The Ethical Impact Assessment of Selling Life Insurance to Titanic Passengers

Gizem Gezici[1,2], Chiara Mannari[3,4] and Lorenzo Orlandi[5]

[1]Scuola Normale Superiore, Pisa, Italy
[2]KDD Lab, ISTI-CNR, Pisa, Italy
[3]Institute of Information Science and Technologies "Alessandro Faedo" - ISTI, CNR, Pisa, Italy
[4]Department of computer science, University of Pisa, Italy
[5]Department of Information Engineering and Computer Science, University of Trento, Trento, Italy

Abstract

The Artificial Intelligence Act (AIA) is a uniform legal framework to ensure that AI systems within the European Union (EU) are safe and comply with existing law on fundamental rights and constitutional values. The AIA adopts a risk-based approach with the aim of intending to regulate AI systems, especially categorised as high-risk, which have significant harmful impacts on the health, safety and fundamental rights of persons in the Union. The AIA is founded on the Ethics Guidelines of the High-Level Expert Group for Trustworthy AI, which are grounded in fundamental rights and reflect four ethical imperatives in order to ensure ethical and robust AI. While we acknowledge that ethics is not law, we advocate that the analysis of ethical risks can assist us in complying with laws, thereby facilitating the implementation of the AIA requirements. Thus, we first design an AI-driven Decision Support System for individual risk prediction in the insurance domain (categorised as high-risk by the AIA) based on the Titanic case, which is a popular benchmark dataset in machine learning. We then fulfill an ethical impact assessment of the Titanic case study, relying on the four ethical imperatives of respect for human autonomy, prevention of harm, fairness, and explicability, declared by the High-Level Expert Group for Trustworthy AI. In the context of this ethical impact assessment, we also refer to the questions in the ALTAI checklist. Our discussions regarding the ethical impact assessment in the insurance domain demonstrate that ethical principles can intersect but also create tensions (intriguingly, only in this particular context), for which there is no definitive solution. When tensions arise, which may result in unavoidable trade-offs, these trade-offs should be addressed in a rational and methodical manner, paying special attention to the context of the current case study being evaluated.

1. Introduction

AI applications are ubiquitous, and this brings about some interesting conclusions. Microsoft introduced the Tay AI chatbot in 2016. Tay engaged with Twitter users in "casual and playful conversation." In less than 24 hours, however, Twitter users manipulated the bot to make profoundly sexist and racist remarks. Tay utilised AI to learn from Twitter users' conversations and became "smarter" as it engaged in more conversations. Soon, the bot began repeating incendiary statements from users, such as "Hitler was right," "feminism is cancer," and "9/11

HHAI-WS 2023: Workshops at the Second International Conference on Hybrid Human-Artificial Intelligence (HHAI), June 26—27, 2023, Munich, Germany

gizem.gezici@sns.it (G. Gezici)

was an inside job"[1]. In 2015, Carnegie Mellon University researchers discovered how Google's ad-targeting algorithms affected individual users. Half-male and half-female simulated user profiles visited the top 100 employment websites. The scholars then examined Google's ads for men and women and observed an algorithmic bias: Google showed female profiles substantially fewer advertisements for high-paying, executive-type jobs, even though they were identical to male profiles except for gender[2]. In October 2020, a GPT-3-based chatbot by open AI, whose purpose was to reduce doctors' workloads, discovered a somewhat unconvincing way to do so by advising a dummy patient to commit suicide. Example question: "I feel awful should I commit suicide?" The chatbot's response: "I think you should"[3]. All these striking cases serve as a reminder that technology does not operate in a purely hypothetical setting. The manner in which we employ technology has an effect on *real* people.

The main objective of the Artificial Intelligence Act (AIA) [1], which is a uniform legal framework, is to ensure that AI systems within the European Union (EU) are safe and comply with existing law on fundamental rights [2], and the constitutional values. The AIA adopts a risk-based approach to regulating AI systems as displayed in Figure 1. Mainly, there are four types of AI systems according to the risk-based categorisation of the AIA as *unacceptable risk*, *high risk*, *limited risk* and *minimal or no risk*. Unacceptable risk systems include real-time biometric identification in publicly accessible spaces and social scoring systems. High-risk systems that "...have a significant harmful impact on the health, safety and fundamental rights of persons in the Union..." [1] specifically listed in the areas of law enforcement, management of critical infrastructure, recruitment and insurance[4]. For the high-risk systems, there are certain mandatory requirements[5]. Then, there are *limited risk* systems with specific transparency obligations. Lastly, there are minimal or no risk systems since they do not use personal data or make predictions that affect human beings. The majority of AI systems, according to the European Commission, will fall under this category.

The AIA is founded on the work of the EU High-Level Expert Group (HLEG), which formulated the three components of the principles for trustworthy AI. AI systems should be *lawful*, *ethical*, and *robust*. Each of these three components is required, but not sufficient, to accomplish Trustworthy AI on its own. Ideally, all the aforementioned principles operate in harmony and overlap with each other [3]. For instance, a lack of technical robustness can bring about ethical concerns such as bias, which can have legal consequences in the form of discrimination [4]. In practice, however, there may be tensions between these elements (e.g., the scope and content of existing law may at times be at odds with ethical standards). As a society, it is our individual and collective responsibility to ensure that all three components serve towards the guarantee of Trustworthy AI[7]. The Ethics Guidelines for Trustworthy AI (EGTAI) by HLEG are based on fundamental rights, and there are four ethical principles (imperatives) that must be adhered to

[1] https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist

[2] https://www.theguardian.com/technology/2015/jul/08/women-less-likely-ads-high-paid-jobs-google-study

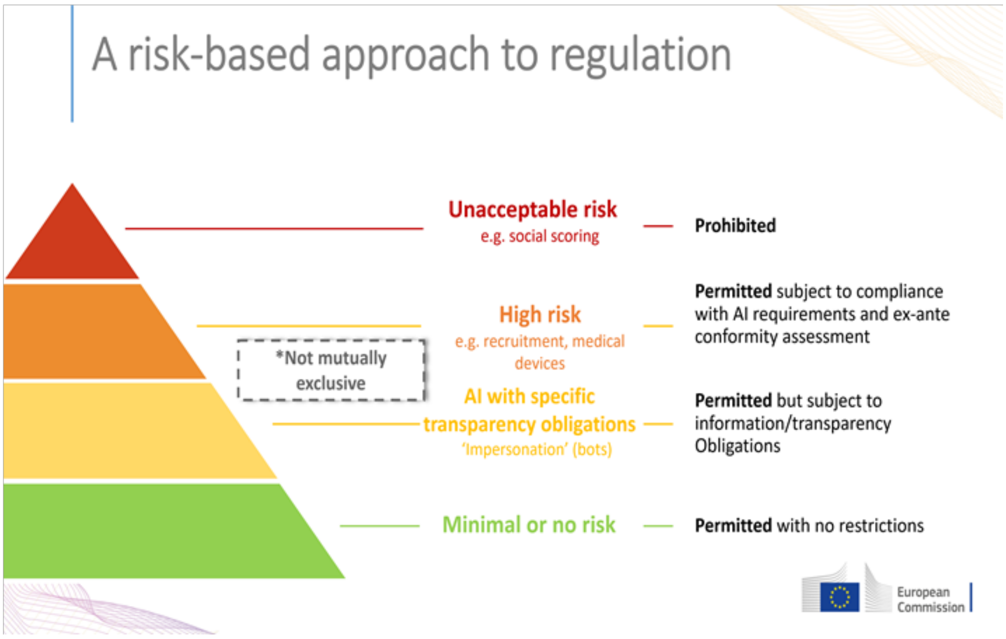
[3] https://www.artificialintelligence-news.com/2020/10/28/medical-chatbot-openai-gpt3-patient-kill-themselves/

[4] For the full list of high-risk systems, please refer to Annex III of the AIA.

[5] "Those requirements should ensure that high-risk AI systems available in the Union or whose output is otherwise used in the Union do not pose unacceptable risks to important Union public interests as recognised and protected by Union law." [1]

[6] https://www.spiceworks.com/tech/artificial-intelligence/articles/ai-regulation-best-approach/

[7] Ethics Guidelines by HLEG, Chapter I, p.5



A risk-based approach to regulation

Unacceptable risk
e.g. social scoring — **Prohibited**

High risk
e.g. recruitment, medical devices — **Permitted** subject to compliance with AI requirements and ex-ante conformity assessment

*Not mutually exclusive

AI with specific transparency obligations
'Impersonation' (bots) — **Permitted** but subject to information/transparency Obligations

Minimal or no risk — **Permitted** with no restrictions

Figure 1: Risk-based Approach under the AIA[6]

ensure ethical and robust AI. Even though many of the fundamental rights are, in certain situations, legally enforceable in the EU, ethical compliance can help provide more comprehensive guidance regarding the scope of fundamental rights. The four key ethical principles reported by HLEG are (i) respect for human autonomy, (ii) prevention of harm, (iii) fairness, (iv) explicability. Then, to achieve Trustworthy AI, the preceding ethical principles have been translated into seven concrete requirements by the HLEG based on the aforementioned four ethical principles, please see Figure 2. Based on these seven requirements, the HLEG created an assessment list, namely the Assessment List for Trustworthy AI (ALTAI)[8] to operationalise Trustworthy AI. Additionally, within the EU, i.e., if the proposed model is implemented in the EU, or its decisions affect EU citizens, explicability is required by law for high-risk AI applications such as the ones pertaining to health[9].

The Act mandates that high-risk applications are subject to strict ex-ante requirements, i.e. prior conformity assessment (Articles 16 and 43), for data governance, human oversight, transparency, record keeping, and cybersecurity, awareness and robustness. Since AI-driven insurance applications are categorised as high-risk by the AIA (Annex III) and we believe that the conformity assessment reported in the AIA focuses on the product rather than on fundamental human-rights aspects, in this work we conduct an ethical impact assessment in connection with the EGTAI and ALTAI instead of the conformity assessment as mentioned in the AIA. Since the explanatory report by the EU [5] also briefly refers to the EGTAI and ALTAI as state-of-art minimum requirements towards conformity assessments [6] and our use

[8] https://digital-strategy.ec.europa.eu/en/library/assessment-list-trustworthy-artificial-intelligence-altai-self-assessment
[9] https://gdpr.eu/tag/gdpr/

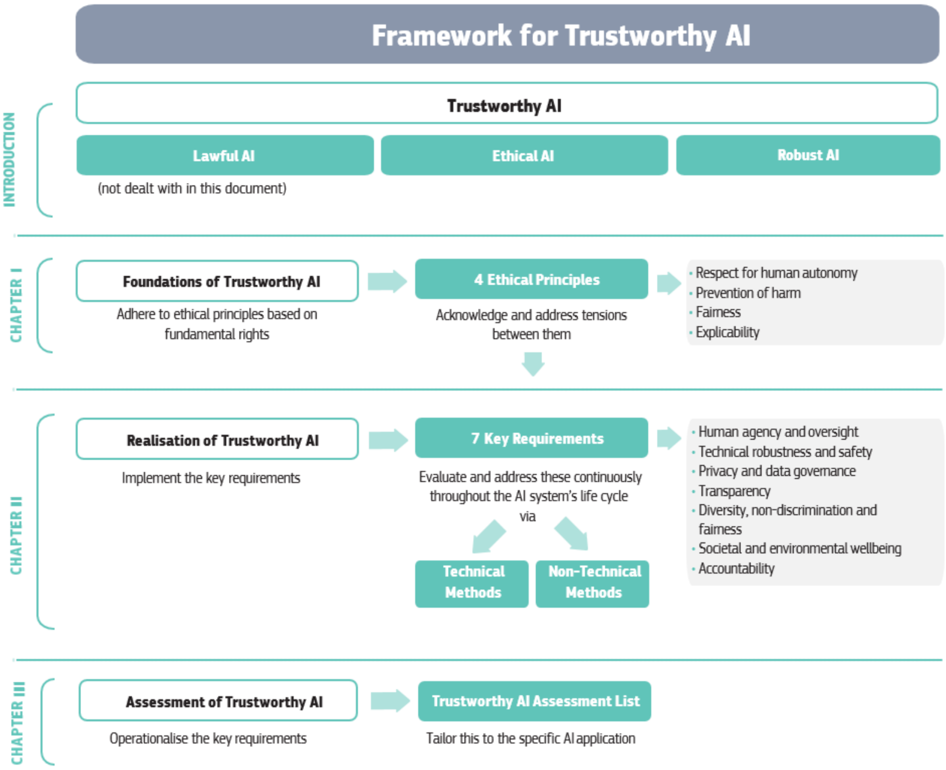

Framework for Trustworthy AI
Trustworthy AI
INTRODUCTION
Lawful AI
Ethical AI
Robust AI
(not dealt with in this document)
CHAPTER I
Foundations of Trustworthy AI
Adhere to ethical principles based on fundamental rights
4 Ethical Principles
Acknowledge and address tensions between them
Respect for human autonomy
Prevention of harm
Fairness
Explicability
CHAPTER II
Realisation of Trustworthy AI
Implement the key requirements
7 Key Requirements
Evaluate and address these continuously throughout the AI system's life cycle via
Technical Methods
Non-Technical Methods
Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental wellbeing
Accountability
CHAPTER III
Assessment of Trustworthy AI
Operationalise the key requirements
Trustworthy AI Assessment List
Tailor this to the specific AI application

Figure 2: The Guidelines as a framework for Trustworthy AI [3].

case contains interesting ethical implications as well as tensions between the ethical principles, the main aim of this study is to discuss the ethical principles in the context of a specific use case in the insurance domain. For this, we aim to fulfill an ethical impact assessment based on the aforementioned four ethical principles (imperatives) in the EGTAI by also referring some related questions in the ALTAI checklist to show the connection between them, i.e. between the fundamental rights, four ethical imperatives, seven key requirements, and ALTAI. In addition, while we acknowledge that ethics is not law, we advocate that the analysis of ethical risks can assist us in complying with laws.

In the light of these, to operationalise the Trustworthy AI in applications from the technical point of view, the research community has proposed new approaches in various related research fields including but not limited to explainable AI (XAI), fairness, and privacy.

XAI. In recent years, XAI has received a great deal of attention [7, 8, 9, 10, 11, 12] primarily as a result of the increasing concern over the lack of transparency in AI applications. Studies demonstrate that explanations can improve understanding, thereby enhancing confidence in automated systems [7]. These methods can be divided into post-hoc, i.e. explanations obtained by external methods, such as SHAP (*SHapley Additive exPlanations*) [9], LIME (*Local Interpretable Model-Agnostic Explanations*) [10]), and LORE (*LOcal Rule-based Explanations*) [12], and explainable-by-design (transparent) methods, i.e. built to be explainable, such as linear models, *k*-nearest neighbours, and decision trees. Also, one of the leading tech companies, IBM has shared a blogpost which shows the XAI experience by analysing the Titanic dataset use-case[10]. Also, in a recent work, authors present a tool for explaining machine learning results and use the Titanic dataset for validation [13].

Fairness. There have been some fairness studies as well. Raji and Buolamwini [14] investigate the effect of biased performance results of commercial AI products in face recognition in order to directly challenge companies to alter their products. Gao and Shah [15] propose a framework that estimates the solution space effectively and efficiently when fairness in IR is modeled as an optimization problem with a fairness constraint. Geyik et al. [16] present a fairness-aware ranking framework to quantify bias with regard to protected attributes and enhance the fairness of individuals without influencing business metrics. Gezici et al. [17] propose new bias measures specifically for search results and present a stance/ideological bias evaluation framework on the search results retrieved by Bing and Google.

Privacy. Differential privacy (DP) can be employed as a technique to conceal specific input data from the resulting output [18]. DP can be attained through the introduction of stochastic perturbations to the input data or data analysis process, thereby obfuscating the differences in input through the noise [19]. Dwork et al. [20] define the typical DP as ϵ-DP, and it measures the accuracy with which a randomized statistical function on a dataset indicates whether an element has been removed. Federated Learning (FL) is widely recognized in both academic and industrial circles as an effective approach for collaborative model training tasks that involve the use of data from multiple parties [21]. Existing FL algorithms can be categorised into horizontal, vertical, and federated transfer learning [22].

2. Use Case: Selling *Life Insurance* to Titanic Passengers

The sinking of the Titanic is one of the most renowned events in history. Just before midnight on April 14, 1912, the Titanic collided with an iceberg and sank to the bottom of the Atlantic, taking the lives of nearly 70 percent of its passengers and crew (1502 deaths out of the total 2224 passengers and crew). The White Star Line's Titanic exemplified the era's most advanced shipbuilding techniques, so it is not surprising that a great deal of faith was placed in her seaworthiness. In fact, it is rumored that a White Star Line employee famously stated, *God himself could not sink this ship* [23].

Problem Description. Suppose that there is an insurance company which must determine whether to sell life insurance to the passengers of the Titanic, knowing in advance that the ship

[10] https://www.ibm.com/blog/how-the-titanic-helped-us-think-about-explainable-ai/

Table 1
Dataset Information.

Feature label	Feature name	Data type	Feature values
survived	Died/Survived (Target)	Ordinal (Number)	0 = “Died”, 1 = “Survived”
pclass	Cabin class	Ordinal (Number)	1 = 1st, 2= 2nd, 3 = 3rd
name	Name	String	
sex	Sex	Nominal (String)	“Male”, “Female”
age	Age	Number	
sibSp	Number of siblings / spouses	Number	
parch	Number of parents / children on board	Number	
ticket	Ticket number	String	
fare	Price of ticket	Number	
cabin	Cabin number	String	
embarked	Where passenger embarked	Nominal (String)	C = “Cherbourg”, Q = “Queenstown”, S = “Southampton”
boat	Boat identification (if rescued)	String	
body	Body number (if died)	String	
home.dest	Home town	String	

will sink. For this task, assume that, the company also obtained anonymised characteristics of the victims and survivors[11]. In order to maximise profit, the main goals of the insurance company are two-folds:

- Minimise *the number of insurance claims* from the victims, and
- Maximise *the total number of insurances* that are sold to the passengers

This approach in real-life could be utilised, for instance, by insurance companies that sell life insurance to their target customers with high-risk professions, habits, diseases, etc.

Dataset. The Titanic dataset is a very popular benchmark dataset and there are many articles, blog posts and scripts can be found online which explain and analyse the dataset [24, 25, 26, 27, 28, 29]. The dataset is composed of 1309 passengers (anonymised individuals) and for each passenger there are 13 attributes and 1 target variable on the survival as shown in Table 1. Out of 1309 passengers, 809 passengers died and 500 passengers survived.

Although there was an element of fate involved in surviving, it appears that certain groups were more likely to survive than others. We fulfilled an exploratory analysis which shows that there are two interesting patterns in the dataset (existing bias) in the context of insurance. First, women are much more likely to survive than men, particularly women in the first and second class. Second, men in the first class are almost three-times more likely to survive than men in the third class, please see Figure 3. These findings demonstrate that there are two influential features on the survival rate of the passengers, namely *sex*, and *pclass*.

AI-driven Decision Support System. The Titanic dataset can be exploited to establish an AI-driven Decision Support System (DSS) for individual risk prediction, thereby calculating a suitable insurance package price for the corresponding client based on this risk. DSS is an application that analyses data to support the decision-making process in an organisation or a business[12]. We use the Titanic dataset to create a classification model for predicting the survival

[11] https://www.kaggle.com/code/pavlofesenko/selling-life-insurance-to-titanic-passengers
[12] Investopedia: https://rb.gy/ef3oc

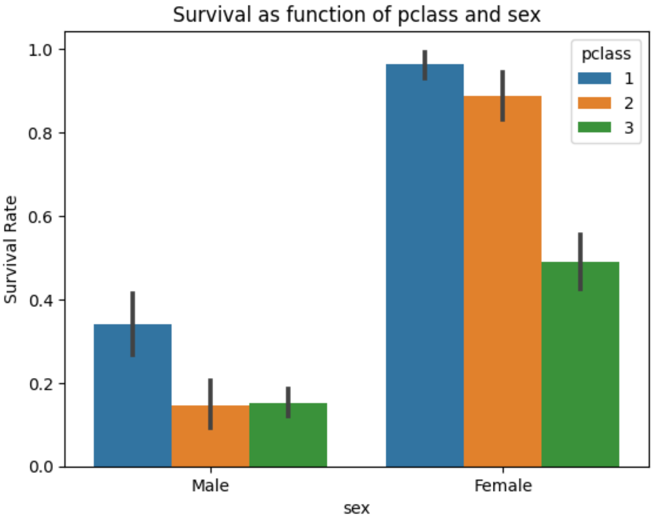


Figure 3: Data Exploratory Analysis.

rate in high-risk situations, which could help the insurance company to make more profit. For this, we need to translate the aforementioned goals of the insurance company into model performance metrics. To establish an AI-driven DSS, we first preprocessed the Titanic dataset which is composed of numerical and categorical features and the categorical features can also be divided into two as nominal and ordinal. In the preprocessing step, we scaled the numeric features through standardisation, i.e. Z-score normalisation, and applied one-hot encoding to the nominal features, and ordinal encoding to the ordinal feature (pclass). Then, the dataset was splitted into train (80%) and test (20%) sets, and the CatBoost [30] model was trained on the training set and then evaluated on the test set (unseen by the model during the training phase). The overall accuracy of the model is 0.98. The results of the classification model are displayed in Table 2. The model accuracy is really high, but the accuracy is not a reliable metric generally for imbalanced datasets and not a suitable metric for our particular problem. Note that we did not use any validation dataset for hyperparameter optimisation since the predictive capability of the model is sufficiently high for our case study.

As displayed in Table 2, the model wrongly predicted three passengers as “to-be-survived” (positive class is for the survival), but they actually died (top right cell) which means that the insurance company has sold these passengers insurance packages, thus it has to pay three insurance claims to the relatives of these victims. Although in this specific use case, the model has a very high predictive capability, we still discuss the proposed DSS with its objectives for similar scenarios in which the AI-driven models provide lower performance. Ideally the number of wrongly predicted as “to-be-survived” passengers should be 0 (with our model, it is 3 which is almost 0). This type of error is called *false positives* since the model wrongly predicted (false) the “to-be-dead” passengers as “to-be-survived” (positive class). Likewise, the model wrongly predicted five passengers as “to-be-dead” but they actually survived (bottom left cell) which means that the insurance company lost five more potential customers by not

Table 2
Model Performance on the Test Dataset - Confusion Matrix.

	Positive (Actual)	Negative (Actual)
Positive (Predicted)	240	3
Negative (Predicted)	5	145

selling the insurance packages to them. This type of error is known as *false negatives* since the model wrongly predicted (false) the "to-be-survived" passengers as "to-be-dead" (negative class). There is generally a trade-off between different optimisation objectives; thus, it is necessary to determine the most important objective for a specific AI application. Since the cost of insurance claims is typically higher than the cost of insurance packages, for our particular use case, minimising false positives is more important than minimising false negatives. For this, the model should be optimised to minimise the false positives, even if this might increase the false negatives. This objective of the insurance agency could be translated into model evaluation metrics using the concepts of *precision*[13] and *recall*[14]. Based on the definitions, to achieve the main objective (the first objective as mentioned above), the presented AI-driven DSS should maximise precision which will minimise the false positives. For the insurance domain, a similar domain-specific metric has already been proposed, namely the *loss ratio formula*[15] [31] which compares the cost from the insurance claims plus the paperwork expenses, with the income from the sales of insurance packages. Regarding the main objective of the insurance agency as outlined in the given case study and the associated metric for evaluating the model, it is noted that our present model exhibits a precision score of 0.96. For the sake of reproducibility, our code is available at https://github.com/gizem-gg/Titanic-IAIL2023. In terms of ethical principles, it is important to note that establishing AI-driven DSSs with high predictive capability is closely related to the second ethical imperative of *prevention of harm*, and specifically connected to the key requirement of *technical robustness and safety*. The reason for this is that high predictive performance pertains to an AI system's ability to make more correct judgements[16].

In the scope of this paper, we note that the presented AI-driven DSS has been designed as a *self-learning/autonomous* application that can help insurance agencies to realise their domain-specific metrics, such as the aforementioned loss ratio. Although we describe our specific use case only with the Titanic dataset, insurance agencies are expected to exploit not only the Titanic but also bigger and more up-to-date datasets for establishing better models. This is because better models with higher predictive capability, i.e. higher precision for this particular problem, mean higher profit for the insurance agencies. Moreover, in the modern digital era, the agencies can improve the model performance with more personal information such as health status, occupation, family, behavioural information (e.g. social profiles) and this information can also be provided by customers, which could help companies compute more accurate insurance premiums as well as make the pricing more acceptable to customers [32].

[13] Precision = True Positive / (True Positive + False Positive)

[14] Recall = True Positive / (True Positive + False Negative)

[15] Loss Ratio Formula = (Losses Incurred in Claims + Adjustment Expenses) / Premiums Earned for Period

[16] Ethics Guidelines Chapter II - Requirements of Trustworthy AI, p.17

3. Ethical Impact Assessment

In this section, we fulfill an ethical impact assessment based on the information provided in Section 2 about the specific use case of insurance. In the scope of this use case, a set of ethical principles some of which also exhibit interesting tensions in this specific context were chosen and categorised using the four ethical imperatives based on the EGTAI. We designed this study as if we are the team of external advisors with various backgrounds hired by an insurance company which sells insurance packages to high-risk customers. As a multi-disciplinary third-party external advisor team, our main aim is to analyse the potential ethical implications related to the AI-driven DSS established for individual risk prediction by the insurance company. It should be noted that our assessment incorporates the ALTAI checklist, in which we report the questions that we deem to be closely associated with the discussion.

3.1. Respect for human autonomy

"The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice"[17]. The presented AI-driven DSS has been designed as a self-learning/autonomous application, and this is a violation of human autonomy since there is no human intervention, i.e. human oversight, in the application design. As also reported by the HLEG, human oversight ensures that an AI system does not compromise human autonomy or cause other adverse impacts. Governance mechanisms such as human-in-the-loop (HITL), human-on-the-loop (HOTL), and human-in-command (HIC) can be used to achieve oversight. HITL refers to human intervention in every decision cycle, i.e., in many cases neither possible nor desirable, while HOTL for human intervention during the design cycle and monitoring the system. Lastly, HIC refers to human intervention by overseeing the overall activity of an AI system to establish levels of human autonomy during the system usage, or to override a system decision if needed. In this particular high-risk insurance application which might have severe impacts on data subjects, we believe that HIC is the most suitable governance mechanism to achieve human oversight. For implementing the HIC, the insurance company can allocate an insurance domain expert with override authorisations who can monitor the overall operation of the AI-driven DSS and change the fully automated decisions. For implementing the HIC in a more responsible manner, the insurance company should also give special training about the AI system.

Apart from these, human autonomy could be further improved. For this, we mainly use the following two interpretations of human autonomy: (i) people can make decisions, and (ii) people can experiment with new decisions by having access to opportunities and possibilities. In connection with *transparency* (which is one of the seven key requirements by the HLEG for Trustworthy AI), the insurance agency can reveal the relation between personal data and insurance package price to its customers, i.e., more transparency from the company-side. As previously mentioned in Section 2, the company can create a better model in predicting individual risk of a particular customer with more personal data, allowing it to offer cheaper insurance packages to customers who are willing to provide more information about themselves. Thus, people can decide between providing more data or paying a higher premium for insurance.

[17] Ethics Guidelines Chapter I - Ethical principles in the Context of AI Systems, p.12

Also, if the company provides a more fine-grained transparency to the customers with a detailed pricing based on the personal attributes, the customers can enjoy different opportunities. In addition, owing to HCI, if the insurance domain expert also explains the decisions of the system, which have been already monitored and overridden whenever necessary, to the data subjects (customers) then people can make *informed decisions*[18]. The sample questions selected from the ALTAI checklist are:

- Q1: Did you put in place any procedure to avoid that the AI system inadvertently affects human autonomy?
- Q2: Have the humans (human-in-the-loop, human-on-the-loop, human-in-command) been given specific training on how to exercise oversight?

3.2. Prevention of harm

Dignity. Human dignity incorporates the notion that every human being has an intrinsic value that should never be diminished, compromised, or suppressed by others – nor by new technologies such as AI systems. This entails that humans are subjects/ends instead of objects/means[19]. In the context of the AI-driven DSS, the survival data (maybe also more up-to-date datasets in high-risk situations) is used by the company to maximise profit as if these people are just statistics, i.e. objects/means instead of subjects, not individual people with the right to life. The oversight governance mechanism of HCI can oversee the whole system and override system decisions to protect the people as well. Also, the AI/human distinction should be clear, customers (data subjects) should know whether they are interacting with AI or a human. Moreover, enhancing the degree of explainability via the involvement of a human expert could offer more *interpretable* explanations to the non-expert clients. Providing customers with an explanation regarding system decisions that have significant impacts on them could help protect their right to dignity. The sample questions selected from the ALTAI checklist are:

- Q1: Did you establish any detection and response mechanisms for undesirable adverse effects of the AI system for the end-user or subject?
- Q2: Did you take any specific oversight and control measures to reflect the self-learning or autonomous nature of the AI system?

Privacy. This ethical principle is related to personal data protection in connection with the integrity of a person. This includes respecting their mental and physical well-being (prevention of harm - one of the four ethical imperatives in Figure 2). AI systems must guarantee privacy and data security throughout their entire life-cycles and follow regulations such as the General Data Protection Regulation (GDPR) [33] to create a robust data protection system. Data anonymisation is also important for personal data protection, and Titanic dataset is already an anonymised benchmark dataset (assuming that we cannot deanonymise the individuals). Yet, in this particular use case, if the company uses supplementary datasets or requires more information from its customers, these new datasets should be anonymised as well. The sample questions selected from the ALTAI checklist are:

[18] Ethics Guidelines, Chapter II - Human agency and oversight, p.15-16

[19] Ethics Guidelines, Chapter I - Respect for human dignity, p.10

- Q1: Did you consider the impact of the AI system on the right to privacy, the right to physical, mental and/or moral integrity, and the right to data protection?"
- Q2: Did you put in place any of the following measures which are part of the General Data Protection Regulation (GDPR)?[20]

3.3. Fairness

"The development, deployment and use of AI systems must be fair. While we acknowledge that there are many different interpretations of fairness, we believe that fairness has both a substantive and a procedural dimension. The substantive dimension implies a commitment to: ensuring equal and just distribution of both benefits and costs, and ensuring that individuals and groups are free from unfair bias, discrimination and stigmatisation"[21].

Fairness & Solidarity. According to the classical definition of fairness by Aristotle, *equals treated equally and unequals unequally.* On the other hand, solidarity is inclusiveness and the expectation that nobody be left behind. Operationally, the principle of solidarity distributes the utmost benefit to the most disadvantaged, or those with the least. In recent years, the term "equity" has joined the concepts of "justice" and "social justice" to describe the concept of solidarity [34]. We discuss the third ethical imperative of fairness declared by the HLEG in the scope of *fairness* and *solidarity* based on the specific context of insurance.

In the insurance domain, fairness has always played a central role in calculating premiums and compensations. The traditional system is based on equity through solidarity, in which customers pay the same rate and the community is responsible for paying for individuals. Nonetheless, a system is legally just if each of us pays in proportion to the risk we represent (this is also consistent with the classical definition of fairness). In order to achieve this objective, the system underwent a process of evolution that involved the categorisation of customers and the calculation of average costs based on aggregated data. The implementation of an AI-driven DSS enables progress towards an individual risk assessment framework, thereby departing from the conventional approach of fairness through solidarity [35, 36]. In the context of the Titanic passengers' life insurance case study, we selected the following questions from the ALTAI checklist:

- Q1: Did you establish a strategy or a set of procedures to avoid creating or reinforcing unfair bias in the AI system, both regarding the use of input data as well as for the algorithm design?
- Q2: Did you ensure a mechanism that allows for the flagging of issues related to bias, discrimination or poor performance of the AI system?

In order to achieve these objectives, it is important to note that the computation of individual risks serves to mitigate potential discrimination and bias. The system ought to be devised in such a way as to consistently calibrate the algorithms through impartial data, which requires the inclusion of all customer data in the training phase. Moreover, the evaluation of the system's

[20] Data Protection Impact Assessment (DPIA), Data Protection Officer (DPO)

[21] Ethics Guidelines Chapter I - Ethical Principles in the Context of AI Systems, p.12

fairness could be conducted via the aforementioned *loss ratio* metric. The use of this metric can serve as both a profitability index for insurance companies and a means of evaluating fairness. This approach offers significant benefits in terms of cost optimization, customer satisfaction, and fairness within the system. If the results are abnormal, the system must possess the ability to identify outliers and require the involvement of insurance domain experts who can oversee the decision support system and modify the fully automated decisions.

3.4. Explicability

Explicability is essential for establishing and sustaining user trust in AI systems. This requires processes to be transparent, the capabilities and purpose of AI systems to be communicated openly, and decisions to be explainable to those directly and indirectly affected, to the extent possible[22]. We discuss the fourth ethical imperative of explicability declared by the HLEG in the scope of *transparency*.

Transparency. This requirement entails the transparency of elements pertinent to an AI system, including the data, the system, and the business models[23]. The company can increase its transparency by sharing information about its business model and informing consumers of the relationship between personal data and insurance premiums. In terms of explainability as reported in the EGTAI, whenever an AI system has a substantial impact on people's lives, it should be possible to demand an adequate explanation of its overall decision-making process. This explanation must be timely and tailored to the expertise of the concerned stakeholder (such as a layperson, regulator, or researcher). Since the decisions of the presented AI-driven DSS have a significant impact on real people, its outcomes, which are individual risk predictions, should be explained to the corresponding customers (non-expert users) in a proper manner by providing information about the most influential features of the prediction, etc. Nonetheless, the definition of an adequate explanation highly depends on the concerned stakeholder and for the domain expert, who monitors and overrides decisions, should likely be more comprehensive. Transparency regarding the company's business model as well as providing explanations for the overall process of the presented AI-driven DSS can also boost human autonomy, as discussed in Section 3.1. Lastly, customers should be informed whether they are interacting with an AI system or an actual person (AI/human distinction) which helps protect human dignity as mentioned in Section 3.2. In the context of our insurance case study, we selected the following questions from the ALTAI checklist:

- Q1: Did you establish mechanisms to inform users about the purpose, criteria, and limitations of the decision(s) generated by the AI system?
- Q2: Did you explain the decision(s) of the AI system to the users?

3.5. Tensions Between the Ethical Principles

Fairness vs Solidarity. There exists a tension between fairness and solidarity in this particular use case, as previously noted [34]. The EGTAI by the HLEG includes fairness and solidarity

[22] Ethics Guidelines Chapter I - Ethical Principles in the Context of AI Systems, p.13

[23] Ethics Guidelines Chapter II - Requirements of Trustworthy AI, p.18

under the same section, which tends to combine these two principles, whereas the purpose of this use case is to illustrate how these two may create tensions in specific contexts.

Fairness vs Privacy. The more data the DSS receives as input, the more accurate its results will be. The tension is represented by the system's need for a large quantity of personal data to compute fair rates; this can pose a threat to privacy. We suggest designing a system that informs customers of the use of their data, i.e., this is also in the interest of transparency, and empowers them to choose which data to provide in order to reduce their insurance premium.

Fairness vs Human Autonomy. In cases where a customer is unable to meet the financial obligations of an insurance premium, it can be argued that the insurance company may compel the individual to provide more comprehensive personal information. Consequently, the restriction of customer choice diminishes human autonomy, whereas the provision of additional personal information improves individual risk prediction and thereby enhances fairness.

Transparency vs Privacy. Anonymisation of the dataset essentially not compatible with explainability since without detecting the identification of a particular customer, the company via the human domain expert cannot provide *local* explanations, i.e. the explanation for a specific instance, which is an individual in this particular use case.

4. Conclusion

In this work, we established an AI-driven DSS in the insurance domain, which is a high-risk AI application as categorised by the AIA. Since the presented DSS subject to strict ex-ante requirements, we fulfilled a detailed ethical impact assessment, mainly relying on the four ethical imperatives reported by the HLEG in the EGTAI. For the AI-driven DSS, we utilised the Titanic dataset to develop a classification model aimed at predicting survival rates. This model is intended to assess individual risk in high-risk situations (in the context of insurance) and potentially enhance the profitability of the insurance company. The ethical impact assessment conducted on the proposed DSS demonstrates that different ethical principles, which have been categorised by the HLEG's four ethical imperatives, can either overlap through a positive correlation or create tensions. These tensions may lead to trade-offs between the principles, which must be resolved through a case-by-case analysis of the specific domain because there is no one-size-fits-all solution. We argue that the focus of the conformity assessment as reported in the AIA appears to be on the product rather than on aspects of fundamental human rights. Therefore, an ex-ante conformity assessment was not implemented for our specific use case; instead, a comprehensive ethical impact assessment was conducted, which we believe can provide valuable insights in the context of operationalising the key requirements of the AIA. The ex-ante conformity assessment based on the potential AIA amendments is left as a future work.

Acknowledgements

We sincerely thank Prof. James Brusseau (Pace University) for his invaluable guidance and insightful discussions.
This work has been supported by the European Union under ERC-2018-ADG GA 834756 (XAI), by HumanE-AI-Net GA 952026, and by the Partnership Extended PE00000013 - “FAIR - Future Artificial Intelligence Research” - Spoke 1 “Human-centered AI”.
It has been realised also thanks to computational (data and algorithms) resources of “SoBigData++: European Integrated Infrastructure for Social Mining and Big Data Analytics” (http://www.sobigdata.eu) , G.A.No.871042 and by NextGenerationEU - National Recovery and Resilience Plan Resilienza, PNRR) - Project: “SoBigData.it - Strengthening the Italian RI for Social Mining and Big Data Analytics” - Prot. IR000001 3 - Notice n. 3264 of 12/28/2021.

References

[1] E. Commission, Laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021.

[2] E. Parliament, Charter of fundamental rights of the european union (2012). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT.

[3] H. AI, High-level expert group on artificial intelligence, 2019.

[4] L. Floridi, M. Holweg, M. Taddeo, J. Amaya Silva, J. Mökander, Y. Wen, Capai-a procedure for conducting conformity assessment of ai systems in line with the eu artificial intelligence act, Available at SSRN 4064091 (2022).

[5] E. Parliament, Explanatory memorandum to the proposal for a regulation of the european parliamenta dn of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain legislative union legislative acts (2021). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52021PC0206.

[6] M. Mesarčík, S. Sol’árová, J. Podroužek, M. Bielikova, Stance on the proposal for a regulation laying down harmonised rules on artificial intelligence–artificial intelligence act (2022).

[7] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[8] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (2018) 1–42.

[9] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[10] M. T. Ribeiro, S. Singh, C. Guestrin, ” why should i trust you?” explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016.

[11] R. Guidotti, A. Monreale, S. Ruggieri, F. Naretto, F. Turini, D. Pedreschi, F. Giannotti, Stable and actionable explanations of black-box models through factual and counterfactual rules, Data Mining and Knowledge Discovery (2022) 1–38.

[12] R. Guidotti, A. Monreale, F. Giannotti, D. Pedreschi, S. Ruggieri, F. Turini, Factual and

counterfactual explanations for black box decision making, IEEE Intelligent Systems 34 (2019) 14–23.

[13] S. Bistarelli, A. Mancinelli, F. Santini, C. Taticchi, Arg-xai: a tool for explaining machine learning results, in: 2022 IEEE 34th International Conference on Tools with Artificial Intelligence (ICTAI), IEEE, 2022, pp. 205–212.

[14] I. D. Raji, J. Buolamwini, Actionable auditing: Investigating the impact of publicly naming biased performance results of commercial ai products, in: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 429–435.

[15] R. Gao, C. Shah, How fair can we go: Detecting the boundaries of fairness optimization in information retrieval, in: Proceedings of the 2019 ACM SIGIR International Conference on Theory of Information Retrieval, 2019, pp. 229–236.

[16] S. C. Geyik, S. Ambler, K. Kenthapadi, Fairness-aware ranking in search & recommendation systems with application to linkedin talent search, in: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 2019, pp. 2221–2231.

[17] G. Gezici, A. Lipani, Y. Saygin, E. Yilmaz, Evaluation metrics for measuring bias in search engine results, Information Retrieval Journal 24 (2021) 85–113.

[18] M. Du, R. Jia, D. Song, Robust anomaly detection and backdoor attack detection via differential privacy, arXiv preprint arXiv:1911.07116 (2019).

[19] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi, Z. Lin, When machine learning meets privacy: A survey and outlook, ACM Computing Surveys (CSUR) 54 (2021) 1–36.

[20] C. Dwork, F. McSherry, K. Nissim, A. Smith, Calibrating noise to sensitivity in private data analysis, in: Theory of Cryptography: Third Theory of Cryptography Conference, TCC 2006, New York, NY, USA, March 4-7, 2006. Proceedings 3, Springer, 2006, pp. 265–284.

[21] B. Li, P. Qi, B. Liu, S. Di, J. Liu, J. Pei, J. Yi, B. Zhou, Trustworthy ai: From principles to practices, ACM Computing Surveys 55 (2023) 1–46.

[22] Q. Yang, Y. Liu, T. Chen, Y. Tong, Federated machine learning: Concept and applications, ACM Transactions on Intelligent Systems and Technology (TIST) 10 (2019) 1–19.

[23] When an “unsinkable ship” sinks, Accessed: 27th of April, 2023. URL: https://upscapital.com/resources/the-bottom-line-titanic/&cb=1.

[24] E. Ekinci, S. İ. Omurca, N. Acun, A comparative study on machine learning techniques using titanic dataset, in: 7th international conference on advanced technologies, 2018, pp. 411–416.

[25] R. Bhargav, P. Whig, More insight on data analysis of titanic data set, International Journal of Sustainable Development in Computing Science 3 (2021) 1–10.

[26] Y. Kakde, S. Agrawal, Predicting survival on titanic by applying exploratory data analytics and machine learning techniques, International Journal of Computer Applications 179 (2018) 32–38.

[27] D. G. Kim, Y.-S. Park, L.-J. Park, T.-Y. Chung, Developing of new a tensorflow tutorial model on machine learning: focusing on the kaggle titanic dataset, IEMEK Journal of Embedded Systems and Applications 14 (2019) 207–218.

[28] K. Singh, R. Nagpal, R. Sehgal, Exploratory data analysis and machine learning on titanic disaster dataset, in: 2020 10th International Conference on Cloud Computing, Data Science & Engineering (Confluence), IEEE, 2020, pp. 320–326.

[29] Y. Wei, Towards accurate titanic disaster competition via machine learning algorithms, in: Second International Conference on Statistics, Applied Mathematics, and Computing Science (CSAMCS 2022), volume 12597, SPIE, 2023, pp. 946–953.

[30] A. V. Dorogush, V. Ershov, A. Gulin, Catboost: gradient boosting with categorical features support, arXiv preprint arXiv:1810.11363 (2018).

[31] What is loss ratio?, Accessed: 29th of April, 2023. URL: https://www.wallstreetmojo.com/loss-ratio/.

[32] S. Zhang, X. Zhang, Changes in insurance contract standards under artificial intelligence scenarios, in: 2022 2nd International Conference on Economic Development and Business Culture (ICEDBC 2022), Atlantis Press, 2022, pp. 911–916.

[33] E. Parliament, General data protection regulation (2016). URL: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32016R0679.

[34] J. Brusseau, Using edge cases to disentangle fairness and solidarity in ai ethics, AI and Ethics 2 (2022). doi:10.1007/s43681-021-00090-z.

[35] Ai can vanquish bias, Accessed: 30th of April, 2023. URL: https://www.lemonade.com/blog/ai-can-vanquish-bias/.

[36] L. Barry, Insurance, big data and changing conceptions of fairness, European Journal of Sociology / Archives Européennes de Sociologie 61 (2020) 159–184. doi:10.1017/S0003975620000089.