps to translate the identified applicable assessment norms from Chapter 5 into actionable controls. In addition, this chapter provides valuable steps for the guidelines presented in the next chapter.

The practical contribution is approaching the design requirements phase for AI system developers. It provides an overview of how controls can be established for an initial AI system. In addition, this chapter provides the first controls for the PU system for the value of fairness that AI system developers can use.

For the PU system, this can be reflected in including data scientists and underwriting specialists in the conversations about designing measurable and actionable requirements. With the aid of academic literature, this process resulted in 11 design requirements. Afterward, three conflicts were identified within these 11 design requirements, which must be analyzed for trade-offs. In addition, five relationships were found that could support the process of making trade-offs between the conflicts.

7

The design guide: from theory to practice

The primary objective of this master's thesis is to offer actionable guidelines to developers involved in the control cycle, enabling them to protect stakeholders' values within the design, development, and implementation of AI systems. The guidelines aim to be an artifact that bridges the gap between theory and practice, enabling the transfer of theoretical knowledge and research insights into practical applications. Therefore, the last sub-question will be answered in this chapter:

"Q5: What design guidelines can guide the development of AI systems?"

This guide aims to set up a process for the identification and control of risks throughout the entire lifecycle of an AI system, starting from its design phase to monitoring. The initial intent of this work was to offer direction to those involved, or who need to be involved, in creating AI systems that consider both technical and societal aspects. This chapter introduces a step-by-step guide. The design cycle for developing this guide ensures the collection of all the results of the previous chapters to construct and operationalize the created value framework.

7.1. A guide as an artifact

AI systems rooted in engineering and computer science traditions use mathematical abstractions to grasp technical problems and their solutions (Dobbe et al., 2021). However, a more holistic approach is required due to the socio-technical complexities and different stakeholder needs of AI systems in practical situations. The guide incorporates a socio-technical system view, acknowledging that AI systems, especially when deployed in real-life contexts, involve complex socio-technical interactions that cannot be fully understood through a technical lens alone. This all-encompassing viewpoint uses the interaction between institutions, stakeholders, and technology. In this setting, a guide appears as an artifact and a tool intended to help AI system developers use an institutional framework and process to prevent risk, protect stakeholders' values, and communicate concretely. These guidelines summarize this research's steps but are transformed so that AI systems developers can follow a manageable process during the AI lifecycle. This process uses assessment norms and is linked to operational norms. Examples from the use case show the guidelines in action and their practical application to give them life. Therefore, this artifact proposes a solution for the problem in the practical field of the life insurance industry in Japan. The problem stated the challenge in determining which social norms should be protected and implemented to control the technical operations of AI systems in organizational settings.

7.2. Guidelines connected to values and norms

Ten guidelines comprise the framework, which supports stakeholders in AI system design, development, and deployment scenarios by emphasizing socio-technical norms. These guidelines identify risks within the AI life cycle and can be tailored to the particulars of AI design, development, and deployment scenarios.

The guidelines provide a concise overview of the research methodology. Table 7.1 provides an overview of the sources from which each design guideline is derived, serving as a means to validate where they came from. The table illustrates the research phases from which information is collected and provides explanations. The explanations consist of arguments supporting the conclusions of each sub-question.

The guidelines structure the identified norms of chapter 5 into practical processes. Table 7.2 provides an

Guideline	Research phase	Explanation of guideline creation
Design guideline 1: Create a multi-disciplinary team to set up the project and go through the control cycle	*Exploring the environment*	This research phase highlights the need for business stakeholders to who could explain and guide through the context, environment, goals, and restrictions of the AI system.
Design guideline 2: Define the scope and business purpose	*Exploring the environment*	Incorporation of law and stakeholder input is needed to set goals and restrictions. The goals and restrictions can support defining the scope of the initiated AI system. In addition, this phase focuses on defining which stakeholder groups are relevant and are included in the scope.
Design guideline 3: Identify the metric(s) for the effectiveness of the AI system.	*Exploring the environment & values*	From the business values and expert interviews, it became clear that an AI system must be ethically responsible and effective. Effectiveness could differ in every use case, depending on the system goal. Therefore, end-users must be included in measuring effectiveness in their work practices.
Design guideline 4: Identify stakeholder criteria and assessments of the development	*Values & norms*	This guideline is supported by the steps depicted in Garst et al. (2022). Information sources must be chosen for each stakeholder group; therefore, stakeholder criteria can be set. The next design guideline explains the stakeholder criteria for evaluating the AI system's impact on individuals.
Design guideline 5: Assess the design of the system	*Norms*	The thesis provides a value framework that combines the concepts of Leveson (2012), Garst et al. (2022), and Mäntymäki et al. (2022) to translate the identified values into norms that fit the organizational context. Participants provided a list of risks with aid from the safety control structure and safety constraints to formulate familiar risks. AI system developers can use this value framework by selecting the data assessment norms that apply to the identified stakeholder criteria from the previous guideline. This guideline identifies required controls to mitigate risks to stakeholders' values relevant to the AI system. This step is derived from the process outlined by Garst et al. (2022), referred to as the 'selecting topics' step.
Design guideline 6: Review assessment with experts and setting requirements	*Design requirements*	As highlighted in the design requirement phase, expert input is needed to understand the contextual application and validation of design requirements while defining the scope, methodology, and objectives. In addition, relationships between requirements must be determined, and trade-offs must be made.
Design guideline 7: Assess the development of the system	*Norms*	This step is the same as design guideline 5, but uses the AI assessment norms. The decision to test the AI assessment norms at this stage was based on empirical research and the knowledge base. Mäntymäki et al. (2022) indicate differentiating data and AI to ensure alignment with the organizational IT infrastructure. During the workshop, experts distinguished between data-related risks during the design phase and AI-related issues during the development phase.
Design guideline 8: Review assessment with experts and setting control mechanisms	*Design requirements*	This guideline is the same as design guideline 6.
Design guideline 9: Monitor risks through feedback channels	*Design requirements*	The research shows that most controls are open-ended in the context of the Japanese insurance industry. One of the reasons is the social norm of 'wholeness,' where reductionism is not preferred. Therefore, monitoring through feedback channels with end-users and experts is needed to review the results.
Design guideline 10: Create continuous improvement within the AI system and its governance	*Designing of design guidelines*	This step reflects on this design guide and AI systems. The continuous improvement stems from the legal and industry values to direct their risk-based approach. This guideline is integrated to comply with this value and improve systems and this guide in small steps.

Table 7.1: Design guidelines and the backward traceability of used steps during research. The first column shows the guidelines. The second column shows the phase of the research's approach where the guideline stems. The last column explains the establishment of the guideline.

overview of the guidelines connected to the operational norms and values. Appendix O shows an overview of the process norms.

Guideline	Norms	Values
Design guideline 1: Create a multi-disciplinary team to set up the project and go through the control cycle	6, 11, 18 & 20	Transparency, usability, accountability, and effectiveness
Design guideline 2: Define the scope and business purpose	1, 3, 4, 9, 14 & 16	Trust, transparency, robustness, usability and privacy
Design guideline 3: Identify the metric for the effectiveness of the AI system.	19	Effectiveness
Design guideline 4: Identify stakeholder criteria and assessment of the development	1	Trust
Design guideline 5: Assess the design of the system	5, 13 & 17	Usability, transparency and security
Design guideline 6: Review assessment with experts and setting requirements	11, 12 & 20	Usability and effectiveness
Design guideline 7: Assess the development of the system	5, 8, 10, 11, 13 & 20	Transparency, robustness, usability and effectiveness
Design guideline 8: Review assessment with experts and setting control mechanisms	11, 12, 13, 15, & 20	Usability and effectiveness
Design guideline 9: Monitor risks through feedback channels	7, 11, 13, 15, 20 t/m 23	Robustness, usability, effectiveness and continuous improvement
Design guideline 10: Create continuous improvement within the AI system and its governance	23	Continuous improvement

Table 7.2: Design guidelines connected to the corresponding process norms and values.

7.3. Guide instructions

This section explains the application of the guidelines through an illustrative example. The guidelines must be read chronologically, as these are reflected in the steps taken in research. The example focuses on the first design guideline.

Design guideline one aims to assemble a multi-disciplinary team to go through the project and control cycle. Here, the objective is stated for the developer to contact the right people to go through the project initiative and the different phases of the control cycle together as a team. The participants that will be involved in this step are also indicated.

The process norms indicated here serve as boundary conditions. The first norm focuses on value accountability and states, ”Form a multi-disciplinary team with business stakeholders and developers to ensure effective guidance through the AI lifecycle.” This norm focuses on the objective of this step, where the multi-disciplinary team will jointly walk through the AI lifecycle.

The second norm, focused on the value of usability, states, ”The team will cultivate a cooperative dynamic, uniting varied expertise across the organization to ensure that risk and information management are handled cohesively.” This norm states that the team will focus on risk management of information and AI systems.

The third norm represents transparency and states, ”The team is tasked with instituting transparent accountability and oversight of the AI system’s lifecycle within the enterprise.” This norm focuses on establishing clear responsibilities throughout the AI cycle.

The last norms represents effectiveness: ”The team will facilitate a synergetic development environment, ensuring the AI system’s accuracy, effectiveness, and alignment with the needs of the business.” This norm indicates a clear goal of the AI system.

Finally, an example was given from the PU system use case to make it practical for users. The following subsections present the guidelines, explaining the actions, the responsible parties, and the potential implementation methods. Furthermore, the interconnectivity of the norms is explained, and their relationship is expounded upon. Finally, illustrative examples are provided to enhance the practicality of the guidelines.

7.4. Design guideline 1: Create a multi-disciplinary team to set up the project and go through the control cycle

Objective: Initiate contact to establish a multi-disciplinary team for the AI project. A multi-disciplinary team shall establish a project initiation and oversee the control cycle. This team will delineate the project's scope, articulate its intended impact, scrutinize stakeholder requirements, institute channels for feedback, and manage risk assessment.

Participants: The team will encompass end-users, business owners, and AI system developers. Each brings valuable insights from their interactions with the existing manual processes.

Process:

- Contact business owner
- Engage in collaboration with the business owner and identify other relevant stakeholders who should also participate.

Norms integration:

- **Accountability**: Form a multi-disciplinary team with business stakeholders and developers to ensure effective guidance through the AI lifecycle.
- **Usability**: The team will cultivate a cooperative dynamic, uniting varied expertise across the organization to ensure that risk and information management are handled cohesively.
- **Transparency**: The team is tasked with instituting transparent accountability and oversight of the AI system's lifecycle within the enterprise.
- **Effectiveness**: The team will facilitate a synergetic development environment, ensuring the AI system's accuracy, effectiveness, and alignment with the business needs.

Application example: In the PU system context, the team will consist of AI developers, sales and underwriting business leaders, and end-users. This composition ensures an understanding of the manual processes and the identification of challenges experienced by stakeholders, thereby fostering an environment conducive to effective AI deployment.

7.5. Design guideline 2: Define the scope and business purpose

Objective: Establish the AI system's scope and business purpose during the design phase, incorporating stakeholder needs, regulatory compliance, and operational drivers.

Participants: The multidisciplinary team.

Process:

- Identify and analyze the challenges faced by business owners and end-users.
- Map the existing manual process to understand stakeholder interactions and workflows.
- Research applicable legal and regulatory requirements influencing the manual process.
- Determine the primary motivations driving the AI system's development, focusing on accuracy and efficiency.
- Evaluate the potential impact on each stakeholder group to ensure inclusion.
- Allocate clear responsibilities and establish accountability mechanisms for all stakeholders involved.

Norms integration:

- **Trust:** Utilize AI systems to meet customer needs, elevate service quality, and provide clear and sufficient explanations about AI-driven decisions, emphasizing the system's role in enhancing customer understanding and satisfaction.
- **Trust:** Design AI systems to support human autonomy and augment employee roles.
- **Transparency:** Maintain transparency about the AI's objectives, ensuring its operations are benevolent and clear to all stakeholders.

- **Robustness:** Integrate human oversight within the AI system to guide and verify model outcomes, ensuring AI complements rather than overrides human decisions.
- **Usability:** AI systems should augment human capabilities, enhancing productivity and decision-making without supplanting human autonomy.
- **Privacy:** AI should assist, not replace, human judgment, ensuring responsible use in decision-making processes.

Application example: The project begins by identifying the need for a system that enhances the accuracy and operational efficiency of the PU system. The team will then detail the current processes, dependencies, applicable data, performance metrics, and regulatory considerations. This understanding will guide the development of an AI system that meets technical specifications and integrates seamlessly with stakeholder requirements, maintaining ethical integrity and operational excellence.

7.6. Design guideline 3: Identify the metric(s) for the effectiveness of the AI system

Objective: To measure the AI system's effectiveness in alignment with organizational goals and user satisfaction.

Participants: The multidisciplinary team.

Process:

- Synthesize insights from the system's defined scope to articulate its intended effectiveness.
- Develop metrics that reflect the AI system's performance in enhancing decision-making, operational efficiency, and end-user and customer satisfaction.
- Implement feedback mechanisms for end-users to capture real-time efficacy data.
- Amend the scope from Design Guideline 2 if new effectiveness criteria emerge.

Norm integration:

- **Effectiveness**: The AI system shall be designed with a clear purpose, supporting and enhancing human decision-making, improving employee efficiency, and ensuring alignment with business objectives.

Application example: For the PU system, effectiveness is gauged by the precision of underwriting decisions and client satisfaction levels. Specific performance metrics and customer feedback loops can be established to measure this, enabling ongoing assessment and refinement post-implementation.

7.7. Design guideline 4: Identify stakeholder criteria and assess the design

Objective: To develop criteria based on stakeholders' needs, ensuring the AI system's design aligns with diverse expectations and requirements.

Participants: The multi-disciplinary team and relevant stakeholders.

Process:

- Categorize stakeholder groups such as businesses, agents/clients, and society to tailor the criteria effectively.
- With the aid of a multi-disciplinary team, identify and codify stakeholder-specific criteria that the AI system must meet.
- Gather criteria from various sources, including business owners, end-users, and documented industry and government standards.

Norm integration:

- **Trust**: Utilize AI systems to meet customer needs, elevate service quality, and provide clear and sufficient explanations about AI-driven decisions, emphasizing the system's role in enhancing customer understanding and satisfaction.

Application example: For the PU system, establish criteria that resonate with:

- Business: Ensure a positive impact on the accuracy and reliability of the results.
- Agents/Clients: Ensure a positive impact on customer trust by avoiding unclear, unfair, and inaccurate decisions.
- Society: Ensure no violations of social norms such as wholeness, contribution to the community, sensitivity, and sincerity.

The criteria presented in this use case are derived from discussions with end-users with strong affiliations with the agents, conversations with management, and a review of relevant academic literature. The criteria can be collected in any format that aligns with the stakeholder's perspective.

7.8. Design guideline 5: Assess the design of the system

Objective: To evaluate the AI system's adherence to privacy, security, effectiveness, and accountability as informed by societal, industrial, and managerial inputs. **Participants:** the multi-disciplinary team **Process**:

- Use the assessment framework based on multi-disciplinary inputs, focusing on privacy, security, effectiveness, and accountability.
- Apply a Multi-Criteria Decision Analysis (MCDA) approach to correlate stakeholder criteria with relevant norms.
- AI developers are responsible for detailing how each applicable norm is addressed and controlled within the system.

Norm integration:

- **Transparency**: Maintain a commitment to transparency and integrity throughout the AI lifecycle on the data and system level, with documentation of the AI system's performance and security metrics.
- **Usability**: Conduct regular assessments of the AI system to gauge its usability and impact, using feedback to drive continuous improvement.
- **Security:** Ensure changes are executed, adequately tested, and promoted to production in a controlled and timely manner to prevent service disruption, security breaches, etc.

Application example: In assessing the PU system, the assessment leverages criteria from Design Guideline 4 to determine applicable norms. Each norm that influences the criteria is included in the assessment, with a detailed explanation of how the norm is covered and what control mechanisms are employed. Specific attention is given to privacy and security due to the system's use of customer data. An example of an applicable norm from the value of privacy is *"Activities related to customer and organization data should be preempted with impact assessments to mitigate risks and safeguard customer interests."*, where we could fill in a Data Impact Assessment provided by the organization to identify the risks.

7.9. Design guideline 6: Review assessment with experts and setting requirements

Objective: To facilitate an expert-led review of the AI system's assessment, identifying and documenting potential risks and devising appropriate control mechanisms.

Participants: This process will involve independent experts, including AI effectiveness reviewers, information risk managers, IT security officers, and legal and compliance advisors, each contributing specialized knowledge to the system's design evaluation.

Process:

- Organize a structured roundtable discussion with experts to examine the AI system assessment, focusing on detailed explanations and risk identification. Submit the completed assessment form ahead of time.
- Document insights and potential risks, ensuring they are communicated to the multi-disciplinary team for further action.

Norms integration:

- **Usability:** Guarantee transparency in AI operations, ensuring end-users understand the system's objectives and its implications for their roles.
- **Usability:** Development and implementation of AI require interdisciplinary expertise, integrating diverse business and technical insights.
- **Effectiveness:** Promote a collaborative environment that involves stakeholders across disciplines in the development process, ensuring that the AI system remains accurate, up-to-date, and relevant to evolving business and user needs.

Application example: In the predictive underwriting system context, a panel includes experts from information risk management, operational risk management, and financial risk management to scrutinize the preliminary risk assessment. The discussion aims to unearth any overlooked risks, with subsequent findings thoroughly recorded and relayed to the broader project team for integration into the development lifecycle. One example from the PU system pertains to information risk management, which emphasizes assessing potential consequences and ensuring developers have compliant access to sensitive data (IRM, October 17, 2023).

7.10. Design guideline 7: Assess the development of the system

Objective: To establish an evaluation of the AI system focused on robustness, transparency, and fairness.

Participants: AI system developers and the multi-disciplinary team are tasked with the assessment.

Process:

- Utilize a value framework akin to a Multi-Criteria Decision Analysis (MCDA) to evaluate the system against stakeholder-defined criteria, focusing on norms directly influencing the stakeholder criteria from design guideline 4.
- For the relevant norms, set design requirements.
- Define trade-offs between the design requirements and the stakeholder criteria.
- Implement control mechanisms for each relevant norm, documenting the rationale and expected outcomes.
- Integrate feedback from AI experts and the multi-disciplinary team to set benchmarks and thresholds for performance metrics like Predictive Parity over time for relevant social classes.

Norms integration:

- **Transparency**: Maintain a commitment to transparency and integrity throughout the AI lifecycle on the data and system level, with documentation of the AI system's performance and security metrics.
- **Robustness**: Develop and implement a risk mitigation strategy, incorporating multiple expert analyses to validate the AI system's reliability.
- **Robustness**: Data scientists must validate AI development methodologies thoroughly before system deployment to ensure methodological soundness.
- **Usability**: Foster collaborative processes across diverse teams for the design, development, and operational management of AI systems, promoting an integrated approach to risk and information management.
- **Usability**: Conduct regular assessments of the AI system to gauge its usability and impact, using feedback to drive continuous improvement.
- **Effectiveness**: Promote a collaborative environment that involves stakeholders across disciplines in the development process, ensuring that the AI system remains accurate, up-to-date, and relevant to evolving business and user needs.

Application example: In the context of the PU system, the development assessment would check for fairness, notably the norm: "Develop AI systems that avoid excessive bias towards specific stakeholders in wealth and society." Control mechanisms such as ongoing monitoring of predictive parity metrics for different social classes will be established, with thresholds informed by stakeholder feedback, ensuring the system's alignment with societal expectations. Nevertheless, conflicts may arise in the design requirements when attempting to balance the objectives of ensuring complete fairness and prioritizing the business perspective, specifically accuracy and efficiency. In the context of the PU system, trade-offs must be made to achieve a balance between accuracy and fairness.

7.11. Design guideline 8: Review assessment with experts and setting control mechanisms

Objective: To facilitate expert evaluation of the AI system's assessment and to establish relevant control mechanisms.

Participants: This involves independent experts across various fields pertinent to the system's design, including AI's robustness, transparency, fairness, security, and legal compliance.

Process:

- Organize review sessions with all pertinent experts simultaneously.
- Iterate the assessment process based on the insights and risk identification provided by these experts.
- Document discussions, expert feedback, and finalized control mechanisms for transparency and traceability.

Norms integration:

- **Usability:** Foster collaborative processes across diverse teams for designing, developing, and operational managing AI systems, promoting an integrated risk and information management approach.
- **Usability:** Guarantee transparency in AI operations, ensuring end-users understand the system's objectives and its implications for their roles.
- **Usability:** Conduct regular assessments of the AI system to gauge its usability and impact, using feedback to drive continuous improvement.
- **Usability:** Development and implementation of AI require interdisciplinary expertise, integrating diverse business and technical insights.
- **Effectiveness:** Promote a collaborative environment that involves stakeholders across disciplines in the development process, ensuring that the AI system remains accurate, up-to-date, and relevant to evolving business and user needs.

Application example: In reviewing the PU system, a round table with domain-specific experts will be convened to discuss and refine the system's fairness, transparency, and robustness. To uphold principles of fairness and transparency, involving an expert in artificial intelligence and an underwriter in the review process is compulsory (End-user, October 17, 2023). The model's robustness can be influenced by information risk management, which considers data quality, and financial risk management, which evaluates the methodology used in the model (IRM, October 17, 2023; FERM, October 17, 2023). We assess these factors as they can impact different business operations, including the ability to meet financial obligations (solvability) (FERM, October 17, 2023).

7.12. Design guideline 9: Monitor risks through communication channels

Objective: To establish an open-ended control system that adapts to the evolving needs and outcomes within the AI lifecycle, allowing for integrating new functionalities and processes (Standish, 2003).

Participants: AI developers and end-users monitor and adapt control mechanisms to ensure system integrity and relevance.

Process:

- Deploy adaptable control mechanisms not confined to predefined pathways or outcomes but can evolve with the system.
- Utilize dashboards that integrate feedback channels and monitor established controls, adjusting them during system updates or continuous monitoring scenarios.

Norms integration:

- **Robustness:** Diligently track and enhance the AI system's performance, promptly rectifying errors or inconsistencies.

- **Usability:** Foster collaborative processes across diverse teams for designing, developing, and operational managing AI systems, promoting an integrated risk and information management approach.
- **Usability**: Regularly assess the AI system to gauge its usability and impact, using feedback to drive continuous improvement.
- **Usability**: Development and implementation of AI require interdisciplinary expertise, integrating diverse business and technical insights.
- **Effectiveness:** Promote a collaborative environment that involves stakeholders across disciplines in the development process, ensuring that the AI system remains accurate, up-to-date, and relevant to evolving business and user needs.
- **Effectiveness:**Implement a systematic approach to assess the AI system's performance and impact, optimizing its efficiency and documenting its business implications to guide strategic decisions and process improvements.
- **Continuous improvement:** Implement continuous monitoring protocols to evaluate and optimize the AI system's impact on stakeholder trust, service quality, and ethical operations, including privacy, security, and fairness.
- **Continuous improvement:** Foster a collaborative environment for risk assessment, leveraging diverse perspectives to understand the full context of the AI system's deployment and operation.

Application example: As defined in Chapter 6, a multitude of design requirements that possess the potential to be amalgamated into cohesive control mechanisms have been identified. For example, we consider the norm of "Creating artificial intelligence systems that mitigate the occurrence of disproportionate favoritism towards particular interest groups within the realms of affluence and societal structures." Design requirements one and two prioritize monitoring two important metrics: the positive predictive value (PPV) and the false positive rate (FPR). By incorporating the third design requirement of requesting input from experts and stakeholders to establish suitable thresholds, we have developed a control mechanism to identify risks associated with the norm.

7.13. Design guideline 10: Create continuous improvement within the AI system and its governance

Objective: To incorporate a philosophy of continuous improvement into the AI lifecycle, enabling the system to adapt to evolving risks and stakeholder values.

Participants: This process involves AI developers, domain experts, and management teams responsible for the AI system's evolution.

Process:

- Yearly review and adjust the value framework to align with current laws, emerging AI technologies, and evolving stakeholder feedback.
- Implement methodologies that foster continuous improvement to allow quick adaptations to changes in the socio-technical environment for AI systems.

Norm Integration:

- **Continuous Improvement**: Foster a collaborative environment for risk assessment, leveraging diverse perspectives to understand the full context of the AI system's deployment and operation.

Application Example: Continuous improvement mechanisms for the PU system will involve regular reviews of the effectiveness of the deployed AI in light of changing legal standards and technological advancements. This iterative process ensures that the AI system adheres to current standards and is primed for future adaptations, as stakeholders and experts recommend. Therefore, as intended in Design Guideline 10, we need to implement open-ended control mechanisms that allow for reflection on the results with experts in their socio-technical context. This approach of using flexible control mechanisms should be similarly applied to managing and evaluating the value framework. The design guide includes design guidelines 5 and 8 based on the norms set by stakeholders. The norms in the value framework can change over time and require adjustments.

7.14. Reflection of experts on guidelines

Reflective meetings were set up with experts to validate whether the design guide solves the problem statement and is practical. Seven participants with different expert knowledge were asked to provide feedback on the design guidelines in individual meetings, and their roles are listed in a table.

Participant function	Stakeholder role	Participant ID
End-user (underwriter)	Operator	1
Information Risk Manager	Examiner	2
Operation Risk Manager	Examiner	3
Data scientist	Developer	5
Data scientist manager	Developer	6
Financial Risk manager	Examiner	8
Technology management	Examiner	9
Risk management	Examiner	10

Table 7.3: Participants of the reflection round.

A presentation was prepared and delivered to the participants to showcase the use of the design guidelines. During the presentation, the design guidelines were presented and visually represented through a process flow and dashboard. The presentation can be found in Appendix P. The process flow was omitted due to the confidentiality of the content, and the dashboard was purely for illustration purposes to provide an example of how design requirements can be measured. This explanation contributes to the understanding of the design guidelines.

From the interviews, several points and questions emerged for reflection:

- The multi-disciplinary team is a logical addition, but is it efficient in developing AI systems if involved throughout the life cycle (Risk Management, November 11, 2023)? Testing this process with a real-life project in the future was recommended.
- The sequence of steps two and three could be smoother. First, the scope must be established, then the effectiveness, and then a step back must be taken to adjust the scope (Risk Management, November 11, 2023; FERM, October 31, 2023). This feedback is incorporated into the guidelines.
- Adding independent experts who fit the use case is a good idea. It was recommended that this be applied to a real-life case (Risk Management, November 11, 2023).
- For design guideline two, the multi-disciplinary team must have access to a database where basic laws and policy documents can be found (Risk Management, November 11, 2023).
- Simplify the guidelines in layman's terms so everyone understands them (Risk Management, November 11, 2023). This feedback was incorporated by creating a guide with steps for what needs to be done.
- Applying such a dashboard requires certain expertise to interpret results (Technology Management, November 11, 2023; FERM, October 31, 2023; Data Scientist and Manager, October 31, 2023).
- Translating the norms into questions is preferred for practicality (Technology Management, November 11, 2023). This advice has yet to be done due to time constraints.

Most of the feedback is used iteratively to improve the guidelines on practicalities. Some feedback is directly used to improve the guidelines. Other feedback intends to use the guidelines in a real-life use case to understand the challenges. Due to time constraints, this iterative step was not made.

7.15. Conclusion

In conclusion, developing the design guidelines presented in this master's thesis is the final step of the design science research approach, integrating the results from the previous chapters to address the socio-technical complexities in AI systems. These guidelines operationalize a value framework emphasizing risk management and stakeholder values throughout the AI lifecycle, from conception to deployment. The guidelines are presented in Chapter 7.4 up until Chapter 7.13. The guidelines are reflected on by experts in the life insurance domain, and part of the feedback is adopted. This chapter aims to answer the following sub-question:

"Q5: What design guidelines can guide the development of AI systems?"

Answering the sub-question resulted in an artifact: guidelines for AI system developers to protect stakeholders' values. Following these guidelines contributes to the safe design, development, and deployment of AI systems and fosters continuous improvement within AI systems. It aims to protect the identified values of the stakeholders applicable to the PU system. This chapter highlights the need to translate the process steps done through this research into the practical environment of the Japanese life insurance organization. From a scientific point of view, this means that the different methods used are too extensive for practical work. The design of the guidelines shows how different parts of the methods used can be standardized and how some parts of the methods can be reused in the practical work field.

Concepts reusable in the guidelines for AI system developers to employ are two steps from the adjusted Garst et al. (2022) method to select relevant topics for documentation. These steps included identifying information sources and selecting topics from the assessment norms. This step is practical as it converges the focus of the controls that must be designed to protect stakeholders's values.

The other methods were standardized to shape the guideline steps. Safety constraints and safety control structures of Leveson (2012) contributed to the value framework to identify risks through the AI lifecycle. In addition, it contributes to where certain identified norms must be placed in the guideline steps. The AI governance framework and definitions proposed by Mäntymäki et al. (2022) were used to structure the guidelines for integrating them into an IT governance format for organizations. Specifically, the framework emphasizes the need to differentiate between data and AI assessment norms.

The guidelines outline a structured approach for stakeholders to establish AI systems that align with the defined norms of Chapter 5. The concept reinforces the principle of continuous improvement, which is central to regulatory risk-based approaches. Furthermore, it guides the effective management of open-ended control mechanisms to align with the principle of wholeness, considered a significant social norm in Japanese society. The content of the guidelines advocates for forming multi-disciplinary teams, defining clear project scopes, establishing criteria based on stakeholder needs, and conducting continuous assessment and refinement of the AI system. These steps are not static but involve iterative evaluations and feedback mechanisms to adapt to new information, challenges, and regulatory changes. The design guidelines show that societal, legal, industry, and business influences must be considered when establishing such a process.

However, difficulties can arise with the implementation of these guidelines. This research incorporates a reflection and validation of these guidelines with experts, which resulted in some possible challenges. First, introducing control mechanisms in new environments, such as fairness and explainability, demands domain-specific expertise that may be lacking within organizational settings. This expertise is crucial for identifying appropriate controls and making sufficient technical trade-offs within the socio-technical environment. Second, the involvement of different stakeholders introduces varied incentives, leading to complex trade-offs regarding openness, value protection, progress, and content (De Bruijn & Heuvelhof, 2018). In addition, including all these stakeholders can slow down the design, development, and deployment process. The dynamic environment surrounding the guidelines, especially in the rapidly evolving field of AI and human-machine interactions, requires that the guidelines be adaptable over time. Third, the design guidelines are derived from institutional constraints and potential risks. Integrating technical constraints, hazards, and the interaction between system design and institutional arrangements is not done comprehensively. Hence, it is recommended that further investigation be conducted into the design requirements and the associated trade-offs. Lastly, the robustness of the guidelines and framework, initially developed based on a single use case, should be tested across various use cases to continuously enhance and refine the guidelines.

Part IV

Conclusion and discussion

8

Conclusion and discussion

This research aimed to develop guidelines that developers can use to control AI systems from the start of the AI life cycle. The results presented in the previous chapters are discussed in this chapter to clarify the meaning of the results and the lessons learned. This chapter presents the main findings, the research's limitations, contributions to the knowledge base and practice, and recommendations for future research.

8.1. Main findings

The motivation for this study stems from a noticeable void in systematic guidance to protect stakeholders' values with AI in Japan's life insurance industry. In addition, an institutional value framework is missing within this particular industry. This void causes the practical problem, which states that it is unknown which social norms should be protected and implemented to control the technical operations of AI systems in organizational settings. The lack of clarity presents significant challenges for organizations in effectively controlling these risks, which, if not addressed, could lead to reputational damage in instances of failure.

Literature provides ways to tackle this issue. Van de Poel (2013) demonstrates from a philosophical perspective how to identify social values and translate them into design requirements. Nevertheless, this approach lacks practical implementation from a protection perspective (Aizenberg & Van den Hoven, 2020). Leveson (2012) presents safety concepts that offer possibilities for translating values into controls. Furthermore, Garst et al. (2022) propose several steps to condense the findings into contextually significant aspects relevant to organizations. Mäntymäki et al. (2022) established an AI governance framework to contextualize and incorporate these points within the organization. However, academic literature lacks a standard process for translating identified high-level values into practical organizational guidelines.

Therefore, this thesis contributes scientifically to an initial setup for identifying social norms for designing, developing, and deploying AI systems to protect and translate them to Japanese life insurance organizations' business operations. To reach this contribution, a use case sheds light on when an AI system is (not) complying with social norms in its perceived socio-technical context. Two steps are taken to control AI systems according to social norms.

First, a framework is created that fits the specific context. This creation is supported by using existing concepts from the traditional ”design for values” theory and concepts from system safety provided by Leveson (2012) to build a value framework. Second, it is ensured that AI developers can use this framework in an organizational setting. This step is taken by adapting concepts of the AI governance framework of Mäntymäki et al. (2022) and reporting standard steps by Garst et al. (2022). By taking these two steps, this thesis provides a process as a societal contribution by setting up practical guidelines for AI system developers to protect stakeholders' values and norms. This process can be repeated and improved in other similar situations. The practical contribution is reflected in creating a value framework for the Japanese life insurance industry and guidelines that AI system developers can follow to protect stakeholders' values and norms. These elements are encapsulated in the main research question of this thesis:

”What design guidelines can developers in Japan's life insurance domain follow to control AI systems while protecting stakeholders' values?”

8.2. The deliverable: guidelines

The objective of this research is to deliver design guidelines. Therefore, a few steps are taken to reach this objective.

First, the thesis started with a design science research approach, employing the PU system as a use case and attaching it to a design for values approach. Initially, an understanding of the environment is gained, focusing on the AI system's goals and restrictions by incorporating laws and stakeholder input. Second, a literature review identified pertinent social norms and values, further enriched by a content and comparison analysis of legal and industry documents. This phase also incorporated questionnaires distributed to the organization's management, leading to the identification of 13 values for safeguarding throughout the AI lifecycle. This step shows the need to identify several information sources to identify the values that must be protected. The third phase entailed translating these values into norms through an empirical study involving questionnaires, workshops, and interviews with eight participants. Utilizing concepts like safety constraints and the safety control structure results in this phase with 54 derived norms. However, for practical application, these were condensed to 31 assessment norms and 23 process norms using Garst et al. (2022) reporting standards, providing clarity for AI system developers on controls and procedural steps. These assessment norms were refined using the Mäntymäki et al. (2022) AI governance framework to align with the IT infrastructure of organizations. This step resulted in 18 data and 13 AI assessment norms. The fourth phase involved converting these norms into design requirements, emphasizing the need for expert insights into law, application, and stakeholder impact. This phase also entailed mapping relationships and trade-offs among these requirements. The final phase was a creative effort, tracing back the actions taken in this research and synthesizing all insights to formulate design guidelines, which served as the thesis's primary output to address the identified problem. This step highlights the need for backward traceability to ensure that the steps taken in this research are included. Section 7.2 shows how the results of each research phase are included in the guidelines. Table 7.1 shows the initial process of integrating the different concepts from the literature into the guidelines. In addition, the table shows the initial process that could be repeated and improved for further research.

The results of this research are guidelines that foster safe socio-technical AI systems' design, development, and deployment, driven by continuous improvement. Following these guidelines protects stakeholders' values, as they are integrated into the guideline design. Reaching this objective contributes to the practical and scientific fields. The guidelines efficiently guide AI system developers and organizations through the AI lifecycle, enhancing their work in the field. The guidelines are an effort to combine socio-cultural influences from society, institutions provided by authorities, and expert knowledge from an organizational point of view.

However, implementing these guidelines has its challenges. This research incorporates a reflection and validation of these guidelines with experts, which resulted in some possible challenges. Firstly, integrating fairness and explainability into new domains requires domain-specific knowledge to identify suitable controls and balance technical trade-offs. Second, the involvement of different stakeholders introduces varied incentives, leading to complex trade-offs regarding openness, value protection, progress, and content (De Bruijn & Heuvelhof, 2018). In addition, including all these stakeholders can slow down the process. Third, the design guidelines are derived from institutional constraints and potential risks. Integrating technical constraints, hazards, and the interaction between system design and institutional arrangements is not done comprehensively. Hence, it is recommended that further investigation be conducted into the design requirements and the trade-offs associated with them. Lastly, the robustness of the guidelines and framework, initially developed based on a single use case, should be tested across various use cases to continuously enhance and refine the guidelines.

8.3. Generalizability of results

This study obtained insights from the application domain, which helps make our results useful in many different situations. First, establishing the value framework was achieved through a method of triangulation. This framework's foundation is based on academic literature, further confirmed by incorporating legal documents, and subsequently validated through engagement with business stakeholders. Second, the information obtained was a synergistic blend of environmental practices and academic knowledge, resulting in a rich and varied viewpoint that remains grounded in reality. Experts provided insights across various domains related to the identified values, which were instrumental in operationalizing these values into

practical norms and guidelines. Lastly, the guidelines developed through this research underwent a validation process by experts. This validation served a function, providing a deeper understanding of how the guidelines could effectively assist the organization in identifying and controlling risks within its AI systems. The expert feedback affirmed the guidelines' relevance and applicability and highlighted their potential to enhance risk control measures within AI systems.

8.4. Limitations

The subsequent section of this report examines the limitations of this study, which aids in comprehending the significance of the research findings and offers guidance for future research.

8.4.1. Design science research

This research employed a design science approach, which encountered limitations in the practical environment. One such limitation was the partial disclosure of technical specifics of the model, a consequence of its classification as a high-risk model within the industry, which led to a potential distortion in understanding its impact compared to reality. Academic literature and discussions with stakeholders were utilized to supplement the system's outcomes to compensate for this.

A second limitation pertained to the selection of participants. For this study, discussions were held with 11 participants, chosen based on their expertise. However, a shortfall in this selection was the absence of a representative for the legal department. This gap meant the selected laws and regulations were primarily based on the underwriters' knowledge. Although the relevant documentation was available in English, some social norms embedded in these laws and regulations might have been overlooked. To mitigate this, individuals within the organization who did not have a direct stake in the research assisted in translating the documentation. Despite this effort, it is recommended for future work to review the framework of the selected laws and regulations with the help of legal and compliance experts.

Another environmental-based limitation was the decision to employ a multi-criteria decision analysis (MCDA). In this process, stakeholders were asked to prioritize their values using the PROMETHEE method. However, some participants found prioritizing challenging as they deemed all aspects important. Thus, the outcomes of the MCDA are used as inputs for a workshop, which makes the process more practical and accessible for the participants.

8.4.2. Design for values

This research utilized a combination of design for values and design science research approaches. Within the methodology of DfV, three limitations need to be acknowledged.

First, a limitation was encountered during the literature selection for the integrative literature review. In identifying social values, available literature was scarce. To address this, an individual within the organization validated the findings from the literature to determine the applicability of the identified values. Nonetheless, further research into the interaction between Japanese society and AI would be beneficial to enrich understanding in this area.

Second, a limitation of the framework was that the empirical findings at the values, norms, and design requirements levels were based solely on internal stakeholders. There was no opportunity to engage with customers or agents to incorporate their perspectives on value protection regarding AI systems. Hence, the end-user perspective is considered. However, it is important to note that this approach inevitably introduces a certain degree of bias.

Third, a limitation concerns the value hierarchy process based on a single use case. While the use case provided valuable insights into its development, the intentions of the system, and its environmental impact, these insights were instrumental in establishing the value framework and guidelines. However, the limitation of relying on just one use case is that it narrows the scope of risks that can be identified, as risks are context-dependent. Incorporating a second use case could have led to more comprehensive insights. Therefore, broader questions were posed during the semi-structured interviews. These questions were focused on more than just the PU system, thereby expanding the scope of the inquiry.

8.4.3. Design guidelines

A notable limitation in the development of the design guidelines must be acknowledged. Specifically, there was restricted access to develop the model further, resulting in challenges in formulating actionable design requirements. Due to the inability to establish fairness metrics, more insights into potential thresholds and conversations for control mechanisms were needed. Consequently, while discussions with end-users about possible scenarios were conducted to form a control mechanism, this approach did not fully comprehend how these mechanisms would function in practice. Further research is thus required to explore these control mechanisms' practical application and effectiveness.

8.5. Research contribution

The contribution of this research can be categorized as a scientific contribution and a contribution to the life insurance industry.

8.5.1. Scientific Contribution

This research introduces a replicable and improvable initial process for incorporating social norms into AI system operations within organizational settings, addressing the gap in existing literature on transforming identified values into practical guidelines in AI. The study refines these abstract concepts into practical guidelines by utilizing Van den Hoven et al. (2015) design for values theory and Van de Poel's (2015) value hierarchy. The research complements the value hierarchy with other methods found in the literature that determine which social norms and how to implement them in the technical operations of AI systems should be protected in an organizational setting. Therefore, system safety concepts by Leveson (2012) assist from a safety lens; Garst et al. (2022) aid in selecting relevant topics to make them manageable in an organizational context; and Mäntymäki et al. (2022) structure the topics on the IT governance level in an organizational context. This approach is validated through empirical study, providing evidence of its effectiveness and utility in practical settings.

In addition, the replicated and improvable initial process also contributes by integrating established knowledge into a novel environment. Different concepts of literature related to the problem statement were used, such as Van den Hoven et al. (2015), Leveson (2012), Garst et al. (2022), and Mäntymäki et al. (2022). Incorporating knowledge from this literature to design for values allows AI researchers to build further research on them.

The second contribution is the identified values in the Japanese life insurance industry. This contribution addresses the gap in the current literature on design for values. There is little research on consolidating these high-level principles and requirements into the Japanese industry context. 54 norms encompassing 13 values were identified from societal, legal, industry, and business perspectives. The synthesized values contribute to the scientific knowledge base as new information that could be reused in other research. The text analysis offers insights into Japanese definitions of values in AI within the life insurance industry. Similarities and differences in values and risk approaches were identified by comparing these values with those found in European documents. This cross-cultural analysis enriches our understanding of global AI development and its diverse impacts.

The last contribution adds empirical evidence of using guidelines through a use case. As mentioned by Mittelstadt (2019), there is a need for further research and empirical data to implement AI systems effectively within their specific contexts. Creating guidelines through empirical evidence leads to another contribution to the scientific knowledge base. This study offers empirical evidence of its approach to operationalizing values in practice from a safety perspective. It is translated into workable guidelines in the AI domain for the Japanese life insurance industry. These guidelines enhance the scientific knowledge base and offer a foundation for further research and adaptation across various AI applications.

8.5.2. Practical contribution

The problem addresses the challenge of determining which social norms should be protected and implemented to control the technical operations of AI systems in organizational settings. Therefore, this thesis contributes to a value framework and design guide consisting of practical guidelines for Japan's life insurance industry. The guidelines are designed for AI system developers in organizations and the Japanese life insurance industry.

First, the practical contributions of the AI system developers in organizations. AI system developers are

involved in AI use cases or are about to start with AI use cases. The guide provides a starting point for AI system developers to start a project and points to evaluating existing AI systems. Using the guide will consolidate a focused view of AI system vulnerabilities that could harm stakeholders' values. The influence of social norms is another factor that shapes the guidelines that make it fit in its environment. This guide will contribute to designing and implementing safe AI systems that foster continuous improvement.

Second, practical contribution to the life insurance industry and regulators. Creating the value framework provides insights into the needs of society, legal, industry, and businesses. The guide aids in anticipating and identifying risks through continuous improvement rather than intervening after the damage has occurred.

Third, the organization where this research is employed. Participants in the evaluation session confirmed that this method significantly contributes to local practices by providing tangible deliverables for risk control activities and system design constraints.

8.6. Recommendation for future research

The first recommendation emerges from the observed limitations and pertains to the guidelines. Notably, since the current guidelines are founded on a single use case, it is recommended that this process be replicated across various use cases. This recommendation is further reinforced by the reflections and validations provided by experts in the field in Chapter 7. An initial step could involve exploring use cases within the same industry, specifically the life insurance sector. This approach would facilitate gradual adjustments to the guidelines, addressing the nuances and differences identified in similar industry contexts. Also, it is advised to incorporate processes with differing objectives. For instance, exploring the application of a large language model that functions as an organizational knowledge base or a system not related to customer data would provide valuable insights. Including a diverse range of use cases would significantly broaden the understanding and applicability of the guidelines and framework. Such an expansion would enhance the robustness of the guidelines and ensure their relevance and effectiveness across varying contexts and operational purposes.

The second recommendation suggests reevaluating the value framework, incorporating perspectives from a broader range of stakeholders. This review should include inputs from the legal department, social stakeholders, agents, customers, and society to discern variations in their interactions and prioritization choices. Such an inclusive approach enriches the understanding of how different stakeholders perceive and prioritize values within the context of AI. Integrating these diverse viewpoints in an empirical setting could create a complex challenge but provide a more robust value framework and a better-focused process.

The third recommendation advocates for research into the interaction dynamics between Japanese society and machines. The existing research highlights a notable need for more literature concerning the values held by Japanese society towards AI. Accordingly, a recommendation is that researchers build upon the value framework from this research by investigating these human-machine interactions with a focus on the socio-cultural aspects across diverse contexts. One practical approach would be to examine the interaction with specific systems, such as the PU system, to understand societal perceptions and attitudes toward such technologies. Delving into these specifics would provide deeper insights into societal expectations and needs regarding AI. This enhanced understanding supports more accurately identifying potential risks and tailoring AI systems to align better with societal values and norms.

The concluding recommendation proposes an extended examination of the system theoretic hazard analysis, part of system safety. The initial phase of this process has been undertaken, utilizing the safety control structure as a methodology within the empirical data collection phase to establish institutional constraints and hazards. Building upon this groundwork, the next logical progression for us involves testing the technical constraints associated with the design of the AI system. Further investigation will validate and refine the established institutional constraints and ensure the AI system's technical aspects are thoroughly evaluated and optimized for safety. In addition, further research would provide a concrete analysis of the conflicts that could arise within the organizational setting and offer possibilities to explore trade-offs.

8.7. Personal reflection

This research journey has been an enlightening experience in applying theoretical frameworks and methodologies in practice. One notable aspect of this process was the discovery that theories sometimes

unfold differently in practical settings than anticipated. A prime example was prioritizing values using the Multi-Criteria Decision Analysis (MCDA) questionnaire with operational stakeholders. This exercise revealed varied responses, with some stakeholders finding it challenging to prioritize and marking all options as equally important, while others completed it as expected. This experience highlighted the complexity and subjectivity involved in such exercises, underscoring the need for flexibility and adaptability in research methodologies. Another example is that I built my research as a constructive technology assessment in the first place. After the first meeting and interaction with my stakeholders, I acknowledged that this approach would only work if it were more intensive in an organizational life insurance setting.

A significant factor in my research was the cultural and linguistic differences encountered within a Japanese organizational setting. Navigating these differences required finding effective communication methods, adapting to the local environment, and being aware of my Western perspective as a bias. This experience, far from being a hindrance, was a motivation for deeper understanding and adaptation. It improved my awareness of approaching various situations and engaging with my environment within and beyond the research context. Building trust with participants, for instance, became crucial. Additionally, experiencing the importance of hierarchy within Japanese organizations and building relationships based on trust provided a more profound understanding than theoretical knowledge alone could offer. These insights underscore the importance of factoring in cultural aspects when conducting socio-technical research, as they significantly influence the research timeline and process.

A personal challenge I faced was a propensity to become distracted by unimportant details, occasionally leading me away from my main objectives. This aspect of my approach sometimes led to a shift in focus away from the end goal. This could be seen in my writing style, where the text's focus was lost. My supervisor's valuable tip is to write down topic sentences and sometimes take a break from writing or working to restructure everything. It is still a difficult task, but I am still trying!

Reflecting on the process, I would certainly approach certain aspects differently with the benefit of hindsight. One such aspect is having a clearer vision of the end product from the outset. The research initially aimed to develop an academic value framework, but it evolved to include a design guide to enhance practicality for the organizational context. Hence, it's important to pose focused questions and provide practical examples of what the end product might entail and how it could function. This approach would facilitate a more streamlined journey toward the research objectives and select the right frameworks, literature, participants, and so on at a faster pace.

Should the opportunity arise again to take on a similar challenge, I would answer with a definite yes. The experience provided me with a wealth of knowledge and insights, enough to write extensive stories about what I learned. This assignment was set in Japan and deeply intertwined with the Japanese context, significantly enriching my academic understanding and awareness.

8.8. Link with the Complex Systems Engineering and Management program

This research is conducted as part of the Complex Systems Engineering and Management (CoSEM) master's program, which emphasizes the design of socio-technical systems through a multidisciplinary lens to address complex real-world challenges. The study explores issues using philosophical concepts, such as design for values, in a socio-technical environment. The study aims to understand how theoretical concepts work in the real application environment. Therefore, the end deliverable, namely the guide, aims to protect stakeholder values within its socio-technical context by combining theoretical concepts.

This research involves understanding the complexities of how social norms must be considered in an AI system. In addition, the study explores the goals and restrictions of the system by using social norms, laws, and the system's technical capabilities. The study will employ a multi-faceted approach, examining technology from an institutional and process lens. The exploration provides insights into the interconnectedness between technical, process, and institutional factors, forming the basis for a value framework that translates stakeholders' values into AI system controls.

Moreover, this research adheres to the principles outlined in the CoSEM program, attempting to create an artifact, namely design guidelines, within the dynamic and complex system by the guidance of a use case. The artifact is constructed by integrating elements from various frameworks and methodologies,

aligning with the CoSEM program’s objective of addressing real-world socio-technical challenges through innovative and interdisciplinary solutions.

9

References

Aggour, K. S., & Cheetham, W. (2005). Automating the underwriting of insurance applications. Innovative Applications of Artificial Intelligence, 27(3), 1451–1458. https://doi.org/10.1609/aimag.v27i3.1891

AI HLEG. (2021). Assessment List for Trustworthy Artificial Intelligence (ALTAI) for self-assessment. In High-level Expert Group on Artificial Intelligence. https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai

Aizenberg, E., & Van Den Hoven, J. (2020). Designing for human rights in AI. Big Data & Society, 7(2), 205395172094956. https://doi.org/10.1177/2053951720949566

Alla, S., & Adari, S. K. (2021). Beginning MLOps with MLFlow. In Apress eBooks. https://doi.org/10.1007/978-1-4842-6549-9

Alzubi, J. A., Nayyar, A., & Kumar, A. (2018). Machine Learning from Theory to Algorithms: An Overview. Journal of Physics, 1142, 012012. https://doi.org/10.1088/1742-6596/1142/1/012012

Ashok, M., Madan, R., Joha, A., & Sivarajah, U. (2022). Ethical framework for Artificial Intelligence and Digital technologies. International Journal of Information Management, 62, 102433. https://doi.org/10.1016/j.ijinfomgt.2021.102433

Avraham, R. (2017). Discrimination and Insurance [PDF]. The Routledge Handbook of the Ethics of Discrimination, 13. https://doi.org/10.4324/9781315681634

Azzutti, A., Ringe, W., & Stiehl, H. S. (2022). The Regulation of AI Trading from an AI Life Cycle Perspective. Social Science Research Network. https://doi.org/10.2139/ssrn.4260423

Baran-Kooiker, A., Czech, M., & Kooiker, C. (2018). Multi-Criteria Decision Analysis (MCDA) Models in Health Technology Assessment of Orphan Drugs—a Systematic Literature review. Next steps in methodology development? Frontiers in Public Health, 6. https://doi.org/10.3389/fpubh.2018.00287

Bauer, J. M., & Herder, P. M. (2009). Designing Socio-Technical systems. In Elsevier eBooks (pp. 601–630). https://doi.org/10.1016/b978-0-444-51667-1.50026-4

Berberich, N., Nishida, T., & Suzuki, S. (2020). Harmonizing artificial intelligence for social good. Philosophy & Technology, 33(4), 613–638. https://doi.org/10.1007/s13347-020-00421-8

Berghoff, C., Neu, M., & Von Twickel, A. (2020). Vulnerabilities of connectionist AI Applications: Evaluation and defense. Frontiers in Big Data, 3. https://doi.org/10.3389/fdata.2020.00023

Bhuiyan, N., & Baghel, A. (2005). An overview of continuous improvement: from the past to the present. Management Decision, 43(5), 761–771. https://doi.org/10.1108/00251740510597761

Birkstedt, T., Minkkinen, M., Tandon, A., & Mäntymäki, M. (2023). AI governance: themes, knowledge gaps and future agendas. Internet Research, 33(7), 133–167. https://doi.org/10.1108/intr-01-2022-0042

Bossen, C. (2018). Socio-technical betwixtness. In Elsevier eBooks (pp. 77–94). https://doi.org/10.1016/b978-0-12-812583-0.00005-5

Bradford, A. (2020). The Brussels effect: How the European Union rules the world. https://scholarship.law.columbia.edu/books/232/

Burston, D., Howell Charlie, Gilchrist, A., & Simpson, P. (2020). Artificial Intelligence: The ethical use of AI in the life insurance sector. Milliman. Retrieved October 21, 2023, from https://us.milliman.com/en/insight/artificial-intelligence-the-ethical-use-of-ai-in-the-life-insurance-sector

Cabinet Secretariat: Council for Social Principles of Human-centric AI. (2019). Social Principles of Human-Centric AI. Integrated Innovation Strategy Promotion Council. Retrieved August 3, 2023, from https://www.cas.go.jp/jp/seisaku/jinkouchinou/pdf/humancentricai.pdf

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., & Cosentini, A. (2022). A clarification of the nuances in the fairness metrics landscape. Scientific Reports, 12(1). https://doi.org/10.1038/s41598-022-07939-1

Ceylan, İ. E. (2022). The effects of artificial intelligence on the insurance sector: emergence, applications, challenges, and opportunities. In Accounting, finance, sustainability, governance & fraud (pp. 225–241). https://doi.org/10.1007/978-981-16-8997-013

Chen, R. J., Wang, J. J., Williamson, D. F. K., Chen, T., Lipková, J., Lu, M., Sahai, S., & Mahmood, F. (2023). Algorithmic fairness in artificial intelligence for medicine and healthcare. Nature Biomedical Engineering, 7(6), 719–742. https://doi.org/10.1038/s41551-023-01056-8

Code of Conduct. (2018). The Life Insurance Association of Japan. Retrieved August 16, 2023, from https://www.seiho.or.jp/activity/sdgs/rule/pdf/pdf.pdf

Cummins, J. D., & Weiss, M. A. (2013). Analyzing firm performance in the insurance industry using frontier efficiency and productivity methods. In Springer eBooks (pp. 795–861). https://doi.org/10.1007/978-1-4614-0155-128

Cyphers, B., & Rodriguez, H. S. K. (2021, May 14). Japan's Rikunabi Scandal Shows The Dangers of Privacy Law Loopholes. Electronic Frontier Foundation. https://www.eff.org/deeplinks/2021/05/japans-rikunabi-scandal-shows-dangers-privacy-law-loopholes

De Bruijn, H., & Heuvelhof, E. T. (2018). Management in networks. In Routledge eBooks. https://doi.org/10.4324/9781315453019

De Pagter, J. (2023). From EU Robotics and AI Governance to HRI Research: Implementing the ethics Narrative. International Journal of Social Robotics. https://doi.org/10.1007/s12369-023-00982-6

De Silva, D., & Alahakoon, D. (2022). An artificial intelligence life cycle: From conception to production. Patterns, 3(6), 100489. https://doi.org/10.1016/j.patter.2022.100489

Deldjoo, Y., Anelli, V. W., Zamani, H., Bellogín, A., & Di Noia, T. (2019). Recommender Systems fairness evaluation via generalized cross entropy. arXiv (Cornell University). https://arxiv.org/pdf/1908.06708.pdf

Dobbe, R. (2022). System safety and artificial intelligence. 2022 ACM Conference on Fairness, Accountability, and Transparency. https://doi.org/10.1145/3531146.3533215

Dobbe, R., Gilbert, T. W., & Mintz, Y. (2021). Hard choices in artificial intelligence. Artificial Intelligence, 300, 103555. https://doi.org/10.1016/j.artint.2021.103555

Eggert, M. (2014). Compliance Management in Financial Industries: A model-based business process and reporting perspective. http://ci.nii.ac.jp/ncid/BB18734071

EIOPA publishes report on artificial intelligence governance principles. (2021, June 17). European Insurance and Occupational Pensions Authority. https://www.eiopa.europa.eu/eiopa-publishes-report-artificial-intelligence-governance-principles-2021-06-17en?source=search

Falco, G., Shneiderman, B., Badger, J., Carrier, R., Dahbura, A., Danks, D., Eling, M., Goodloe, A. E., Gupta, J. P., Hart, C., Jirotka, M., Johnson, H., Lapointe, C., Llorens, A. J., Mackworth, A. K., Maple, C., Pálsson, S., Pasquale, F. A., Winfield, A. F. T., & Yeong, Z. K. (2021). Governing AI safety through independent audits. Nature Machine Intelligence, 3(7), 566–571. https://doi.org/10.1038/s42256-021-00370-7

Fang, K., Jiang, Y., & Song, M. (2016). Customer profitability forecasting using Big Data analytics: A case study of the insurance industry. Computers & Industrial Engineering, 101, 554–564. https://doi.org/10.1016/j.cie.2016.09.011

Financial Service Agency. (2023). Comprehensive Guidelines for Supervision of Financial Instruments Business Operators, etc. In Financial Service Agency. FSA. Retrieved October 20, 2023, from https://www.fsa.go.jp/common/law/guide/kinyushohineng.pdf

Financial Services Agency. (2021). Comprehensive Guidelines for Supervision for Insurance Companies. In FSA. FSA Japan. Retrieved September 20, 2023, from https://www.fsa.go.jp/common/law/guide/enins.pdf

Financial Services Agency. (2023). https://www.fsa.go.jp/en/refer/cold/index.html. Cold Calling. Retrieved September 20, 2023, from https://www.fsa.go.jp/en/refer/cold/index.html

Følstad, A. (2017). Users' design feedback in usability evaluation: a literature review. Human-centric Computing and Information Sciences, 7(1). https://doi.org/10.1186/s13673-017-0100-y

Friedman, B., Harbers, M., Hendry, D. G., Van Den Hoven, J., Jonker, C. M., & Logler, N. (2021). Eight grand challenges for value sensitive design from the 2016 Lorentz workshop. Ethics and Information Technology, 23(1), 5–16. https://doi.org/10.1007/s10676-021-09586-y

Garst, J., Maas, K., & Suijs, J. (2022). Materiality assessment is an art, not a science: Selecting ESG topics for sustainability reports. California Management Review, 65(1), 64–90. https://doi.org/10.1177/00081256221120692

General Life Insurance Association of Japan. (2020). Matters to Note regarding Insurance Solicitation (Compliance guide for solicitation). GIAJ. Retrieved September 20, 2023, from https://www.sonpo.or.jp/en/about/ue089i0000002nnb-att/ComplianceguideforsolicitationSection4.pdf

Glenn, B. J. (2003). Postmodernism: the basis of insurance. Risk Management and Insurance Review, 6(2), 131–143. https://doi.org/10.1046/j.1098-1616.2003.028.x

Greenleaf, G., & Shimpo, F. (2014). The puzzle of Japanese data privacy enforcement. International Data Privacy Law, 4(2), 139–154. https://doi.org/10.1093/idpl/ipu007

Haakman, M., Cruz, L., Huijgens, H., & Van Deursen, A. (2020). AI Lifecycle Models Need To Be Revised. An Exploratory Study in Fintech. Empir. Software Eng., 26. https://doi.org/10.1007/s10664-021-09993-1

Habuka, H. (2023). Japan's Approach to AI Regulation and Its Impact on the 2023 G7 Presidency. In Center for Strategies and International Studies. CSIS. Retrieved September 29, 2023, from https://www.csis.org/analysis/japans-approach-ai-regulation-and-its-impact-2023-g7-presidency

Hevner, A. R. (2007). A three cycle view of design science research. Scandinavian Journal of Information Systems, 19(2), 4. http://community.mis.temple.edu/seminars/files/2009/10/Hevner-SJIS.pdf

Hiroshi, O. (2019). AML/CFT and new technologies: Challenges in Japan. Journal of Financial Compliance, 2(4), 342–361.

Imai, M. (2007). Gemba Kaizen. A commonsense, Low-Cost approach to management. In Gabler eBooks (pp. 7–15). https://doi.org/10.1007/978-3-8349-9320-52

Ishihara, K. (2006). Reputation management in the Japanese insurance marketplace. Geneva Papers on Risk and Insurance-issues and Practice, 31(3), 446–453. https://doi.org/10.1057/palgrave.gpp.2510089

Jacob, F., Pez, V., & Volle, P. (2021). Principles, methods, contributions, and limitations of design science research in marketing: Illustrative application to customer journey management. Recherche Et Applications En Marketing, 37(2), 2–29. https://doi.org/10.1177/20515707211032537

Jaiswal, R. (2023). Impact of AI in the General Insurance underwriting factors. Central European Management Journal, 31(2), 697–705. https://doi.org/10.57030/23364890.cemj.31.2.72

Johannesson, P., & Perjons, E. (2014). An introduction to design science. In Springer eBooks. https://doi.org/10.1007/978-3-319-10632-8

Kallus, N., Mao, X., & Zhou, A. (2022). Assessing Algorithmic Fairness with Unobserved Protected Class Using Data Combination. Management Science, 68(3), 1959–1981. https://doi.org/10.1287/mnsc.2020.3850

Katirai, A. (2023). The ethics of advancing artificial intelligence in healthcare: analyzing ethical considerations for Japan's innovative AI hospital system. Frontiers in Public Health, 11. https://doi.org/10.3389/fpubh.2023.1142062

Kelley, K. H., Fontanetta, L. M., Heintzman, M., & Pereira, N. (2018). Artificial intelligence: Implications for social inflation and insurance. Risk Management and Insurance Review, 21(3), 373–387. https://doi.org/10.1111/rmir.12111

Khambatta, P., Matz, S., & Wang, D. (2021). AI and Algorithmic Decision Making: Exploring Their Promises, Perils, And Pitfalls. Proceedings - Academy of Management, 2021(1), 12824. https://doi.org/10.5465/ambpp.2021.12824symposium

Kozuka, S. (2019). A governance framework for the development and use of artificial intelligence: lessons from the comparison of Japanese and European initiatives. Uniform Law Review, 24(2), 315–329. https://doi.org/10.1093/ulr/unz014

Kroes, P., Franssen, M., Van De Poel, I., & Ottens, M. (2006). Treating socio-technical systems as engineering systems: some conceptual problems. Systems Research and Behavioral Science, 23(6), 803–814. https://doi.org/10.1002/sres.703

Kumar, N., Srivastava, J. D., & Bisht, H. (2019). Artificial intelligence in insurance sector. Journal of the Gujarat Research Society, 21(7), 79–91. http://gujaratresearchsociety.in/index.php/JGRS/article/view/405

Kurshan, E., Shen, H., & Chen, J. (2020). Towards self-regulating AI. https://doi.org/10.1145/3383455.3422564

Larsson, S. (2020). On the Governance of Artificial Intelligence through Ethics Guidelines. Asian Journal of Law and Society, 7(3), 437–451. https://doi.org/10.1017/als.2020.19

Lee, M. S. A., & Singh, J. (2020). The landscape and gaps in open source fairness toolkits. Social Science Research Network. https://doi.org/10.2139/ssrn.3695002

Leveson, N. G. (2012). Engineering a safer world: systems thinking applied to safety. Choice Reviews Online, 49(11), 49–6305. https://doi.org/10.5860/choice.49-6305

Lindholm, Mathias and Richman, Ronald and Tsanakas, Andreas and Wuthrich, Mario V., What is Fair? Proxy Discrimination vs. Demographic Disparities in Insurance Pricing (May 24, 2023). Available at SSRN: https://ssrn.com/abstract=4436409 or http://dx.doi.org/10.2139/ssrn.4436409

Loi, M., & Christen, M. (2021). Choosing how to discriminate: navigating ethical trade-offs in fair algorithmic design for the insurance sector. Philosophy & Technology, 34(4), 967–992. https://doi.org/10.1007/s13347-021-00444-9

López-Paz, D., Bouchacourt, D., Sagun, L., & Usunier, N. (2022). Measuring and signing fairness as performance under multiple stakeholder distributions. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2207.09960

Lundin, M., & Eriksson, S. (2016a). Artificial Intelligence in Japan (R &D, Market and Industry Analysis). In EU-Japan. Tokyo: EU-Japan Centre for Industrial Cooperation.

Lundin, M., & Eriksson, S. (2016b). Artificial Intelligence in Japan (R &D, Market and Industry Analysis). In EU-JAPAN CENTRE FOR INDUSTRIAL COOPERATION. EU-JAPAN CENTRE FOR INDUSTRIAL COOPERATION.

Maier, M. E., Carlotto, H., Saperstein, S., Sanchez, F., Balogun, S., & Merritt, S. A. (2020). Improving the Accuracy and Transparency of Underwriting with Artificial Intelligence to Transform the Life□Insurance Industry. Ai Magazine, 41(3), 78–93. https://doi.org/10.1609/aimag.v41i3.5320

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

Mamiko, Y. A. (2020). The Impact of Big Data and Artificial Intelligence (AI) in the Insurance Sector. In Policy Commons (20.500.12592/cwxpqb). OECD Organisation for Economic Co-operation and Development. Retrieved August 7, 2023, from https://policycommons.net/artifacts/3864431/the-impact-of-big-data-and-artificial-intelligence-ai-in-the-insurance-sector/4670386/

Mäntymäki, M., Minkkinen, M., Birkstedt, T., & Viljanen, M. (2022). Defining organizational AI governance. AI And Ethics, 2(4), 603–609. https://doi.org/10.1007/s43681-022-00143-x

McGeveran, W. (2016). Friending the Privacy Regulators. ARIZONA LAW REVIEW, 58(959). https://scholarship.law.umn.edu/cgi/viewcontent.cgi?article=1627&context=facultyarticles

McStay, A. (2021). Emotional AI, ethics, and Japanese spice: contributing community, wholeness, sincerity, and heart. Philosophy & Technology, 34(4), 1781–1802. https://doi.org/10.1007/s13347-021-00487-y

Metcalf, J., Moss, E., Watkins, E. A., Singh, R., & Elish, M. C. (2021). Algorithmic Impact Assessments and Accountability. FAccT. https://doi.org/10.1145/3442188.3445935

Methnani, L., Brännström, M., & Theodorou, A. (2023). Operationalising AI Ethics: Conducting socio-technical assessment. In Lecture Notes in Computer Science (pp. 304–321). https://doi.org/10.1007/978-3-031-24349-316

Michelfelder, D. P., & Doorn, N. (2020). The Routledge Handbook of the Philosophy of Engineering. In Routledge eBooks. https://doi.org/10.4324/9781315276502

Microsoft. (n.d.). Microsoft Copilot for Microsoft 365. Your AI Assistent at Work. Retrieved November 1, 2023, from https://www.microsoft.com/en-us/microsoft-365/enterprise/copilot-for-microsoft-365tabs-oc2a1ctab1

Minkkinen, M., Zimmer, M. P., & Mäntymäki, M. (2021). Towards ecosystems for responsible AI. In Lecture Notes in Computer Science (pp. 220–232). https://doi.org/10.1007/978-3-030-85447-820

Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. Nature Machine Intelligence, 1(11), 501–507. https://doi.org/10.1038/s42256-019-0114-4

Miyashita, H. (2016). A Tale of Two Privacies: Enforcing Privacy with Hard Power and Soft Power in Japan. In Law, governance and technology series. Springer International Publishing. https://doi.org/10.1007/978-3-319-25047-25

Miyashita, H. (2021). Human-centric data protection laws and policies: A lesson from Japan. Computer Law & Security Review, 40, 105487. https://doi.org/10.1016/j.clsr.2020.105487

Morfoulaki, M., & Papathanasiou, J. (2021). Use of PROMETHEE MCDA method for ranking alternative measures of sustainable urban mobility planning. Mathematics, 9(6), 602. https://doi.org/10.3390/math9060602

Morita, A. (2012). A neo-communitarian approach on human rights as a cosmopolitan imperative in East Asia. Filosofia Unisinos. https://doi.org/10.4013/fsu.2012.133.01

Mourtzis, D., Angelopoulos, J., & Panopoulos, N. (2022). A Literature Review of the Challenges and Opportunities of the Transition from Industry 4.0 to Society 5.0. Energies, 15(17), 6276. https://doi.org/10.3390/en15176276

Mullins, M., Holland, C. P., & Cunneen, M. (2021). Creating ethics guidelines for artificial intelligence and big data analytics customers: The case of the consumer European insurance market. Patterns, 2(10), 100362. https://doi.org/10.1016/j.patter.2021.100362

Murata, K. (2019). Japanese Traditional Vocational Ethics: Relevance and meaning for the ICT-dependent society. In Tetsugaku companions to Japanese philosophy (pp. 139–160). https://doi.org/10.1007/978-3-319-59027-17

Narayanan, A. (2018). 21 fairness definition and their politics. Proc. Conf. Fairness Accountability Transp, 3.

Nishikino, H., & Kanazawa, K. (2018). https://www.clo.jp/wp-content/uploads/2018/03/Insurance2018Jp.pdf. Chuo Sogo Law Office. Retrieved September 20, 2023, from https://www.clo.jp/wp-content/uploads/2018/03/Insurance2018Jp.pdf

Papenmeier, A., Kern, D., Englebienne, G., & Seifert, C. (2022). It's Complicated: The Relationship between User Trust, Model Accuracy and Explanations in AI. ACM Transactions on Computer-Human Interaction, 29(4), 1–33. https://doi.org/10.1145/3495013

Pedrini, M., & Ferri, L. M. (2019). Stakeholder management: a systematic literature review. Corporate Governance, 19(1), 44–59. https://doi.org/10.1108/cg-08-2017-0172

Proposal for a REGULATION OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL LAYING DOWN HARMONISED RULES ON ARTIFICIAL INTELLIGENCE (ARTIFICIAL INTELLIGENCE ACT) AND AMENDING CERTAIN UNION LEGISLATIVE ACTS. (2021). In EUR-Lex (No. 52021PC0206). European Commission. Retrieved August 29, 2023, from https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex

Quang, D. X. (2017). Predictive Underwriting vs Traditional Underwriting with Data Science approach. Ressources Actuarielles, 1–105.

Radu, R. (2021). Steering the governance of artificial intelligence: national strategies in perspective. Policy and Society, 40(2), 178–193. https://doi.org/10.1080/14494035.2021.1929728

Richardson, H. S. (1994). Practical Reasoning about Final Ends. https://doi.org/10.1017/cbo9781139174275

Rothstein, M. A., & Joly, Y. (2009). Genetic information and insurance underwriting: contemporary issues and approaches in the global economy. In The Handbook of Genetics & Society: Mapping the New Genomic Era (p. 108). Routlegde. hrefhttps://books.google.co.jp/books?id=KLZ8AgAAQBAJ&printsec=frontcover=onepage&q&f=false

Ruf, B., & Detyniecki, M. (2021). Towards the right kind of fairness in AI. arXiv (Cornell University). https://arxiv.org/pdf/2102.08453.pdf

Sapraz, M., & Han, S. (2022). Translating Human Values to Design Requirements: The Case of Developing Digital Government Collaborative Platform (DGCP) for Environmental Sustainability in Sri Lanka. DG.O 2022. https://doi.org/10.1145/3543434.3543455

Schwartz, R., Vassilev, A., Greene, K., Perine, L., Burt, A., & Hall, P. B. (2022). Towards a standard for identifying and managing bias in artificial intelligence. https://doi.org/10.6028/nist.sp.1270

Shilton, K. (2012). Values levers. Science, Technology, & Human Values, 38(3), 374–397. https://doi.org/10.1177/0162243912436985

Signorello, L. B., Cohen, S. S., Williams, D. R., Munro, H. M., Hargreaves, M., & Blot, W. J. (2014). Socioeconomic Status, Race, and Mortality: A Prospective Cohort study. American Journal of Public Health, 104(12), e98–e107. https://doi.org/10.2105/ajph.2014.302156

Solove, D. J. (2023). Data is what data does: regulating use, harm, and risk instead of sensitive data. Social Science Research Network. https://doi.org/10.2139/ssrn.4322198

Stahl, B. C., & Leach, T. (2022). Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: an empirical test of the European Union Assessment List for Trustworthy AI (ALTAI). AI And Ethics, 3(3), 745–767. https://doi.org/10.1007/s43681-022-00201-4

Taeihagh, A. (2021). Governance of artificial intelligence. Policy and Society, 40(2), 137–157. https://doi.org/10.1080/14494035.2021.1928377

Tamiya, N., Noguchi, H., Nishi, A., Reich, M. R., Ikegami, N., Hashimoto, H., Shibuya, K., Kawachi, I., & Campbell, J. C. (2011). Population ageing and wellbeing: lessons from Japan's long-term care insurance policy. The Lancet, 378(9797), 1183–1192. https://doi.org/10.1016/s0140-6736(11)61176-8

The Ministry of Economy, Trade, and Industry. (2021). AI Governance in Japan Ver. 1.1. In METI. METI.

Torraco, R. J. (2005). Writing Integrative Literature Reviews: Guidelines and Examples. Human Resource Development Review, 4(3), 356–367. https://doi.org/10.1177/1534484305278283

Toshmurzaevich, Y. O. (2020). Developing the underwriting process in life insurance. European Journal of Business and Management Research, 5(6). https://doi.org/10.24018/ejbmr.2020.5.6.657

Truby, J., Brown, R. D., & Dahdal, A. (2020). Banking on AI: mandating a proactive approach to AI regulation in the financial sector. Law And Financial Markets Review, 14(2), 110–120. https://doi.org/10.1080/17521440.2020.1760454

Van De Poel, I. (2013). Translating Values into Design Requirements. In Philosophy of engineering and technology (pp. 253–266). https://doi.org/10.1007/978-94-007-7762-020

Van De Poel, I. (2018). Design for value change. Ethics and Information Technology, 23(1), 27–31. https://doi.org/10.1007/s10676-018-9461-9

Van De Poel, I. (2020). Embedding Values in Artificial Intelligence (AI) Systems. Minds and Machines, 30(3), 385–409.
https://doi.org/10.1007/s11023-020-09537-4

Van Den Hoven, M., Vermaas, P. E., & Van De Poel, I. (2015). Handbook of ethics, values, and technological design: sources, theory, values and application domains. In Springer eBooks.
http://ci.nii.ac.jp/ncid/BB20251988?l=en

Van Der Heide, A. (2023). Dealing in uncertainty: Insurance in the Age of Finance. Policy Press.

Veluwenkamp, H., & Van Den Hoven, J. (2023). Design for values and conceptual engineering. Ethics and Information Technology, 25(1). https://doi.org/10.1007/s10676-022-09675-6

Vries, D. M. M. (2009). Translating Customer Requirements into Technical Specifications. In Elsevier eBooks (pp. 489–512). https://doi.org/10.1016/b978-0-444-51667-1.50022-7

Wan, M., Zha, D., Liu, N., & Zou, N. (2021). Modeling Techniques for Machine Learning Fairness: A survey. arXiv (Cornell University). https://doi.org/10.48550/arxiv.2111.03015

Wang, F. Y. (2020). Cooperative Data Privacy: The Japanese Model of Data Privacy and the EU-Japan GDPR Adequacy Agreement. Harvard Journal of Law & Technology, 33(2), 661–690. https://jolt.law.harvard.edu/assets/articlePDFs/v33/33HarvJLTech661.pdf

Weber, R. L. (1990). Basic content analysis. In SAGE Publications, Inc. eBooks.
https://doi.org/10.4135/9781412983488

Williamson, O. E. (2000). The new institutional economics: taking stock, looking ahead. Journal of Economic Literature, 38(3), 595–613. https://doi.org/10.1257/jel.38.3.595

Wirtz, B. W., Weyerer, J. C., & Sturm, B. (2020). The Dark Sides of Artificial Intelligence: An Integrated AI Governance Framework for Public Administration. International Journal of Public Administration, 43(9), 818–829. https://doi.org/10.1080/01900692.2020.1749851

Wolters, A. (2022). Guiding the specification of sociotechnical Machine Learning systems: Addressing vulnerabilities and challenges in Machine Learning practice. In Repository TU Delft.
https://doi.org/10.4121/19793968.v1

Yi, L., Duan, X., Zhao, C., & Da Xu, L. (2012). Systems Science: Methodological Approaches. https://digitalcommons.odu.edu/itdsbooks/4/

Zednik, C. (2019). Solving the Black Box Problem: a normative framework for explainable artificial intelligence. Philosophy & Technology, 34(2), 265–288. https://doi.org/10.1007/s13347-019-00382-7

A

List of selected literature for integrative literature review

The following page presents the selected literature for the integrative literature review.

Title	Reference	Theme
Designing Socio-Technical Systems	Bauer and Herder (2009)	Socio-technical systems
Harmonizing artificial intelligence for social good	Berberich et al. (2020)	Japanese society & AI
Vulnerabilities of connectionist AI Applications: Evaluation and defense	Berghoff et al. (2020)	AI values
From EU Robotics and AI Governance to HRI Research: Implementing the ethics Narrative	De Pagter (2023)	AI ethics & values
Governing AI safety through independent audits	Falco et al. (2021)	AI ethics & values
Treating socio-technical systems as engineering systems: some conceptual problems	Kroes et al. (2006)	Socio-technical systems
Emotional AI, ethics, and Japanese spice: contributing community, wholeness, sincerity, and heart	McStay (2021)	Japanese society & AI
A neo-communitarian approach on human rights as a cosmopolitan imperative in East Asia	Morita (2021)	Japanese society & AI
Japanese traditional vocational ethics: Relevance and meaning for the ICT-dependent society	Murata (2019)	Japanese society & AI
Data is what data does: regulating use, harm, and risk instead of sensitive data	Solove (2023)	AI values
Assessing the ethical and social concerns of artificial intelligence in neuroinformatics research: an empirical test of the European Union Assessment List for Trustworthy AI (ALTAI)	Stahl and Leach (2022)	AI values
Governance of artificial intelligence	Taeihagh (2021)	AI values
Handbook of ethics, values, and technological design: sources, theory, values and application domains.	Van den Hoven et al. (2015)	AI ethics & values

Table A.1: Selected literature list to identify societal values.

B

Translated LIAJ code of conduct

On the subsequent page, the translated LIAJ code of conduct is presented. The yellow mark indicates the elements that have been introduced. The text that is marked out has been eliminated.

1. **Encouraging appropriate response to customers throughout the stages from proposal and provision of products to claims payment.**
   To establish customer satisfaction and trust, life insurers shall provide quality products which meet customer's needs, render services from the customer's point of view, and award appropriate insurance payouts.

1.1 Life insurers should adequately recognize customer's needs and strive to develop and provide high quality products which can surely deliver the "sense of security" to customers. This focuses on providing customer safety.

1.2 Life insurers should develop and publish a solicitation policy as well as taking measures to ensure an appropriate solicitation. Also, life insurers should make an appropriate and sufficient explanation to enable customers to accurately understand details of products in order to select the best product.

1.3 Life insurers should provide information related to contract details and each procedure to customers throughout the stages from the conclusion of a contract to claims and payment of insurance money and benefits on a timely basis and in a comprehensible manner.

1.4 Life insurers should recognize that the payment of insurance money and benefits is the most important and fundamental function in the life insurance business and provide the service in a quick, accurate, fair, and careful manner. Life insurers should make a sufficient explanation in order to gain understanding and satisfaction from customers where they cannot pay out the benefits.

1.5 Life insurers should cultivate their employees who can render appropriate services from the customer's point of view throughout the stages from provision of contracts to claims payment.

2. **Promoting mutual understanding with customers and society**
   Life Insures shall provide a wide variety of stakeholders and society with information related to business activities in an accurate and proactive manner, listen extensively to customer input, respond to it sincerely, and reflect it in the management.

2.1 Life insurers should provide customers and society with proactive and accurate information related to financial conditions and business activities such as listening extensively to customer input in an accurate and proactive manner in order to enable customers to accurately understand the life insurance business.

2.2 In light of changes in the social environment, actively provide information that contributes to the improvement of financial literacy. Life insurance should strive to contribute to the stability and improvement of the life of the association.

2.3 Life insurers should listen extensively to stakeholders' input and respond to the opinions and requests sincerely in order to improve their operations, products and services.

3. **Handling properly and protecting thoroughly the customer information**
Life insurers shall recognize the materiality (~~highly confidential and significant nature~~) of information received from customers through life insurance business and strive to properly handle and thoroughly protect such information.

3.1 Life insurers should recognize that they handle the important personal information related to customer's life, corporeity and property and ensure its appropriate treatment and protection in order to enable customers to provide the information with a sense of security.

3.2 Life insurers should also recognize the importance of information of legal entities and organizations obtained through each transaction and ensure its appropriate treatment and protection.

3.3 Life insurers should appropriately treat the personal information based on the Personal Information Protection Law, the Personal Information Protection Commission, the guidelines set by the Financial Service Agency (FSA) and the LIAJ as well as laws and provisions including guidelines.

4. ~~Promoting compliance~~ **Ensuring fair business activities**
In order to establish the firm trust of customers and society, life insurers shall conduct fair and impartial business activities and confirm with the norms of society, including all relevant laws and regulations.

4.1 Life insures should comply with relevant laws for the protection of life insurance policyholders and consumers as well as social norms in order to conduct fair business activities.

4.2 Life insurers should comply with the Anti-Monopoly Act and compete fairly and freely in order to promote the protection of general customer's ~~benefits~~ profit (not monetary) and sound market development.

4.3 In international business activities, life insurers should comply with international rules and laws as well as respecting a local culture. Also, they should pay attention to any of its impacts on a local society and economy.

4.4 Life insurers should conduct fair business activities by strict adherence to compliance and implementation of effective governance.

5. **Blocking off the relationship with anti-social forces**
Life insurers shall completely block off the relationship with any anti-social forces that could jeopardize public order and safety.

5.1 In order to block off the relationship with anti-social forces, life insurers should take appropriate measures as an organization, for example, resolutely refusing unreasonable demands from anti-social forces with the cooperation of external professional organization.

5.2 In order to prevent terrorism financing and money laundering, life insurers should take appropriate actions to confirm identification of customers and report suspicious specific ~~activities~~ transactions.

6. **Engaging in safe and profitable asset management with due consideration for its social nature**

Life insurers shall engage in asset management seeking to ensure safety, profitability, and liquidity taking into consideration its social and public nature.

6.1 Life insurers should engage in asset management seeking to ensure safety, profitability, and liquidity in order to comply with the mandate of customers.

6.2 Given the public nature of life insurance business, life insurers should engage in asset management with due consideration for its social public nature.

6.3 As a major participant in both financial and capital markets at home and abroad, life insurers should engage in asset management taking its impacts on each market and economy into consideration.

6.4 In order to contribute to the resolution of social issues and achieve a sustainable society, we strive to manage assets while considering environmental, social, and governance (ESG) factors.

6.5 As a responsible investor, we strive to fulfill our stewardship responsibilities by engaging in purposeful dialogue with investee companies, with the aim of achieving sustainable growth for those companies.

7. **Promoting efforts to address environmental issues.**

Life Insurers shall address environmental issues voluntarily and proactively based on the recognition that addressing the issues are important tasks to be undertaken commonly by all humankind.

7.1 Life insures should address environmental issues voluntarily and proactively by promoting savings of energy and resources in business activities.

7.2 Life insurers should enhance employee's awareness of environmental issues through environmental education and support them to participate in environmental conservation activities.

8. **Promoting social service activities**

To achieve a sound and sustainable development of society which serves as the basic infrastructure of life insurance business activities, life insurers shall actively take part in social service activities or programs as a good corporate citizen.

8.1 Life insurers should become aware that they are a member of a local community and actively take part in social service activities or programs as a good corporate citizen to achieve a

sound and sustainable development of society in order to create a society that is rich and filled with a sense of security.

8.2 Life insurance should contribute to resolution of social issues cooperating with non-profit organizations, non-governmental organizations and local communities as well as participating in social service activities as the industry and business community.

9. Respect for all human rights

Respect the human rights of all people and act with consideration for the impact of your activities on human rights.

9.1 Respect the human rights of all people, having understood the internationally recognized human rights.

9.2 Be aware of business actions for the impact of your activities on human rights, not only towards customers but also towards all stakeholders such as business partners.

10. **Respecting the human rights of employees and achieving a vigorous work environment**

Realize a work style that enhances the capabilities of employees and respects their personality, individuality, and diversity, and ensure a healthy and safe workplace environment that is easy to work in.

10.1 Life insures should ensure a fair work environment where there is no discrimination or harassment while respecting the human rights and privacy of employees.

10.2 Life insurers should ensure a vigorous work environment which allows each employee to exercise his/her full abilities, while enhancing abilities of each person through career formation and competence development.

10.3 Given the acceleration of the aging population, life insures should achieve a vigorous work environment by supporting employees who give birth to a child, raise children or care for aging parents and promoting their flexible working style.

10.4 Life insurers should promote employment which helps to pursue social participation of diverse human resources.

10.5 Ensure a comfortable and safe working environment that takes into account employee health and well-being.

11. Strengthening risk management measurers

Life insurers shall strengthen risk management measures under the leadership of executives with proper operation and continuous improvement to be able to meet obligations to customers and establish trustworthiness.

11.1 Life insurers should establish the risk management system in order to recognize and evaluate various risks, and deal with them appropriately under the leadership of executives with an

aim to fulfill obligations to customers, then they should reexamine it to verify if the system functions properly and implement continuous improvement.

11.2 Life insures should strengthen risk management measures responding to each risk such as insurance underwriting risks, asset management risks, operational risks and system risks, to make sure that proper risk management is used.

11.3 Life insurers should establish the risk management system responding to crisis management and large-scale disasters for any situations that the normal risk management system is not enough and prepare a system of appropriate payouts of benefits through smooth transaction of business and establish a system that can reliably make insurance payments and other payments.

12. More effective prevention of recurrence of mismanagement and fulfillment of accountability

In the cases of the events that shall affect customers or society, under a solid leadership of executives, life insurers shall strive to thoroughly determine the causes, prevent a recurrence and fulfill their accountability and responsibility to explain to customers and society.

12.1 Life insurers should establish a company structure to promptly and properly respond to a situation that affects customers or society, for example by amending a protocol or procedures in easy explainable manners.

12.2 Life insurers should conduct fact-finding research and investigate the cause of any situation that affects customers or society under the responsibility of the management. Also, they should deal with it promptly and appropriately as well as putting out efforts to prevent reoccurrence to restore trust. In addition, they should fulfill accountability and responsibility, giving a clear and quick explanation to the customers and the society.

C

Code list: legal and industry documents

Code

Accessibility
Accountability
Accuracy
Adaptability
Continuous improvement
Diversity
Diversity: Discrimination
Diversity: Diversity
Effectiveness
Efficiency
Empowerment
Fairness
Fairness: Dignity
Fairness: Discrimination
Fairness: Discrimination (2)
Fairness: Equality
Fairness: Equity
Fairness: Exclusion
Fairness: Fairness
Fairness: Financial exclusion
Fairness: Inclusion
Fairness: Inequalities
Fairness: Inequality
Fairness: Non-discrimination
Fairness: Population discrimination
Fairness: Procedural fairness
Fairness: Social equity
Fairness: Social inclusion
Fairness: Social inequalities
Fairness: Social justice
Fairness: Vulnerability
Human autonomy
Human oversight
Inclusiveness
Integrity
Learning
Privacy
Privacy: Anonymized personal information
Privacy: Confidentiality
Privacy: Consumer privacy
Privacy: Data Protection
Privacy: Personal data privacy
Profitability
Proportionality
Redundancy
Reliability
Reliability: Creditworthiness
Reliability: Reliable
Responsibility
Responsibility: Compliance
Responsibility: Corporate social responsibili
Responsibility: Environmental responsibility
Responsibility: Financial responsibility
Responsibility: Liability
Responsibility: Responsible
Safety
Safety: Control
Safety: Safe
Safety: Well-being
Security
Security: Availability
Security: Consumer protection
Security: Control
Security: Cybersecurity
Security: Robustness
Stability
Transparency
Transparency: Auditability
Transparency: Clarity
Transparency: Explainability
Transparency: Opacity
Transparency: Transparent
Trustworthy
Trustworthy: Credibility
Trustworthy: Trustworthiness
Understandability

D

Privacy rules

Subject	APPI	EU
Definition of Personal Information	Article 2: presents a comprehensive definition of personal information, encompassing data capable of uniquely identifying an individual, such as their name, date of birth, or other descriptive details.	Article 4: Personal data encompasses any information about a natural person who is identified or can be identified. The definition of the APPI is narrower than this one.
Consent	Article 16: the collection and use of personal information, including for underwriting purposes, necessitate consent.	Article 6 mandates that consent is necessary to process personal data. However, it establishes a more stringent criterion for obtaining valid consent, which entails an unambiguous affirmative action.
Sensitive data	Not specifically mentioned.	Article 9 explicitly acknowledges the existence of distinct categories of personal data, such as health data, and imposes more stringent processing obligations on them.
Data minimization	Article 16: mandates collecting and utilizing personal data solely for specified and legitimate purposes. Data collection and utilization in insurance underwriting are restricted to underwriting and associated activities.	Article 5: is like the APPI.
Data subject rights	Article 25: Right to access Article 26: Right to correction Article 27: Right to suspension of use Articles 26 and 27: Right to deletion Article 76: Right to Complain to the Personal Information Protection Commission	Articles 13 and 14: Right to be informed Article 15: Right of access Article 16: Right to rectification Article 17: Right to Erasure Article 18: Right to restrict processing Article 20: Right to object Article 22: Right to Automated Decision-making and Profiling Article 77: Right to complain
Data protection officers (DPO)	Not specifically mentioned. However, the law uses similar terms, such as Personal Information Protection Manager.	Article 37 and 39 specifies that a DPO must be appointed in cases where the processing of personal data is carried out. It also points out the details about the tasks and responsibilities of the DPO. Additionally, the DPO is involved in all data protection matters and is independent.

E

Questionnaire template

Role:

	Shall the focus of the AI Assessment process be on ethics or also on the business impact?						
	✱✱✱	✱✱	✱	No star	Definition	Is this definition different to your understanding?	Explanation of the number of stars.
Accessibility					Refers to designing and developing artificial intelligence systems and technologies that can be easily used and accessed by people.		
Accountability					Refers to the responsibility individuals and organization needs to take for one's actions and decisions, and being answerable to others for the consequences of those actions and decisions.		
Accuracy					The quality or state of being correct or precise.		
Adaptability					Being able to adjust to new conditions.		
Autonomy					Refers to the ability of an AI system to operate independently, without human intervention or control, in order to perform a specific task or achieve a specific goal.		
Diversity					Refers to the presence of differences among individuals or groups, such as differences in race, gender, ethnicity, religion, or cultural background.		
Effectiveness					Refers to the ability of designed systems to achieve its intended objectives and produce meaningful outcomes.		
Efficiency					Refers to the ability of an AI system to perform its intended tasks and functions in a timely and resource-efficient manner.		
Empowerment					Refers to the ethical principle actions and decisions empower individuals and communities to make informed decisions and take meaningful action.		
Fairness					Refers to the ethical principle that decisions and processes are fair and non-discriminatory, and that promote social justice, equality and equity.		
Human autonomy					Refers to the ethical principle of respect and promote the autonomy of individuals, and that do not unduly restrict or manipulate their decision-making or behavior.		
Inclusiveness					Refers to the extent to which individuals feel valued, respected, and included in a group or organization, regardless of their differences.		
Integrity					Integrity is regarded as the honesty and truthfulness or accuracy of one's actions.		
Learning					The acquisition of knowledge or skills through study, experience, or being .		
Privacy					Someone's right to keep their p ersonal matters secret.		
Profitability					Refers to the ability of AI systems and related businesses to generate profits and revenue streams.		
Proportionality					Related actions should be proportional to the intended objectives and goals, and should not result in unnecessary harm or negative consequences.		
Redundancy					Refers to the practice of building backup systems and processes into AI systems, in order to ensure that the system can continue to function even in the event of a failure or disruption.		
Reliability					The ability of AI systems to consistently produce accurate and trustworthy results, and to function as intended in a variety of different contexts and scenarios.		
Responsibility					Refers to the obligation of individuals, organizations, and governments.		
Safety					The condition of being protected from or unlikely to cause danger, risk, or injury.		
Security					The state of being free from danger or threat.		
Stability					To operate reliably and consistently over time, and to avoid unexpected or unstable behavior.		
Transparency					The characteristic of being easy to see through.		
Trustworthy					Refers to the ability to operate in a way that is reliable, ethical, and consistent with the expectations of users and society.		
Understandability					Refers to the ability of AI systems to present information and decision-making processes in a way that can be easily understood by users and stakeholders.		
Continued improvement					Refers to the ongoing efforts to improve the performance, reliability, and ethical standards of AI systems over time.		

Missing values: *Missing values....*

F

Business view value prioritization

Values chosen by business	**Societal values**	**Participants's comment**	Legal documents
Accountability	x	Clear accountability is needed to ensure that governance and control mechanisms are properly implemented and effective. Additionally, the ability to justify decisions to stakeholders is fundamental to maintaining trust and integrity within the system's framework.	(Code of Conduct, 2018, p. 7)
Understandability	Explainability	Participants emphasize the connection to transparency and the need for understandable processes in AI's delivery of outcomes.	(EIOPA Publishes Report on Artificial Intelligence Governance Principles, 2021, p. 8)
Reliability	Robustness	We must show that the outcomes are reliable.	
Accessibility	Usability and security	A balanced approach towards the accessibility and usability of certain systems or technologies. It advocates for making these systems user-friendly yet proposes restricting access to individuals with certain knowledge or expertise.	(Ministry of Justice, 2003, p. 13)
Accuracy	Robustness	System outcomes must be demonstrated and correct.	(EIOPA Publishes Report on Artificial Intelligence Governance Principles, 2021, p. 52)
Responsibility	x		(Code of Conduct, 2018, p. 6)
Transparency	Explainability	Understanding the model's process and decision-making is important to enhance explainability.	
Inclusiveness	Fairness	Same as fairness.	
Stability	Robustness	Is connected to reliability.	(EIOPA Publishes Report on Artificial Intelligence Governance Principles, 2021, p. 64)
Effectiveness	x	Is considered very important as it evaluates the desired outcomes of the system.	

Table F.1: Values from business view connected to the social values with argumentation from legal documents and participants

G

Informed consent templates

Participants information/Opening statement for interviews

You are being invited to participate in a research study titled Design guidelines to protect stakeholders' values in AI systems - Based on a use case in the Japanese Life Insurance Industry. This study is being done by Shan Amin from the TU Delft.

This research study's purpose is to design a value framework for the Japanese insurance industry to assess AI models responsibly, and it will take approximately 60 minutes to complete. The data will support defining values, norms, and guidelines to assess AI and the use AI. We will be asking you to participate in the interviews, where questions will be asked about the topic.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by anonymizing any transcribed workshop or focus group; your information will not be re-traceable. Names, e-mails, and anonymized pseudo-transcribed interviews collected during the study will stored in a TUD approved storage solution. Only aggregated conclusions will be made publicly available with the MSc thesis. The thesis will be reviewed by the organization before publication to ensure that no confidential information is published. In the thesis, you will be referred to by a generalized job title. Signature on the consent form and pseudo-anonymous transcripts will be accessible only to the TUD research team and will be deleted at the latest two years after the end of the project. The recordings will be deleted as soon as the transcription is complete.

Your participation in this study is entirely voluntary, and you can withdraw at any time. You are free to omit any questions.

Corresponding researcher
Shan Amin

Responsible researcher
Sander Renes

Figure G.1: Informed consent template for interviews.

Participants information/Opening statement for questionnaires

You are being invited to participate in a research study titled Design guidelines to protect stakeholders' values in AI systems - Based on a use case in the Japanese Life Insurance Industry. This study is being done by Shan Amin from the TU Delft.

This research study's purpose is to design a value framework for the Japanese insurance industry to assess AI models responsibly, and it will take approximately 60 minutes to complete. The data will support defining values, norms, and guidelines to assess AI and the use AI. We will be asking you to participate in the questionnaires, where questions will be asked about the topic.

As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by anonymizing any transcribed workshop or focus group; your information will not be re-traceable. Names, e-mails, and anonymized pseudo-transcribed interviews collected during the study will stored in a TUD approved storage solution. Only aggregated conclusions will be made publicly available with the MSc thesis. The thesis will be reviewed by the organization before publication to ensure that no confidential information is published. In the thesis, you will be referred to by a generalized job title. Signature on the consent form and pseudo-anonymous transcripts will be accessible only to the TUD research team and will be deleted at the latest two years after the end of the project. The recordings will be deleted as soon as the transcription is complete.

Your participation in this study is entirely voluntary, and you can withdraw at any time. You are free to omit any questions.

Corresponding researcher
Shan Amin

Responsible researcher
Sander Renes

Figure G.2: Informed consent template for questionnaires.

Participants information/Opening statement for workshop

You are being invited to participate in a research study titled Design guidelines to protect stakeholders' values in AI systems - Based on a use case in the Japanese Life Insurance Industry. This study is being done by Shan Amin from the TU Delft.

This research study's purpose is to design a value framework for the Japanese insurance industry to assess AI models responsibly, and it will take approximately 60 minutes to complete. The data will support defining values, norms, and guidelines to assess AI and the use AI. We will be asking you to participate in the workshop, where questions will be asked about the topic.

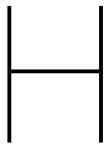
As with any online activity, the risk of a breach is always possible. To the best of our ability, your answers in this study will remain confidential. We will minimize any risks by anonymizing any transcribed workshop or focus group; your information will not be re-traceable. Names, e-mails, and anonymized pseudo-transcribed interviews collected during the study will stored in a TUD approved storage solution. Only aggregated conclusions will be made publicly available with the MSc thesis. The thesis will be reviewed by the organization before publication to ensure that no confidential information is published. In the thesis, you will be referred to by a generalized job title. Signature on the consent form and pseudo-anonymous transcripts will be accessible only to the TUD research team and will be deleted at the latest two years after the end of the project. The recordings will be deleted as soon as the transcription is complete.

Your participation in this study is entirely voluntary, and you can withdraw at any time. You are free to omit any questions.

Corresponding researcher
Shan Amin

Responsible researcher
Sander Renes

Figure G.3: Informed consent template for workshops.



Workshop template

	Values	Risk could be...	Rule: The AI System should not ....
Business strategy, need and values			
Business purpose: Supporting a more efficient sales process with a competitive advantage and a more accurate underwriting process.			
Gathering data: financial, medical and personal data from external sources and internal sources.			
Exploring data quality and the relationship between data variables.			
Cleaning data to improve the quality and accuracy of the data.			
Model development through AI software.			
Evaluation performance testing accuracy, sensitivity and specificity.			
Model outcomes with a decision on whether or not to accept the risk per customer.			
Assessment of eligibility through **Zan-s** application.			
Create list with eligible customers after assessment.			
Provide list with eligible customers for underwriting to agents.			

General AI lifecycle

Exploring the 'AI life cycle' offers valuable insights through which organizations can deepen their understanding and awareness of the risks tied to complex socio-technical systems (Azzutti et al., 2022). Understanding the AI life cycle allows stakeholders to grasp the entire journey of an AI solution, from conceptualization to real-world deployment. This approach also fits the informal institution of 'wholeness,' where we try to understand the whole system. Furthermore, every phase of the AI life cycle comes with its own set of challenges and risks. Organizations can identify, mitigate, and systematically manage risks by exploring the life cycle.

The AI life cycle consists of three phases: design, development, and deployment, supported by academic literature and practical experience. Figure I.1 illustrates the AI cycle, derived from synthesizing various scholarly sources and use case practices. The design phase encompasses idea generation and exploration of potential impacts stemming from a problem. The development stage signifies the implementation of the previous stage. During the deployment phase, the primary objective is to standardize and improve the accessibility of the service and solution for all stakeholders and end-users.

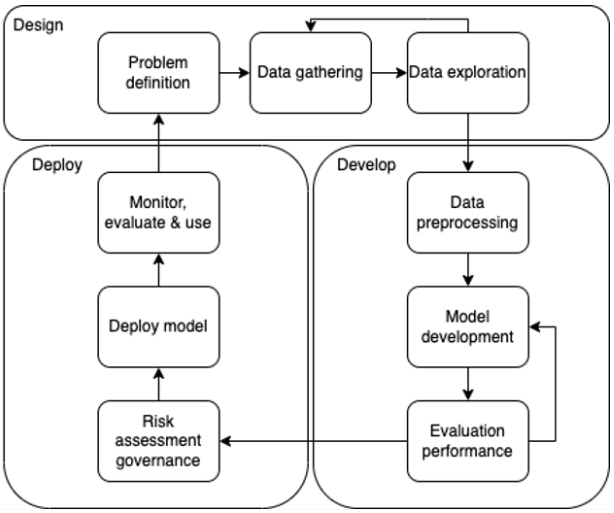


Figure I.1: General AI lifecycle

Different phases require specific human expertise, such as the roles of AI/data scientist for design, AI/ML scientist for development, and AI/ML engineer for deployment (De Silva & Alahakoon, 2022). De Silva and Alahakoon (2022) emphasize enhancing AI system design through secondary enabling roles that contribute to value, inclusivity, and quality. The roles encompass an ethics committee, a project manager, a pool of domain experts, legal counsel specializing in AI and data, and a steering and advisory committee with comprehensive oversight. The subsequent paragraphs delineate each phase, highlighting the stakeholders involved. These insights support the guidance of the value framework in a later phase of the research.

I.0.1. Design phase

The design phase begins by identifying and defining the problem that requires a solution. According to Haakman et al. (2022), two primary approaches, namely the ”Innovation push” and the ”Technology push,” are observed in practice. The concept of ”innovation push” refers to the situation where stakeholders bring up questions or issues that need to be addressed, resulting in the formation of a team to create a solution using suitable Machine Learning techniques. On the other hand, the ”Technology push” approach involves teams discovering new data sources or Machine Learning techniques to improve business value and address organizational challenges. This approach aims to optimize processes, minimize manual labor, enhance model performance, and generate innovative business prospects. The use case arises from a technology-driven approach, wherein developers conceptualized the AI system in collaboration with the business stakeholders.

During the design phase, stakeholders' collaboration is paramount to define the problem to solve (De Silva & Alahakoon, 2022). Teams closely collaborate with stakeholders to evaluate the suitability of employing an AI system to address the identified problem. This collaborative effort supports the participants in completing a project document that outlines crucial details such as the problem statement, goals, and the corresponding business case. In addition, domain experts contribute their expertise to the teams, ensuring a comprehensive understanding of the problem and the potential ethical and business risks.

I.0.2. Development phase

Once the data is collected and deemed representative of the problem, it is prepared for modeling (Haakman et al., 2021). Developing an AI model begins by identifying a suitable AI algorithm that accurately represents the required AI capabilities for the intended application. Researchers have categorized all existing practical applications into one or a combination of four capabilities: prediction, classification, association, and optimization.

The model learns from the training data set, evaluates its performance on the testing data set, and uses the validation data set to select or tune the model's hyperparameters for improved performance (Alla & Adari, 2021). The training begins by instructing the model to learn and perform its designated task. After completing the training phase, the subsequent step involves either evaluation or validation. Developers assess the model's performance using accuracy, precision and recall metrics during the evaluation step. This evaluation step is accomplished by providing the model with test data. The selection of the validation metric is contingent upon the specific context. This phase is an iterative process that employs various models and fine-tunes the parameters. Model explainability, or explainable AI (XAI), is crucial for transparency during the development phase. Transparency is crucial, particularly for complex models such as gradient boosting or neural networks, as numerous layers of computations conceal the information flow. XAI methods can be broadly classified into intrinsic and extrinsic categories. These methods aid AI scientists in comprehending the contributions of attributes, learning processes, and model parameters to the anticipated outcome of the AI.

I.0.3. Deployment phase

AI/ML engineers receive the AI model after evaluating it through various metrics and testing its explainability. To achieve broader deployment, the task performance of the system must be both effective and computationally efficient. The evaluation considers secondary metrics, including CPU and memory performance, time complexity, ethical considerations, and convergence metrics. Another set of metrics assesses risk factors, including privacy, cybersecurity, trust, robustness, explainability, interpretability, usability, and social implications. Researching topics like adversarial robustness is crucial for effectively mitigating cyber threats. The decision-making process entails choices concerning model compression, data handling, and deployment location.

AI fairness metrics are essential for addressing algorithmic bias and ethical challenges. According to Narayanan (2018), variations in definition exist for these metrics, requiring specialized terminology and techniques for auditing, detecting biases, and implementing mitigation strategies. AI technology providers facilitate the connection between research and practice by providing various tools. These tools include AI Fairness 360 (AIF360), the What-If Tool, OpenAI Gym, and benchmarks for adversarial attacks (Lee & Singh, 2020).

After evaluating the model, it transitions into operational use, known as model serving, model scoring, or

model production. Typically, this operational deployment occurs on a smaller scale than the subsequent description of full-scale ”operationalization.” Crucial deployment considerations include:

- Determining whether the AI model will be used in real-time or batch processing;
- Assessing the number and types of end users and applications;
- Defining expected output formats and;
- Estimating lead time and frequency of usage.

The system will be monitored and assessed during this final phase of the life cycle. The evaluation criteria include the representation of the technology, its utilization by diverse individuals in diverse settings, and the value generated. The technology is evaluated through developers’ model drift and model staleness assessment. The criterion of people is assessed based on end-user activity, whereas value generation is quantified by return on investment (ROI).

J

Interview questions

Interviewee 1: end-user
17/10/2023 - *Norms*

• How is accountability and responsibility arranged in the manual?
• How are accountability and responsibility arranged in the PU model?
• How is accuracy measured in the manual underwriting process?
• How often is the data updated in the PU model?
• How often is it enough to update the data in the context of underwriting?
• What kind of human control is adapted in the PU model?
• Do you have the ability to change the output of the system?
• How do you use the output of the model?
• Are you involved in the design, development, or deployment of the system?
• Do you trust fully on this output, or do you do your own controls before using the output?
• How do we measure the effectiveness of the manual underwriting process?
• How do we measure the effectiveness of the PU model?
• Is the PU model accepted by our agencies?
• What different groups can be categorized from our PU model?
• We are using financial and medical data. Can you guide me through the process of how you handle these decisions and how you come to specific business rules?
• How do make sure that we are using the financial and medical data legally because we use sensitive data?
• Any idea how we attain privacy in the manual underwriting process and PU model?
• How is the reliability of the manual underwriting process measured? • How is the reliability of the PU model measured?
• How is security maintained in PU and manual processes?
• On what aspects do you expect transparency? Design, development, output?
• When is something trustworthy regarding your definitions?
• How do we know that our data is equally distributed?
• Would you use the system if the input and output of the models are explainable to some extent, but the AI is a black box?
• When is someone in the manual process classified as a high risk?
• To what extent do we rely on the model's outcome?
• How do we measure fairness in the manual process?

Interviewee 2: IRM
17/10/2023 - *Norms*

• How do we control privacy?
• How do control the following guidelines from the APPI:
o Article 25: Right to access
o Article 26: Right to correction
o Article 27: Right to suspension of use
o Article 26 and 27: Right to deletion
o Article 76: Right to Complain to the Personal Information Protection
• How is the accuracy of the data being controlled in the current state? Can we use this mechanism for AI systems?
• Where within the IRM processes is human control necessary with respect to the PU model/system?
• Can we use the standards/procedures of the Business Impact Assessment to measure the integrity of our data for the AI systems?
• What safety concerns do arise within the IRM field regarding AI?
• What security mechanisms are used to protect data from outside threats?
• What security mechanisms are used to protect data from system errors? • What are other security concerns regarding the PU model?
• How do we control data against manipulation and control of information provision, such as unfair customer targeting, and biased customer segmentation?
• How do we control biased data, such as filter bubbles (diversity in the data set)?
• How do we make sure we have consent for using customers' information for new models such as the PU model?
• How do control that the business is only using needed personal information for their AI design, development, deployment, and use?
• Where do you need transparency within the topics of privacy and security?

Interviewees 3 & 4: ORM manager and analyst
17/10/2023 - *Norms*

• How do we manage to achieve the goals of the management?
• How do we control accessibility and on what criteria do we decide who may access?
• How do we control security in terms of cyber threats, data leakage, and misuse of sensitive data?
• How do we control the integrity and quality of the data?

Interviewees 5 & 6: Data scientist and manager 17/10/2023 - *Norms*

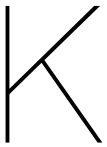
• Accountability: What must be documented according to accountability?
• Accountability: How do we determine accountability?
• Accuracy: data must be accurate and the decisions themselves, but how do we control and manage this in the current state?
• Accuracy: data must be accurate and the decisions themselves, but how do we control and manage this in the future state?
• Effectiveness: do we already have certain tools to measure the effectiveness of projects?
• Effectiveness: do we document the effectiveness of certain projects in a way?
• Effectiveness: what measurement/metrics do we need to take for the effectiveness of the PU model?
• Fairness: do we measure fairness in a certain way in the current state?
• Fairness: do we need to make a standard measurement for fairness?
• Fairness: how do we make fairness measurable?
• Fairness: how do we improve fairness or understand fairness in a continuous way?
• Privacy: Are there any privacy issues you are concerned about?
• Privacy: how do we document privacy in the current state?
• Reliability: how do we measure reliability for the PU model? What metric?
• Security: current state
• Security: future state
• Transparency: how to document and control transparency?
• Transparency: who is accountable and responsible for transparency?
• Trustworthy: when is something trustworthy from your expert role perspective?
• Trustworthy: how do we ensure trust within our systems right now?
• Trustworthy: how do we want to ensure trust within our systems for AI in PU?

Interviewee 7: D&AI manager
17/10/2023 - *Norms*

• Understandability: How do we manage in the current state of transparency to understand our projects/processes within the business?
• Understandability: How do we manage in the current state the loss of predictability and control of our business processes?
• Usability: AI may use certain knowledge or expertise: how do we control the use of AI by educated employees in this field? How do we make sure that only educated people are working with the AI systems?
• Usability: How do we control the misuse of AI in the current state?
• Accountability/Responsibility: How do we handle the replacement of the human workforce?
• Transparency/Effectiveness: How do we measure financial feasibility in the PU model?
• Continued improvement: How do we handle organizational resistance to data sharing?
• Continued improvement: What must be considered for the AI strategy of the PU use case?
• Accountability: How to control accountability and responsibility within the design, development, deployment, and use of AI? What control mechanisms does already exist?
• Usability/Fairness: How do we translate the perception of human values in AI to the stakeholders?
• Fairness: How should we reflect on fairness?
• Robustness: Must standard statistical tests be performed for every AI design project?
• Trust: How do we control the negative impact on people's self-worth?
• Fairness: How do we control human vs. machine value judgment?

Interviewee 8: FERM manager
17/10/2023- *Norms*

• What are the main risks that we need to control from the FERM perspective?
• Is the methodology from the PU system clear? So no, what do we need to explain more?



Value framework

Value	Norm	Norm type	Assessment type	law	interview	Standardized control
Trust	*Utilize AI systems to meet customer needs, elevate service quality, and provide clear and sufficient explanations about AI-driven decisions, emphasizing the system's role in enhancing customer understanding and satisfaction.*	*Process*		LIAJ		
Trust	*Commit to transparency throughout the AI development process, from objectives to outcomes, ensuring that all aspects of AI are clear, comprehensible, and integrate seamlessly into user workflows, thus supporting the system's benevolence and clarity for all stakeholders.*	*Process*			x	x
Trust	*Design AI systems to support human autonomy and augment employee roles.*	*Process*			D&AI manager	
Transparency	*Maintain transparency about the AI's objectives, ensuring its operations are benevolent and clear to all stakeholders.*	*Process*			End-user	
Transparency	*Verify that the AI development process and its results are transparent and understandable for all stakeholders, ensuring effective integration into user workflows.*	*Assessment*	*AI governance*		End-user	
Transparency	*Maintain a commitment to transparency and integrity throughout the AI lifecycle on the data and system level, with comprehensive documentation of the AI system's performance and security metrics.*	*Process*		FSA		x
Transparency	*Establish and communicate clear accountability and responsibility for the AI's development and operational processes within the organization.*	*Process*		FSA	Everyone	x
Transparency	*Engage independent audits to affirm the AI system's transparency and user comprehensibility, ensuring consistent trustworthiness.*	*Assessment*	*AI governance*	FSA	FERM, End-user and ORM	x
Transparency	*Implement routine expert evaluations to maintain and verify the AI system's explainability and performance.*	*Assessment*	*AI governance*		End-user	x
Robustness	*Monitor and verify the AI system's quality, reliability, and accuracy throughout its development stages, ensuring robustness aligned with the socio-technical environment.*	*Assessment*	*AI governance*		FERM, End-user, ORM, D&AI	x
Robustness	*Diligently track and enhance the AI system's performance, promptly rectifying errors or inconsistencies.*	*Process*			ORM	x
Robustness	*Develop and implement a risk mitigation strategy, incorporating multiple expert analyses to validate the AI system's reliability.*	*Process*			ORM	x
Robustness	*Implement ongoing assessments and monitoring of the AI system's data quality, integrity, availability, accessibility, and confidentiality.*	*Assessment*	*Data governance*	APPI	IRM	x
Robustness	*Integrate human oversight within the AI system to guide and verify model outcomes, ensuring AI complements rather than overrides human decisions.*	*Process*			D&AI manager	
Robustness	*Establish a certification protocol for data access that safeguards data integrity across all user interactions with the system.*	*Assessment*	*Data governance*	APPI	D&AI manager	x
Robustness	*Mandate thorough validation of AI development methodologies by data scientists prior to system deployment to ensure methodological soundness.*	*Process*			FERM & Data scientist	x
Robustness	*Perform regular audits to assess the effectiveness of the structured policy framework in managing data, ensuring it consistently improves data quality and accessibility.*	*Assessment*	*Data governance*		D&AI manager	x
Robustness	*Conduct scheduled evaluations to verify that the AI system and its data inputs are regularly updated, ensuring the system's accuracy and reliability are upheld.*	*Assessment*	*AI governance*		Data scientist	x
Usability	*Foster collaborative processes across diverse teams for designing, developing, and operationalizing AI systems, promoting an integrated approach to risk and information management.*	*Process*			End-user	
Usability	*Provide comprehensive training for employees, particularly those interfacing with the AI system, to deepen their understanding and effective usage of AI tools.*	*Assessment*	*Data governance*	SP Japan	D&AI manager	x
Usability	*Guarantee transparency in AI operations, ensuring end-users clearly understand the system's objectives and its implications for their roles.*	*Process*			End-user	
Usability	*Conduct regular assessments of the AI system to gauge its usability and impact, using feedback to drive continuous improvement.*	*Process*			FERM	
Usability	*AI systems should augment human capabilities, enhancing productivity and decision-making without supplanting human autonomy.*	*Process*			D&AI manager	
Usability	*Development and implementation of AI require interdisciplinary expertise, integrating diverse business and technical insights.*	*Process*			Data scientist	
Privacy	*Ensure that the use and documentation of customer and company data are purpose-driven and transparent, with access rights and usage intent clearly defined.*	*Assessment*	*Data governance*	APPI		x
Privacy	*Data use for business operations must be contingent upon explicit customer consent, ensuring ethical data practices.*	*Assessment*	*Data governance*	APPI		x
Privacy	*Where feasible, personal data should be anonymized to protect customer privacy.*	*Assessment*	*Data governance*	APPI		
Privacy	*Activities related to customer and organization data should be preempted with impact assessments to mitigate risks and safeguard customer interests.*	*Assessment*	*Data governance*		ORM & IRM	x
Privacy	*Access to personal data should be role-specific within the organization to ensure privacy and relevance of data processing.*	*Assessment*	*Data governance*		ORM & IRM	x
Privacy	*Customers retain rights over their data, including access, rectification, suspension, and deletion.*	*Assessment*	*Data governance*	APPI		x
Privacy	*AI should assist, not replace, human judgment, ensuring responsible use in decision-making processes.*	*Process*		APPI		

Value	Norm	Norm type	Assessment type	law	interview	Standardized control
Fairness	*Develop AI systems that avoid excessive bias towards specific stakeholders in wealth and society.*	*Assessment*	*AI governance*	SP Japan		
Fairness	*Prevent unfair competition and excessive data collection by dominant companies in AI.*	*Assessment*	*AI governance*	LIAJ & SP		
Fairness	*Incorporate only those biases into AI system outcomes based on transparent and verifiable objective criteria, subject to regular review and consensus among stakeholders.*	*Assessment*	*AI governance*	APPI, SP Japan	End-user	
Fairness	*Adopt a product selection that is transparently designed and executed based on verifiable medical data and expert consensus, avoiding reliance on subjective judgment.*	*Assessment*	*AI governance*		End-user	
Fairness	*Establish pricing strategies that are demonstrably fair and inclusive, designed to reflect the equitable application of AI assessments and market standards, ensuring affordability and accessibility for all customer segments.*	*Assessment*	*AI governance*	LIAJ & FSA		x
Security	*Conduct regular security assessments to ensure the integrity and confidentiality of all personal and sensitive data within the AI system, evaluating the effectiveness of safeguards, monitoring protocols, and secure connections to thwart unauthorized access and data breaches.*	*Assessment*	*Data governance*	FSA	ORM	x
Security	*Ensure the AI system's stability and responsiveness by utilizing robust, validated infrastructure and implementing a thorough testing and approval process for any system changes to safeguard against disruptions and maintain operational integrity.*	*Assessment*	*Data governance*	FSA	ORM	x
Security	*Schedule systematic evaluations of the access control and activity monitoring framework to confirm its comprehensiveness in managing user entitlements and activities, thereby safeguarding the AI system against improper use and potential misuse.*	*Assessment*	*Data governance*	FSA	ORM	x
Security	*Adopt a proactive risk management stance by evaluating the potential impacts of security incidents, deploying timely countermeasures against cyber threats, and enforcing strict security standards for all third-party vendors.*	*Assessment*	*Data governance*	FSA	ORM	x
Security	*Obtain the required certification to gain access to the necessary data.*	*Assessment*	*Data governance*	FSA	D&AI manager	x
Security	*Ensure changes are executed, adequately tested, and promoted to production in a controlled and timely manner to prevent service disruption, security breaches, etc.*	*Process*			ORM	x
Accountability	*Form a multidisciplinary team with business stakeholders and developers to ensure effective guidance through the AI lifecycle.*	*Process*			End-user	
Accountability	*Create an independent review panel consisting of legal, compliance, and risk management experts to continuously assess AI systems' technical, social, and regulatory aspects, providing recommendations for improvements and confirming system readiness before launch.*	*Assessment*	*AI governance*		Data scientist	x
Accountability	*Set up a centralized communication channel to assign clear accountability for AI system-related decisions, ensuring that all changes and recommendations are documented and addressed by the responsible parties.*	*Assessment*	*Data governance*		D&AI manager	
Accountability	*Uphold data governance practices, requiring developers to obtain appropriate approvals for the use of customer data, ensuring adherence to privacy standards and security protocols.*	*Assessment*	*Data governance*		D&AI manager	x
Accountability	*All AI systems must undergo a multi-tiered approval process involving data science teams and expert committees to validate the system's readiness for deployment.*	*Assessment*	*Data governance*		Data scientist	x
Effectiveness	*The AI system shall be designed with a clear purpose, supporting and enhancing human decision-making, improving employee efficiency, and ensuring alignment with business objectives.*	*Process*			End-user	
Effectiveness	*Establish continuous feedback mechanisms and dual human judgment reviews to validate the reliability and accuracy of the AI system's outcomes.*	*Assessment*	*AI governance*		End-user	x
Effectiveness	*Promote a collaborative environment that involves stakeholders across disciplines in the development process, ensuring that the AI system remains accurate, up-to-date, and relevant to evolving business and user needs.*	*Process*			End-user	
Effectiveness	*Implement a systematic approach to assess the AI system's performance and impact, optimizing its efficiency and documenting its business implications to guide strategic decisions and process improvements.*	*Process*			FERM	
Continues improvement	*Implement continuous monitoring protocols to evaluate and optimize the AI system's impact on stakeholder trust, service quality, and ethical operations, including privacy, security, and fairness.*	*Process*		FSA & LIAJ	End-user, FERM, ORM, IRM, Data scientist	
Continues improvement	*Foster a collaborative environment for comprehensive risk assessment, leveraging diverse perspectives to understand the full context of the AI system's deployment and operation.*	*Process*		FSA		
Continues improvement	*Engage independent experts to evaluate new AI system designs thoroughly, ensuring all technical, institutional, and procedural standards are met before implementation.*	*Assessment*	*AI governance*		End-user	x

AI governance framework and definition by Mäntymäki et al. (2022)

To strengthen informal institutions within an organization and fit a value framework within the organizational structure, we first must understand the existing governance systems and how AI governance fits within the organizational setup. AI governance includes ethical and legal aspects developed by AI developers for regulatory agencies. In doing so, the literature has multiple definitions of AI governance. Several conceptualizations exist, such as characterizing AI governance as tools and strategies influencing AI development and applications (Minkkinen et al., 2021). However, these definitions mainly focus on macro-level perspectives and do not address how organizations should govern their AI systems.

Mäntymäki et al. (2022) reviewed several academic pieces of literature and defined AI governance from a governance perspective. The definition combines other literature where the technical, ethical, regulatory, organizational, and cultural components occur. This definition is a combination of the previous literature that matches this research with the following concise definition of organizational AI governance:

"AI governance is a system of rules, practices, processes, and technological tools that are employed to ensure an organization's use of AI technologies aligns with the organization's strategies, objectives, and values; fulfills legal requirements; and meets principles of ethical AI followed by the organization (Mäntymäki et al., 2022, p. 604)."

The definition emphasizes the integration of a set of institutions within an organization and the requirements for its practical implementation. From a research perspective, it is imperative to establish a value framework that encompasses a comprehensive system of rules, necessitating the implementation of processes and technical tools for effective control. Furthermore, the definition emphasizes legal, business, and ethical obligations. The absence of a customer/agent focus, also known as the stakeholder focus, is the sole deficiency in this context. This aspect pertains to attaining both understandable and fair outcomes, as previously elucidated in chapter 3.3.3.

Furthermore, the definition connects the ties with the three other areas of governance: corporate governance, IT governance, and data governance. Figure L.1 depicts the relations between an organization's governance areas. The figure shows that AI governance is a subset of corporate and IT governance, partially overlapping with data governance. Corporate governance serves as the overarching structure within an organization. AI systems, being a form of IT systems, fall under the purview of IT governance. While data is integral to AI operations, AI governance extends beyond traditional data governance, encompassing the broader aspects of AI system management. Given the use case, it is imperative to consider the technical governance institutions in the current scenario when designing the value framework.

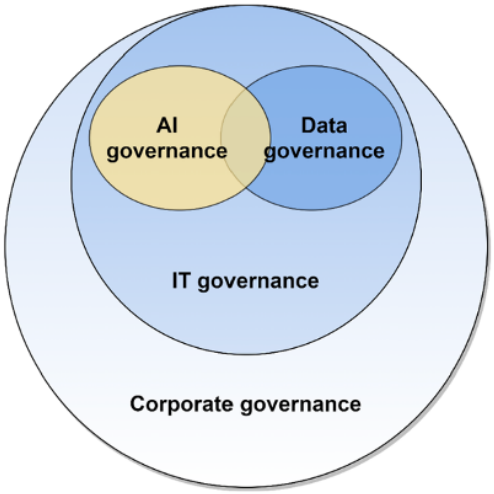


Figure L.1: The governance framework for AI governance adapted by Mäntymäki et al. (2022).

M

Data assessment for PU system

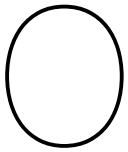
Data governance assessment						
Value	**Norm**	**Norm type**	**Assessment type**	**Ensure a positive impact on accuracy and reliability of the results**	**Ensure a positive impact on customer trust by avoiding unclear, unfair, and inaccurate decisions**	**Ensure no violation of social norms**
Robustness	*Implement ongoing assessments and monitoring of the AI system's data quality, integrity, availability, accessibility, and confidentiality.*	*Assessment*	*Data governance*	x		
Robustness	*Establish a certification protocol for data access that safeguards data integrity across all user interactions with the system.*	*Assessment*	*Data governance*			
Robustness	*Perform regular audits to assess the effectiveness of the structured policy framework in managing data, ensuring it consistently improves data quality and accessibility.*	*Assessment*	*Data governance*	x	x	
Usability	*Provide comprehensive training for employees, particularly those interfacing with the AI system, to deepen their understanding and effective usage of AI tools.*	*Assessment*	*Data governance*			
Privacy	*Ensure that the use and documentation of customer and company data are purpose-driven and transparent, with access rights and usage intent clearly defined.*	*Assessment*	*Data governance*		x	
Privacy	*Data use for business operations must be contingent upon explicit customer consent, ensuring ethical data practices.*	*Assessment*	*Data governance*		x	
Privacy	*Activities related to customer and organization data should be preempted with impact assessments to mitigate risks and safeguard customer interests.*	*Assessment*	*Data governance*		x	
Privacy	*Access to personal data should be role-specific within the organization to ensure privacy and relevance of data processing.*	*Assessment*	*Data governance*		x	
Privacy	*Where feasible, personal data should be anonymized to protect customer privacy.*	*Assessment*	*Data governance*		x	
Privacy	*Customers retain rights over their data, including access, rectification, suspension, and deletion.*	*Assessment*	*Data governance*			
Security	*Conduct regular security assessments to ensure the integrity and confidentiality of all personal and sensitive data within the AI system, evaluating the effectiveness of safeguards, monitoring protocols, and secure connections to thwart unauthorized access and data breaches.*	*Assessment*	*Data governance*	x		
Security	*Ensure the AI system's stability and responsiveness by utilizing robust, validated infrastructure and implementing a thorough testing and approval process for any system changes to safeguard against disruptions and maintain operational integrity.*	*Assessment*	*Data governance*			
Security	*Schedule systematic evaluations of the access control and activity monitoring framework to confirm its comprehensiveness in managing user entitlements and activities, thereby safeguarding the AI system against improper use and potential misuse.*	*Assessment*	*Data governance*			
Security	*Adopt a proactive risk management stance by evaluating the potential impacts of security incidents, deploying timely countermeasures against cyber threats, and enforcing strict security standards for all third-party vendors.*	*Assessment*	*Data governance*	x		
Security	*Obtain the required certification to gain access to the necessary data.*	*Assessment*	*Data governance*			
Accountability	*Set up a centralized communication channel to assign clear accountability for AI system-related decisions, ensuring that all changes and recommendations are documented and addressed by the responsible parties.*	*Assessment*	*Data governance*			
Accountability	*Uphold data governance practices, requiring developers to obtain appropriate approvals for the use of customer data, ensuring adherence to privacy standards and security protocols.*	*Assessment*	*Data governance*	x	x	
Accountability	*All AI systems must undergo a multi-tiered approval process involving data science teams and expert committees to validate the system's readiness for deployment.*	*Assessment*	*Data governance*	x		

N

AI Assessment for PU system

AI GOVERNANCE ASSESSMENT

Value	Norm	Norm type	Assessment type	Ensure a positive impact on accuracy and reliability of the results	Ensure a positive impact on customer trust by avoiding unclear, unfair, and inaccurate decisions	Ensure no violation of social norms
Transparency	*Verify that the AI development process and its results are transparent and understandable for all stakeholders, ensuring effective integration into user workflows.*	*Assessment*	*AI governance*	x	x	
Transparency	*Engage independent audits to affirm the AI system's transparency and user comprehensibility, ensuring consistent trustworthiness.*	*Assessment*	*AI governance*			
Transparency	*Implement routine expert evaluations to maintain and verify the AI system's explainability and performance.*	*Assessment*	*AI governance*	x		
Robustness	*Monitor and verify the AI system's quality, reliability, and accuracy throughout its development stages, ensuring robustness aligned with the socio-technical environment.*	*Assessment*	*AI governance*	x	x	
Robustness	*Conduct scheduled evaluations to verify that the AI system and its data inputs are regularly updated, ensuring the system's accuracy and reliability are upheld.*	*Assessment*	*AI governance*	x	x	
Fairness	*Develop AI systems that avoid excessive bias towards specific stakeholders in wealth and society.*	*Assessment*	*AI governance*			x
Fairness	*Prevent unfair competition and excessive data collection by dominant companies in AI.*	*Assessment*	*AI governance*			
Fairness	*Incorporate only those biases into AI system outcomes based on transparent and verifiable objective criteria, subject to regular review and consensus among stakeholders.*	*Assessment*	*AI governance*	x	x	x
Fairness	*Adopt a product selection that is transparently designed and executed based on verifiable medical data and expert consensus, avoiding reliance on subjective judgment.*	*Assessment*	*AI governance*	x	x	x
Fairness	*Establish pricing strategies that are demonstrably fair and inclusive, designed to reflect the equitable application of AI assessments and market standards, ensuring affordability and accessibility for all customer segments.*	*Assessment*	*AI governance*			
Accountability	*Create an independent review panel consisting of legal, compliance, and risk management experts to continuously assess AI systems' technical, social, and regulatory aspects, providing recommendations for improvements and confirming system readiness before launch.*	*Assessment*	*AI governance*	x		
Effectiveness	*Establish continuous feedback mechanisms and dual human judgment reviews to validate the reliability and accuracy of the AI system's outcomes.*	*Assessment*	*AI governance*	x		
Continues improvement	*Engage independent experts to evaluate new AI system designs thoroughly, ensuring all technical, institutional, and procedural standards are met before implementation.*	*Assessment*	*AI governance*	x		



Process norms

Nr.	Value	Norm	Norm type	Assessment type	law	interview	Standardized control	Design guideline
1	**Trust**	*Utilize AI systems to meet customer needs, elevate service quality, and provide clear and sufficient explanations about AI-driven decisions, emphasizing the system's role in enhancing customer understanding and satisfaction.*	*Process*		LIAJ			**2 & 4**
2	**Trust**	*Commit to transparency throughout the AI development process countinously, from objectives to outcomes, ensuring that all aspects of AI are clear, comprehensible, and integrate seamlessly into user workflows, thus supporting the system's benevolence and clarity for all stakeholders.*	*Process*			x	x	**11**
3	**Trust**	*Design AI systems to support human autonomy and augment employee roles.*	*Process*			D&AI manager		**2**
4	**Transparency**	*Maintain transparency about the AI's objectives, ensuring its operations are benevolent and clear to all stakeholders.*	*Process*			End-user		**2**
5	**Transparency**	*Maintain a commitment to transparency and integrity throughout the AI lifecycle on the data and system level, with comprehensive documentation of the AI system's performance and security metrics.*	*Process*				x	**5&8**
6	**Transparency**	*Establish and communicate clear accountability and responsibility for the AI's development and operational processes within the organization.*	*Process*		FSA	Everyone		**1**
7	**Robustness**	*Diligently track and enhance the AI system's performance, promptly rectifying errors or inconsistencies.*	*Process*			ORM	x	**10**
8	**Robustness**	*Develop and implement a risk mitigation strategy, incorporating multiple expert analyses to validate the AI system's reliability.*	*Process*			ORM	x	**7&8**
9	**Robustness**	*Integrate human oversight within the AI system to guide and verify model outcomes, ensuring AI complements rather than overrides human decisions.*	*Process*			D&AI manager		**2**
10	**Robustness**	*Data scientists must validate AI development methodologies thoroughly prior to system deployment to ensure methodological soundness.*	*Process*			FERM & Data scientist	x	**7&8**
11	**Usability**	*Foster collaborative processes across diverse teams for designing, developing, and operationalizing AI systems, promoting an integrated approach to risk and information management.*	*Process*			End-user		**1, 6, 7, 8, 9, 10**
12	**Usability**	*Guarantee transparency in AI operations, ensuring end-users clearly understand the system's objectives and its implications for their roles.*	*Process*			End-user		***6 & 9***
13	**Usability**	*Conduct regular assessments of the AI system to gauge its usability and impact, using feedback to drive continuous improvement.*	*Process*			FERM		**5, 8, 9, 10, 11**
14	**Usability**	*AI systems should augment human capabilities, enhancing productivity and decision-making without supplanting human autonomy.*	*Process*			D&AI manager		**2**
15	**Usability**	*Development and implementation of AI require interdisciplinary expertise, integrating diverse business and technical insights.*	*Process*			End-user		**6 & 9**
16	**Privacy**	*AI should assist, not replace, human judgment, ensuring responsible use in decision-making processes.*	*Process*		APPI			**2**
17	**Security**	*Ensure changes are executed, adequately tested, and promoted to production in a controlled and timely manner to prevent service disruption, security breaches, etc.*	*Process*	*Data governance*		ORM	x	**5**
18	**Accountability**	*Form a multidisciplinary team with business stakeholders and developers to ensure effective guidance through the AI lifecycle.*	*Process*			End-user		**1**
19	**Effectiveness**	*The AI system shall be designed with a clear purpose, supporting and enhancing human decision-making, improving employee efficiency, and ensuring alignment with business objectives.*	*Process*			End-user		**3**
20	**Effectiveness**	*Promote a collaborative environment that involves stakeholders across disciplines in the development process, ensuring that the AI system remains accurate, up-to-date, and relevant to evolving business and user needs.*	*Process*			End-user		**1, 6, 7, 8, 9, 10**
21	**Effectiveness**	*Implement a systematic approach to assess the AI system's performance and impact, optimizing its efficiency and documenting its business implications to guide strategic decisions and process improvements.*	*Process*			FERM		**10**
22	**Continues improvement**	*Implement continuous monitoring protocols to evaluate and optimize the AI system's impact on stakeholder trust, service quality, and ethical operations, including privacy, security, and fairness.*	*Process*			End-user, FERM, ORM, IRM, Data scientist		**10**
23	**Continues improvement**	*Foster a collaborative environment for comprehensive risk assessment, leveraging diverse perspectives to understand the full context of the AI system's deployment and operation.*	*Process*		FSA			**10&11**

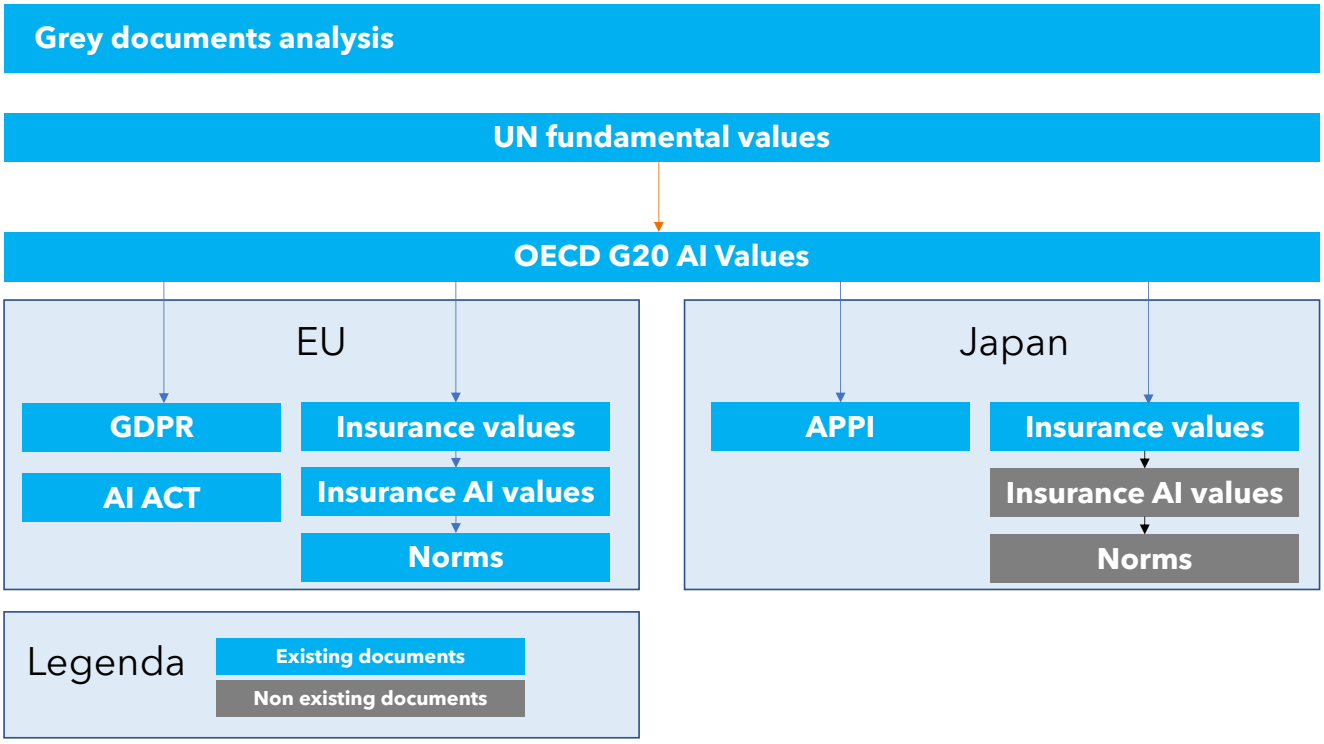
P

Presentation example feedback session with experts

Designing a value framework

Problem statement

The utilization of AI has been demonstrated to enhance decision-making accuracy and optimize operational efficiency within our business processes. However, the potential risks of AI usage remain not completely understood.


Grey documents analysis
UN fundamental values
OECD G20 AI Values
EU
GDPR
AI ACT
Insurance values
Insurance AI values
Norms
Japan
APPI
Insurance values
Insurance AI values
Norms
Legenda
Existing documents
Non existing documents

Future state

Framework

Trust		
Fairness & Non-Discrimination	Transparency, explainability & Understandability	Accountability & responsibility
Privacy	Robustness	Continuous improvement
Usability	Effectiveness	Security

Value framework

Framework

10 values

53 norms

Guide

Example

Fairness: Develop AI systems that avoid excessive bias towards specific stakeholders in walth and society.

- Developers must measure the middle and high socioeconomic classes, the age between 40 and 65, and individuals without disabilities using the predictive parity rate.
- Developers must measure the middle and high socioeconomic classes, the age between 40 and 65, and individuals without disabilities using the false positive rate.
- Developers must monitor both metrics and establish relevant thresholds with experts' and stakeholders' feedback.

Control dashboard

PU AI system

Search

Security

Downtime Causes

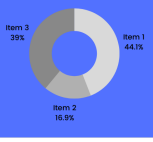
- 25% Broken Machine
- 8% Human Error
- 12% Personal Breaks

Robustness

Accuracy

89%

Reliability



Data quality

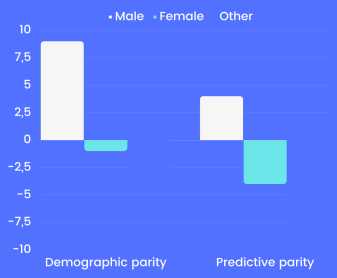
Privacy

Level 3:
Confidential

Accountable

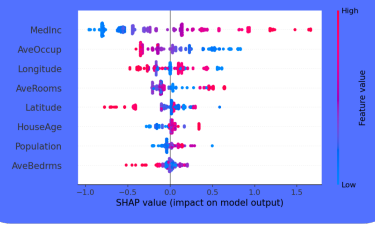
Deve
End-
Busir
L&C:
BSO:
IRM:
ORM:
FERM
Data

Fairness



Predictive parity bias **0.7**

Explainability



Feedback end-user

Agent A & F don't want to work with this model

Design guidelines

D1	Create a **multidisciplinary team** to set up the project and through the AI design cycle.	
D2	Define the business purpose of the AI system and problem statement with the multidisciplinary team and describe it in terms of effectiveness.	
D3	Define the scope with the multidisciplinary team: stakeholders and business process (current situation), impact of the AI system (future situation~~), related laws and industry guidelines.~~	
D4	Assess your design on applicable laws and internal controls	Use the standardized guidelines from the value framework for privacy, trust, accountability & responsibility, and security to identify these specific risks
D5	Identify the potential risks to the AI system that may compromise its societal and business value.	Have a talk with **AI experts** about the filled in assessment **AI experts**: IRM, IT security, Chief Privacy officer, ORM, FERM, D&AI.
D6	Identify stakeholder criteria for the system from each stakeholder view through information sources as the government, business stakeholders and customers.	

D6	Examine the value framework guidelines from the perspective from each stakeholder and select only those that are relevant.	Use the standardized guidelines from the value framework for transparency, fairness and robustness
D7	Formulate the selected guidelines to qualitative and quantitative outputs.	
D8	Create feedback channels for each measurement to the right expert and end-user to verify and validate the measurements.	
D9	Monitor the measurements through qualitative feedback from experts and end-users.	
D10	Create continuous improvement within the AI system by monitoring and improving the systems performance controlled by the end-user.	

Questions?