The Insurance Industry: insights into challenges for indirect discrimination

Research project on the use of automated decision-making by insurers to meet consumers' needs

Freyja van den Boom, PhD

"This material is based upon work supported in whole or in part by The Notre Dame-IBM Tech Ethics Lab. Such support does not constitute endorsement by the sponsor of the views expressed in this publication."

Section 1
Introduction

Building upon the insights gained from research that there is a need to help insurers ensure their use of algorithm decision-making remains lawful and ethical, this report shows how a bias detection tool together with a decision-making framework can help insurers to gain insights and decide upon the use of algorithms for decisionmaking.

This project builds and expands upon previous research on whether the European Union (EU) legal and regulatory framework is fit for purpose in relation to the take-up of telematics insurance.[1] As an example of use-based insurance, telematics insurance is made possible by continuing advancements in data analytics and sensor technologies as well as the shift in the automotive industry towards the development of a business ecosystem around the connected car and value proposition that user-generated car data brings.

With a focus on balancing the different interests of the stakeholders involved for competition, privacy, and innovation recommendations to improve the legal framework include providing more clarity about the scope of the relevant rights and responsibilities for the principal stakeholders (consumers, insurers, and car manufacturers), especially for the data and information-sharing duties they have towards each other. The regulation of competition and enforcement must be improved to address the potentially disruptive effects of the shift towards business ecosystems and non-traditional market players. Furthermore, the overall coherence of the regulatory environment needs attention by clarifying for stakeholders how to comply in the case of conflicting requirements that stem from the different regulations that apply.

Responding to these recommendations, the following report presents the development of tools that are aimed at helping insurers decide whether to use algorithm decision-making in their consumer risk assessments.

Discrimination when someone is treated unfairly because of a protected characteristic, such as sex or race is unlawful. This is known as direct discrimination. However, insurers need to be able to differentiate between consumers if they want to price their premiums to cover the cost.[2] Therefore not all discrimination is unlawful. Indirect discrimination is when there are rules that apply to a group of people, which in theory could be everyone but in practice are less fair to a certain protected characteristic. A well-known example in the insurance industry is the use of postcode which has been shown could lead to discrimination based on race. Indirect discrimination is lawful when insurers can provide an objective justification.[3] We will go into more detail about this in section two.

[1] Van den Boom, F (2022) Driven by digital innovations: Regulating In-vehicle data access and use, in M Borghi and R Brownsword (Eds): Informational Rights and Informational Wrongs: A Tapestry for Our Times, Routledge

[2] Xin, Xi and Huang, Fei, (2022) Anti-Discrimination Insurance Pricing: Regulations, Fairness Criteria, and Models

[3] It should be noted that the distinction between direct and indirect discrimination is increasingly difficult to make in practice.

What became clear from the discussions with insurers is that they struggle to know whether their decision-making is unlawful, especially as a result of their increasing use of automated decision-making. To help insurers become aware of potential bias and unlawful decision-making, the aim of this report and the underlying research was to provide insurers with a set of tools for bias detection and ethical decision-making. The tool can identify direct bias and the flows of discrimination through the other variables in the case of indirect discrimination. The results presented here show that the tool is useful for insurers to help them identify bias and both direct and indirect discrimination.

Approach

This report is based on independent and collaborative research between researchers, several Dutch insurance companies and the Dutch Insurance Association.
The first tool development and tests were done as part of a master's thesis research project and described in section two.[4] Further testing and evaluation were done in collaboration and complemented with new insights gained from further document analysis and semi-structured interviews. The initial tool was developed and the additional research was for the most part completed at the beginning of 2022.

The following sections of this report present the tool and evaluation of its functionality (section two) an analysis of ethical models for decision-making (section three) and concludes with considerations based on the insights gained from the previous sections (section four)

Automated decision-making and the insurance industry

Insurance is a contract between consumers and insurers whereby the insurer allows the consumer to exchange the uncertainty of damage (risk) with the certainty of paying a monthly fee (the premium). To assess the risk for a consumer to call upon insurance, the insurer has to accurately estimate the probability that the damage will occur and the extent of that damage. The more accurately they can score people, the more profitable they will be.[5]

There are specific challenges for insurers in terms of insurance pricing. This includes t*he existence of information asymmetry.* To meet the insurer's need to be able to make a proper assessment of the risks, consumers are required by law to answer truthfully to questions about facts that the insurer considers important (material) to be able to make a proper assessment of the risks. As the consumer was the one who had this information and might not be willing to share it given its importance in terms of lower costs and premiums, this information imposed an obligation on the consumer to respond truthfully to asymmetry. The need for information about the consumer and their behaviour is linked to *the risk of adverse selection*. If an insurer cannot properly assess the risk a person poses, it may result in people with high risks paying too little and people with low risks paying too much to cover their respective costs. As a result, it is understood that this insurance becomes attractive for high-risk persons and low-risk persons may decide not to insure themselves or go to competitors. To remain cost-effective, the insurer must

[4] Bernt van Walree & Rogier Potter van Loon (2020) Bias detection tool, report for the vereniging verbond verzekeraar; paper demo 3D tool ref 2022-876322534-38276/jscha

[5] Not knowing your personal risk score may make it easier to accept being placed in a risk pool. T. Timmer, I. Elias, L. Kool & R. van Est (2015). Berekende risico's. Verzekeren in de datagedreven samenleving, Den Haag, Rathenau Instituut; I. Kerr & J. Earle, 'Prediction, Preemption, Presumption: How Big Data Threatens Big Picture Privacy' (2013) 66 Stanford Law Review Online 65

increase its premiums with the possibility of even more low-risk people will leave and only the high risks remain. Finally, having insurance has been shown to lead to a change in behaviour in such a way that their risk increases leading to an increase in cost for the insurer and premiums no longer being accurate.[6]

As a result of these factors, the premium for consumers does not only include a calculation of the actual risk but also internal costs and economic interests. With the development of AI Insurers have become better at analysing premiums making them more fair and personalised, however, this is also leading to a growing awareness of the risk of bias, discrimination and unfair pricing practices.

Automated decision-making and lawful discrimination.

Insurers have a responsibility to show whether their insurance practices including their use of innovative practices involving personal data and algorithms are lawful. An insurer using automated processes may have to demonstrate why the algorithm selects certain factors as being relevant in determining the risk and that these factors are not chosen because they are predictive of otherwise protected characteristics. In addition, we argue it's becoming increasingly important to monitor outcomes and the impact of decisions in society to ensure these on a whole do not have adverse effects such as being unlawfully discriminating.[7]

Most countries have anti-discrimination legislation in place to protect people. In The Netherlands for example, this is regulated in the General Equal Treatment Act. Direct discrimination on the grounds of religion, belief, political affiliation, race, gender, nationality, sexual orientation or marital status is generally prohibited unless one of the exceptions applies. Given that in most cases it will be difficult for an insurer to explain why race or someone's nationality is the decisive factor, direct discrimination on the basis of protected grounds is prohibited. This is different from indirect discrimination when there can be an objective justification by insurers to discriminate between people based on their risk profile.
This objective justification can also be used when there is discrimination on other grounds than those mentioned in the law such as age and disability to justify direct and/or indirect discrimination.

For indirect discrimination to be justified, there are three criteria to be met namely:

1. the aim must be legitimate
2. the means must be appropriate
3. the means must be necessary

For insurers, one of the main concerns about their use of automated decision-making and increasingly complex algorithms is how to ensure they do not directly discriminate and when the discrimination is indirect to be able to demonstrate that there is an objective justification to do so. The proposed tools were developed with the aim to help insurers address these challenges.

The following section presents the findings based on previous research done in the context of a master's project on developing an algorithm audit tool for bias detection and measurement.[8]

---

[6] Van den Boom, F (2022)

[7] Statistics Netherlands (CBS) provides reliable statistical information and data. Avaialble at https://www.cbs.nl/en-gb

[8]

Section 2
Differentiation detection in insurance pricing models

Using AI can lead to optimization through personalizing prices which can have a negative impact on solidarity in society and lead to biased and differentiating pricing. Bias is defined here as *the systematic and unfair discrimination against certain individuals or groups of individuals based on attributes such as Age, Gender, Nationality or Ethnicity.*

There are several reasons why the outcome of a model leads to bias. This can be because the underlying training data was biased, the algorithm itself is programmed to be biased, or the algorithm with which the model is created may have an alignment problem between its goal and means to achieve it. [9] For example, if the algorithm is designed to achieve the highest possible accuracy, it can amplify certain biases in the data, in order to achieve that goal.[10]

In the case of insurance, the pricing models are often so complex that even its developers do not fully understand how they produce their outcomes. Bias, therefore, can be hard if not impossible to detect by human oversight alone. There are several ways to solve this problem of biased AI algorithms, including using more representative training datasets and removing features that could otherwise be used to identify protected characteristics.[11] These are not always useful in the case of insurance models, however. The problem with excluding protected characteristics in the data set used by the pricing model does not address the issue of proxy discrimination. P*roxy discrimination* happens when the variable although excluded still impacts the model outcome because there is another variable that is used for the same effect. The colour of a car, for example, can be highly indicative of gender, so when this information is included instead of directly including a person's gender the outcome may still have the same discriminatory outcome. To help insurers become more aware and able to check whether their models contain bias, we have developed the so-called bias detection tool.[12]

Automated decision-making: bias detection tools.

As mentioned, discrimination happens when someone or a particular group in society is treated less fairly than others, however indirect differentiating on a protected feature may be legal if the differentiating can be justified. The tool described in more detail below, helps insurers to know which features are used by their pricing model and the importance (weight) of these leading to a certain outcome.

[9] See for more on this Russel: the 3 issues of value alignment and proposes a sullen to have algorithms be human aligned.

[10] M. Hildebrandt, 'Algorithmic Regulation and the Rule of Law' (2018) 376 Philosophical Transactions of the Royal Society A; R. Leenes, E. Palmerini, B.-J. Koops, A. Bertolini, P. Salvini & F. Lucivero (2017). Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues, Law, Innovation and Technology. R. Leenes & F. Lucivero (2014). Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design, in Law, Innovation, and Technology

[11] Other sollutions include modifying the dataset through feature engineering that makes protected characteristics undifferentiated; masking the features in the dataset by random shuffle, and/or using data augmentation.Van Walree & Potter van Loon (2020)

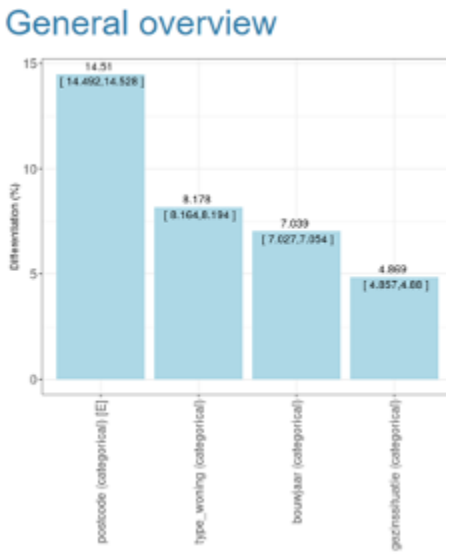
[12] Idem

Empirical findings

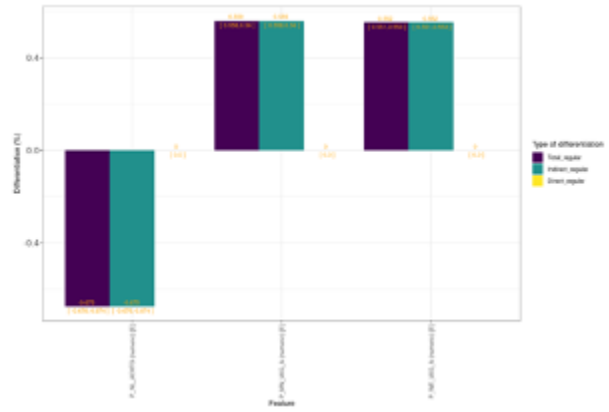
The tool consists of two components; The programming code (R-script) which can be used to detect and measure differentiating and an app (Shiny) that can visualize the results of the code.

The R-script requires as input a user data set and a model (in R). The user of the tool must make some preparations before they can run the script. This mostly comprises entering 'TRUE' or 'FALSE' in a couple of fields corresponding to what kind of model the user provides. After that, the user can run the script. Usually, this takes between 10 and 20 minutes depending on the model of the user and the given data set. The user then uploads the saved file to the shiny app for a graphical overview of all the findings produced by the R-script.

The tool uses CBS data to get (extra) information on discrimination. In the app, the user has several options for what they want to see. This includes what type of discrimination (direct, indirect or total) and measurements (a couple of distinct variables or just an overview of a selected number of most differentiating variables)



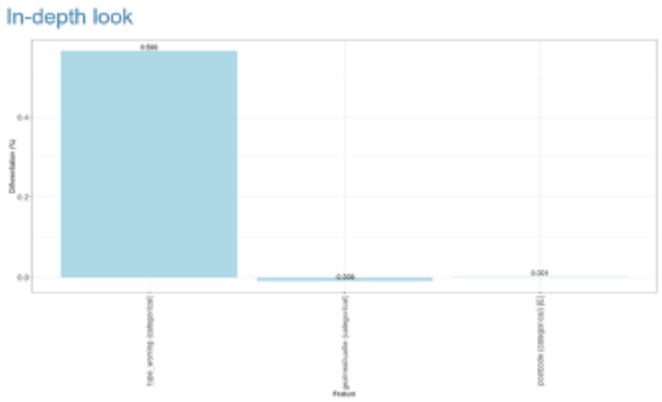
The first graph shows the 4 most differentiating variables regarding total discrimination.



The second graph using as an example the results for non-Western migrants shows all the discrimination results for three variables from the CBS data set regarding ethnicity. The measurements further show how much the insurance premiums change on average when the input feature changes by a standard deviation or by category. The second graph further shows that on average non-Western migrants pay a higher premium than other groups. The discrimination measurement shows that when the percentage of non-Western migrants in an area increases, the insurance premium increases. The tool enables the user to question why this is by providing the opportunity to see through which other variables and differentiation flows.

Direct differentiation is 0 because the model does not directly differentiate on non-Western migrants because it doesn't use CBS data as input.[13]

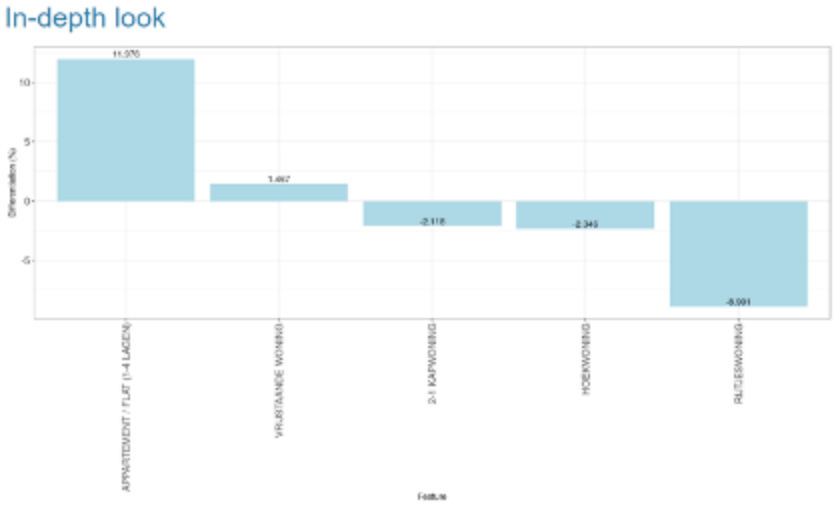
The results in the following graph show that almost all the discrimination comes from what type of house people live in.



The type of house corresponds to a higher number of non-western migrants, which leads to a higher insurance premium. This prompts the question for insurers of what type of house has a higher discrimination measurement or warrants a higher premium. The user can then go into more detail to see what type of house leads to a higher differentiation measurement or warrants a higher premium

In the app, the user can see that consumers who live in an apartment pay a significantly higher insurance premium than people living in other types of accommodation.

[13] Statistics Netherlands (CBS) provides reliable statistical information and data.



The tool enables users to get a better understanding of how the model behaves in terms of differentiating between people based on certain features. These insights can then prompt action on whether to accept the discrimination or if the model leads to unacceptable outcomes.

The example of non-Western immigrants paying a higher premium based on housing flags the need to look further into whether it's warranted to differentiate based on the such characteristic. The insurer should for example look at the relationship between non-Western migrants and the type of house people live in as the app shows that there is a strong relationship between the two features. Next, the user should check the effect each type of housing has on the insurance premium. In this example, it looks like non-Western migrants usually live in apartments and that apartments lead to a higher insurance premium. Whether it is just to differentiate based on these factors depends on the justification that this is indeed the case in reality and that the insurer can justify using these variables based on research that shows its relevance for the insurance risk.[14]

Bias Detection Tools comparison.[15]

Although both tools have the same aim to detect (unwanted) discrimination of models used in the insurance industry they work differently.

Tool A consists of two components namely an R-script to produce the discrimination detection results of the model; and a Shiny app which will visualize the results the R-script produces. The user runs the script on their own model and data set. After providing some details about the model the user runs the script which produces a file that can be uploaded into the Shiny app to see a visualization of all the findings.[see screenshot]

Tool V is more simple in its use because it only requires the user to upload a preprepared excel file with information on postal code, house number and measurement (model outcome). The excel file is uploaded

[14] Paper demo 3D tool ref 2022-876322534-38276/jscha
[15] Van Walree & Potter van Loon (2020)

in a Shiny app and the user is presented with a graphical overview of the discrimination detection measurements.

Comparing the tools using the same model and user data set shows the following results: Both tools use CBS data to get (extra) information on discrimination. The app for Tool A gives the user a few options. They can choose to see the results of either direct, indirect or total discrimination and between either the measurements of a couple of distinct variables or to just receive an overview of a selected number of most differentiating variables. We decided to continue with tool A and discontinue tool V

Evaluation and user feedback

The second step was to have insurers test the tool in practice. They were provided with the tool and information on how to install, use and interpret the outcome of the tool.

Three middle-size insurers were invited (whereas only two responded in time) to give their feedback and respond to the following questions:

A. Does the code work and does it provide relevant insights?
B. What are the issues a middle-size insurer faces when using the tool; and what is required for insurers to adopt the tool?
C. Do they expect to be using the tool and how often, or is it too complicated, and unhelpful in practice?

Their detailed responses are in Annex II

Despite the need for some improvements, the overall assessment of the bias detection tool is positive. The tool was tested on six pricing models. Four of these models were created in a simulation and the other two are actual models used by an insurance company. The results show that the tool can decompose discrimination into direct discrimination and indirect discrimination. This can be further dissected into categories if the feature is categorical. The tool works well to help insurers explain how their models function in terms of discrimination between the different variables in a given data set. The tool can also show the flows of discrimination through the other variables and in the case of indirect discrimination show whether the models are differentiating. Compared to existing solutions, the tool is easy to use and provides users with more targeted information than similar tools.

There is some need for improvements, most notably in relation to step functions and "noise discrimination" because of random correlation. Furthermore, there remains an issue with the causality implied in indirect discrimination that needs further attention.

When asked if they would adopt the tool as is, the insurers' responses were mixed where the main reason mentioned for not using this tool was that they prefer to keep the development of such tools in-house. Still, the tool has shown to be helpful also to raise awareness in the industry about the need for attention to the risks of bias and unlawful discrimination in using automated decision-making. The project therefore also contributed to raising awareness amongst these insurers to review their models and decision-making processes.

.

The second contribution of the project is to provide recommendations for insurers on making ethical decisions in response to the outcome of the bias detection tool. The next section first discusses several relevant models for decision-making and concludes this report with recommendations for adoption by insurers.

Section 3

Ethical decisionmaking tools

In response to the concerns about the harm caused by AI systems, companies, governments and organisations around the world have been developing principles and guidelines to mitigate and redress any harm caused. For example, The U.S. government has introduced principles specifically addressing fairness and non-discrimination.

"Agencies should consider in a transparent manner the impacts that AI applications may have on discrimination. AI applications have the potential of reducing present-day discrimination caused by human subjectivity. At the same time, applications can, in some instances, introduce real-world bias that produces discriminatory outcomes or decisions that undermine public trust and confidence in AI. [..]."[16]

The EU High-Level Expert Group (HLEG) has published the following set of principles and recommendations for responsible AI.[17]

1. Human agency and oversight 2. Technical robustness and safety 3. Privacy and data governance 4. Transparency	5. Diversity, non-discrimination and fairness 6. Societal and environmental well-being 7. Accountability

The Association of Insurers in the Netherlands has published an ethical framework binding for its member, for data-driven applications.[18] The framework expanded upon the recommendations for responsible AI as proposed by the EU. Members commit themselves not to use any artificial intelligence (AI) or other data-driven products and processes in their relationship with customers that are contrary to the principles. In case an insurer does not act according to the framework, the Association of Insurers will sanction the insurer. Consumers who face discrimination can also file a complaint with the Dutch Financial Services Complaints Tribunal (KiFiD), which is an independent complaint handling body for financial services.[19]

[16] Practical guidance on how to do this was not available until recently with its AI Bill of Rights Note that this was only published after much of the report had been written so we could not take this into consideration.
[17] In the US, "automated decision system impact assessments" have been proposed by Congress as part of the Algorithmic Accountability Act of 2019.
[18] Toolkit Ethisch Kader, available online at https://www.verzekeraars.nl/publicaties/
[19] https://www.kifid.nl/about/

The framework, principles and standards

The purpose of the Association of Insurers framework is to ensure the development and use of safe and reliable data-driven applications in the Dutch Insurance sector. It aims to do so by providing insurers tools for an ethics assessment of their intended AI applications and by giving consumers the confidence they need that the use of AI by insurers will be in their best interests.

Requirements for responsible AI[20]	**Sub Requirements**	**Insurance standard**
Human agency and oversight	Use of AI	1. Before insurers use data-driven applications, they carry out an adequate compliance assessment, in which they make a conscious choice with regard to identified risks compared to more traditional techniques and processes. 2. When using data-driven applications such as chatbots, where necessary insurers will mention that the customer is dealing with a system and not a human being, to avoid any confusion or ambiguity.
Technical robustness and safety	Cyber security	3. Insurers will ensure that appropriate security measures are in place for data-driven applications (including data management). 4. Insurers will ensure that data-driven applications are technically safe and robust and that 'self-learning' only takes place under supervision and within a clear oversight framework.
	Fall-back and general security	5. If a data-driven application is not or no longer considered technically safe or robust, insurers will take measures as soon as possible to ensure that the application does comply.
	Reliability and reproducibility	6. Insurers monitor whether data-driven systems in use work in accordance with pre-defined goals, objectives and intended applications.
	Data quality and integrity	7. Insurers will ensure adequate quality (including evaluation of the data quality criteria completeness, correctness, timeliness, adequacy and representativeness) of data and training data used for data-driven applications. 8. When using data-driven applications, insurers make a the well-considered choice on whether or not to use biometric data, data generated from 'affective computing, social media data, web history, IP address and IoT data and will inform customers transparently when required.
	Access to data	9. Insurers will ensure responsible data management and guarantee good data governance.

[20] Adapted from the DAI publication, *Ethical framework for data-driven applications by insurers,*

Privacy and data governance	Respect for privacy and data protection	10. When using personal data for data-driven applications, insurers work in accordance with the General Data Protection Regulation (AVG), the Dutch GDPR Implementation Act (UAVG) and the Code of Conduct for the Processing of Personal Data by Insurers (Gedragscode Verwerking Persoonsgegevens Verzekeraars). 11. Prior to the purchase, development and/or commissioning of data-driven applications, insurers carry out a data protection impact assessment (DPIA) where necessary. 12. Insurers opt for data-driven systems that process as little potentially sensitive data or personal data as possible (data minimisation) and/or offer the possibility to increase privacy through, for example, encryption, the use of pseudonyms, anonymity or aggregation. 13. Insurers provide thorough protection of (training) data from degradation, contamination or hacking.
	Human control	14. Insurers provide adequate training for employees working with data-driven applications, in particular, to prevent 'confirmation bias' (preference for confirmation) and to preserve human autonomy.
	Human supervision	15. In practice, the use of data-driven applications always takes place under adequate human supervision and responsibility, for example by retraining AI where necessary.
		16. New techniques will first be tested in a familiar setting, to see whether margins of error and other risks increase compared to alternative methods and processes.
transparency		17. Before insurers deploy data-driven systems, they consider how to explain the results of the application to customers in the best possible way. 18. When using data-driven applications, human intervention can always be called upon and customers can have the results of an application explained.
Diversity, nondiscrimination and fairness	Prevent unjust bias	19. When violations of fundamental rights, including unjustified discriminatory bias, cannot be avoided or excluded in data-driven applications, insurers will not deploy an application.
	Accessibility and inclusive design	20. When opting to use data-driven systems, insurers pay attention to diversity and inclusiveness, especially for people at risk of exclusion or disadvantage due to special needs and/or a disability.
Societal well-being	Social consequences	21. Insurers will internally monitor the effects of the use of data-driven decision-making for groups of clients.

	Society and democracy	22. Insurers strive to keep as many customers as possible insurable and will inform customers who are more difficult to insure or uninsurable about ways to reduce risks or alternative ways to cover risks.
Accountability	Verifiability	23. Insurers provide an internal control and accountability mechanism for the use of data-driven applications and the data sources used. 24. Insurers promote the knowledge of directors and internal regulators on data-driven applications. 25. Insurers ensure thorough internal communication on the use of data-driven systems.
	Minimisation and reporting negative impacts	26. For all data-driven applications, insurers carry out a risk and impact assessment amongst primary stakeholders. 27. Insurers promote the expertise of their employees working in the field of accountability and control of data-driven systems through an education programme. 28. Insurers ensure an open culture within their company, in which employees are encouraged to make ethical decisions within a sound system where any negative consequences of the use of a data-driven application can be reported and dealt with adequately
	Documentation of considerations	29. Insurers will set out the choices made regarding the use of data-driven decision-making in their internal policy, whereby the decisive factors are made transparent.
	Complaints	30. Insurers inform customers of the possibilities of reporting complaints regarding the use of data-driven applications, first to the company and then to designated dispute resolution bodies.

Each insurer remains responsible for its own practices in adopting and applying the framework. Taking a pragmatic approach, it is recommended that insurers apply the framework to existing applications to understand how the framework works and become able to asses new developments and the potential need for further actions to remain compliant.[21] The proposed Bias detection tool is developed to help insurers with compliance.

Several tools have already been developed to help insurance understand how to comply with the framework in practice. Important to mention is that these tools are not intended to be used as a standard code of conduct but for insurers to gain a better understanding of how to adopt and incorporate the ethical framework within their own decision-making processes.

Ethical decision-making tools: examples

[21] Toolkit Ethisch Kader, p3

The following presents a brief analysis of relevant models that can be adapted by insurers to decide how to respond to the outcome of the bias detection tool and for making ethical decisions. The first example presented here is developed as a generic model, whereas the second example was developed for use by the Dutch insurance industry in the context of AI. Based on an analysis of these models we present some recommendations for insurers.

Data Ethics Decision Aid (DEDA)

The DEDA is one example designed to help decision-makers with their responsible use of data management, models, algorithms and related issues.[22] It aims to do so by making the ethical issues concerning data projects explicit and to help those involved to reflect on and justify their decisions.

To achieve its purpose, DEDA provides a set of tools including a handbook an online survey app and a worksheet to map ethical issues in data projects, document the deliberation process and help improve accountability towards the various stakeholders and the public.
Although insurers may have access to the same information for their risk assessments the outcome can still be different. This can be in part because of the different values and moral theories each insurer holds that influence its decision-making process.[23] Although not designed with the insurance industry in mind, we found The DEDA model is useful for the purpose of helping insurers think about ethical challenges including bias, and to align the impact their decisions may have for consumers with their company values.

The DEDA handbook [..] A biased dataset, model or algorithm can produce results that diverge from the reality it is attempting to describe and represent.[24] Existing biases are sometimes included in interpretations of datasets during data collection, analysis or storage, or in the decisions made on the basis of the data.' The handbook further provides a more detailed description of the different types of biases. Confirmation and in-group bias occur because people are inclined to agree with the most dominant views in a particular group; and selection bias and feedback loops can lead to negative consequences when a project's results, either intentionally or unintentionally, are somehow reused as new data by accident.

The DEDA app is a useful online application for insurers to further gain a better understanding of the different questions to consider when using automated decision-making. The app addresses the different phases of a data project with a set of open questions. Depending on the user's answers, the app responds with new questions, addresses concerns and points to action points. It also documents the user's approach to ethical issues. The report it produces can be used to further scrutinise the ethical compliance of a data project and/or be archived for documentation, transparency and accountability.

The DEDA worksheet addresses the various phases of a project and the different ethical issues that might emerge including bias.[25] It is therefore useful for insurers to use both within their teams and to engage with relevant stakeholders for feedback on how to best respond to the dilemmas they face.

---

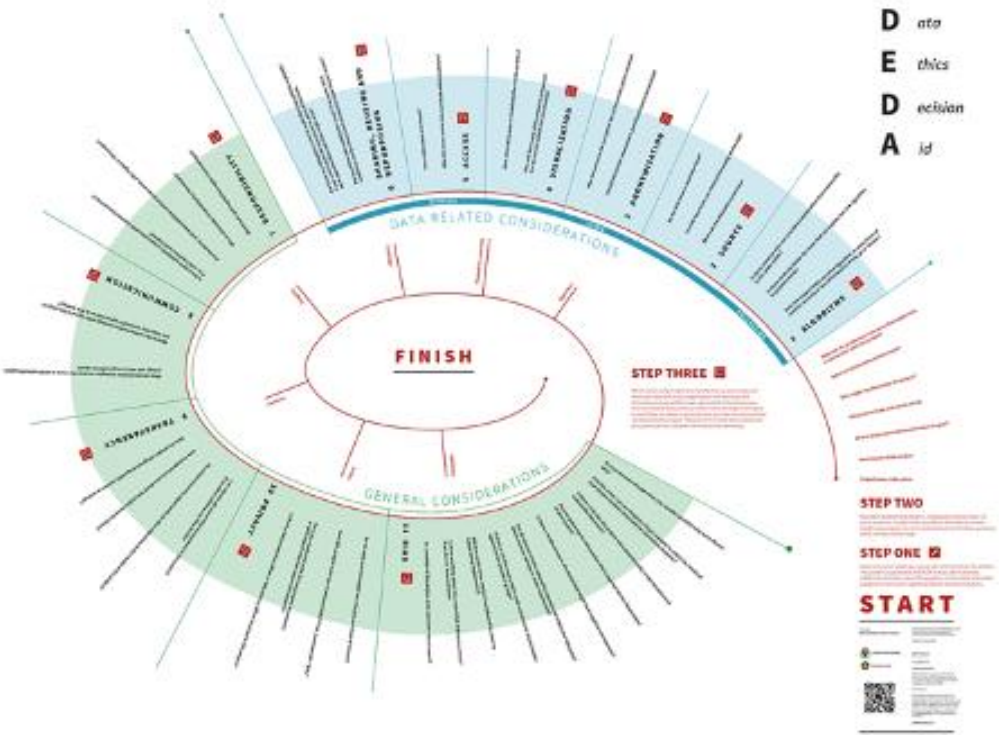
[22] The 'Data Ethics Decision Aid' (DEDA) has been developed by the Utrecht Data School and Utrecht University.. Available at https://dataschool.nl/en/deda

[23] A helpful contribution of the DEDA guidebook is informing the insurers about different moral theories including utilitarianism, relativism, virtue ethics and Kantian theory.

[24] Utrecht Data School, Utrecht University 2020 DEDA - Version 3.1 June 2020 Handbook

[25] Available at https://dataschool.nl/en/deda/worksheet/?lang=en

Image Deda worksheet Utrecht University[26]



The SIVI Checklist for automated decisionmaking

SIVI, the standardisation institute for digital cooperation and innovation, has also developed a checklist for insurers to help them gain insights into the quality of their automated decision-making processes.[27] The checklist can be used a) at the start of a project to identify potential issues; b) during the project for an overview of what actions to take, the responsibilities and status; c) upon completion as a framework for evaluation. The checklist, however, does not address the content of specific domains or products, so it has been extended (the extended checklist) with a specific focus on the adoption and impact of AI within the insurance industry. (see also annexe I)

As with the DEDA, the SIVI Checklist does not provide a qualification of 'right or wrong' but instead gives a structured overview of the most important risks and insights into the mitigation of these risks. The Checklist gives insight into the insurer's internal processes and how it deals with important themes such as laws and regulations, technological risks, testing, traceability and monitoring. By answering the questions, the insurer can get a full picture of their applications with sufficient detail. The Checklist helps

[26] https://dataschool.nl/en/deda/worksheet/?lang=en
[27] https://www.sivi.org/platform-kwaliteit-onbemenste-toepassingen/

insurers to become aware of the issues and can be used as a template for internal and external reports on how the organization deals with the quality requirements of its automated systems.

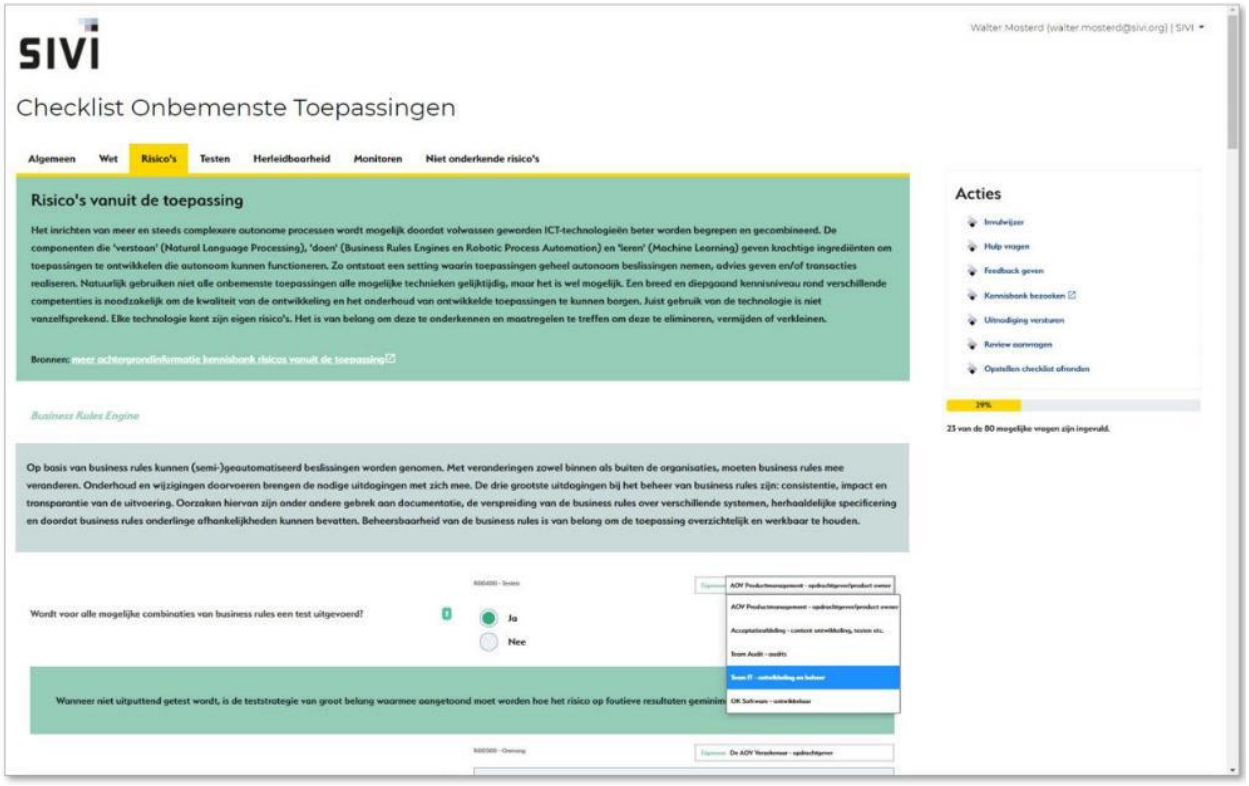


Image of the checklist [in Dutch]

The extended SIVI Checklist

The extended checklist provides an additional list of `checks' that, if answered properly, should test the explainability and transparency of AI model applications, as well as highlight potential weaknesses and areas for improvement. The aim is twofold:

a) It provides insurers insights about the quality and completeness of their AI application with regard to its explainability and transparency. It works as a guide to evaluate if all the relevant elements are in place for a well-designed explainable and transparent AI application. The outcome indicates whether the application meets the quality standards or is incomplete and further actions are required.
b) It provides insurers with the opportunity to share the outcome with third parties (clients or companies) to demonstrate and provide insight into the quality of their AI applications in terms of explainability and transparency.

The additional checklist is relevant to accompany the Algorithm Bias Detection tool we have developed. It helps raise awareness amongst developers and users of AI systems about the risks and need for quality assurance. It triggers the user to consider what further actions are required based on the outcome of the model. Following the perspective that [..] the responsibility of the algorithms' creator does not only

concern its accuracy but also its interpretability and transparency.' the checklist contributes to the growing body of work on Explainable Artificial Intelligence (XAI).[28]

Similar to the proposed guidelines for trustworthy AI, the SIVI extended checklist aim is to give general guidance for the responsible use of AI. The added contribution and value of the checklist are that it makes the outcome actionable for working developers. The checklist has been developed to have this practical relevance, using non-expert terminology and being balanced between being as broad as possible to be appropriate for most users without sacrificing the precision needed to get informative information.

The resulting set of questions is grouped under the subject of transparency, purpose, development, impact, ante- and post-hoc methods, explanations output type, stakeholders, redress, bias and expertise. For the full checklist see Annex II

Upon review of the checklist and the underlying research, it is a good contribution and workable tool for insurers to gain actionable insights into the quality of their algorithms in terms of transparency and explainability. The questions prompt insurers to think about how understandable the inner workings of the applications used are and whether it is possible to see its individual components and interrelations (level of transparency). It further prompts insurers to look at whether it is understandable for the expert user/consumer how the input leads to the output. Taking both of these together the checklist will help insurers to take action to improve the accuracy of their models, justify their use, prevent and uncover biases and help prevent and correct errors. [29]

Further research is needed to look at the lack of diversity and how to improve stakeholder participation from end-customers, consumers and lawmakers in the development of the checklist. [30]
Another concern is the speed with which industry developments take place. Although at the moment there is not much use yet of advanced AI such as Deep Learning within the insurance industry in the Netherlands this may change quickly.[31] The main reasons for this lack of adoption are that insurers consider Deep Learning algorithms compared to machine learning to be less explainable and to require a higher level of expertise to use and understand them; Deep Learning Algorithms are less transferable and given the limited resources insurers find that improving their current processes is more favourable in the short term.[32]

Ethical decision-making framework: Recommendations

For insurers having a policy and value statement to act in the best interest of consumers and society, is not enough for its employees to also understand how to decide in a specific case of conflict. Actionable

[28] Koster et al. (2021) A Checklist for Explainable AI in the Insurance Domain, QUATIC 2021 conference available at https://arxiv.org/abs/2107.14039
[29] Koster et al. (2021) A Checklist for Explainable AI in the Insurance Domain, QUATIC 2021 conference available at https://arxiv.org/abs/2107.14039
[30] research shows for example that most AI audits do not consider multiple stakeholders or the broader social context (Mittelstadt, 2019; Selbst et al., 2018
[31] Koster et al. (2021) A Checklist for Explainable AI in the Insurance Domain, QUATIC 2021 conference available at https://arxiv.org/abs/2107.14039
[32] Koster et.al (2021) A Checklist for Explainable AI in the Insurance Domain, QUATIC conference paper 2021, p 6.

information and practical guidance on how to decide in a specific case that is in line with the company values are often lacking. This leaves the decision-making up to the user of the system who may not be aware of their own biases, or in case of conflicting interests may prioritise their own. In addition documentation on how a decision was made including what stakeholders' interests have been considered and how to justify for example indirect discrimination is also not part of the decision-making process. These issues have made it difficult for independent audits and for third parties to challenge the decisions made by the insurer. This is problematic especially in the case of discrimination to see whether the used justification is lawful.[33]

The proposed Algorithm Bias Detection tool could assist insurers to explain and show

a) what the results of their AI application are used for. The more impact the result and subsequent decision will have on a person's life the more scrutiny there should be of the system. Is the application fully autonomous in terms of decision-making or is the outcome (score) only used by the insurer as advice? If used as advice people will be more likely to notice an error or bias but this does require that the person is in a position to override the outcome of the system without detriment.

b) what *consequences the bias may have and who is most likely at risk.*
How does the insurer mitigate these risks and provides redress? Do these measures outweigh the risk and severity of the potential impact caused when the system does indeed contain a level of bias?

c) enable users to question the decision-making process. It should be possible based on the information that is available to them, for consumers to call out any inconsistencies with the company values the insurer holds.

[33] Koster et.al (2021

Section 4
For discussion

Algorithm Bias Detection: Tool considerations

As more industries including insurance are using AI so do the risks for harm in society. Already examples are plenty where the use of algorithms and automated decision-making has caused harm to people.[34] On the one hand, this is likely to continue to be the case, whereas on the other hand solutions to reduce the risk and avoid certain harms are also being developed.

The tool documented in this report was developed to help insurers detect and measure possible differentiation of insurance pricing models. Insurers can use the tool to explain and motivate their decisions to consumers in compliance with the regulation on automated decision-making.

Because insurers may not want to disclose how their algorithms work for various reasons, however, they must be open and transparent in what are the boundaries within which their decisions remain lawful and ethical or fair. In practice this is difficult for various reasons such as lack of understanding or definition for what constitutes bias or how to apply it to the notion of fairness in any given society.[35]

Proposed further work needs to be done *on incentives* for insurers to adopt tools and enable third-party audits; to become more transparent about their use of AI and automated decision-making as well as to provide information on the impact their policies have on society in terms of insurability and solidarity. *On compliance* and to establish industry guidelines on the adoption of trustworthy AI within the insurance industry to ensure its use is trusted by consumers and other stakeholders in society. *On improving trust, which* continues to be lacking amongst key stakeholders. Given that a lack of consumer trust may lead to people refusing to accept otherwise beneficial innovations for the insurance industry there is an urgent need to improve the trustworthiness of insurers. For insurers to trust is equally important because of the concerns about a breach of confidence and loss of trade secrets. Furthermore improving understanding of how insurance works and what people expect and need is something that is urgently needed also to avoid companies being punished for their transparency which may reveal decisions that although compliant not everyone agrees with.

Algorithm Bias Detection: transparency and information

Information about the use of personal data and consequences for consumers is considered to contribute not only to demonstrating but also preventing unacceptable use of data by insurers and the risks arising from certain innovations in the field of Big data analysis.

[34] M. Hildebrandt, 'Algorithmic Regulation and the Rule of Law' (2018) 376 Philosophical Transactions of the Royal Society A; R. Leenes, E. Palmerini, B.-J. Koops, A. Bertolini, P. Salvini & F. Lucivero (2017). Regulatory challenges of robotics: some guidelines for addressing legal and ethical issues, Law, Innovation and Technology. R. Leenes & F. Lucivero (2014). Laws on Robots, Laws by Robots, Laws in Robots: Regulating Robot Behaviour by Design, in Law, Innovation, and Technology

[35] T. Timmer, I. Elias,
L. Kool & R. van Est (2015). Berekende risico's. Verzekeren in de datagedreven samenleving, Den Haag, Rathenau Instituut;

To protect citizens against the misuse of personal information the General Data Protection Regulation stipulated that the processing of personal data must comply with certain principles, including principle(s) of lawfulness, fairness and transparency.[36] Furthermore, the GDPR aims to empower citizens to take control over 'their' personal data by providing them with certain rights, which include rights to information.[37]

Being well-informed will help consumers to make better decisions about whether to consent to the processing of personal data in exchange for products and services. In the context of insurance, one could argue that consumers need actionable information to decide whether to take out insurance, what coverage they need and what the consequences are. To go even further, we would argue that consumers also have to be informed about how their behaviour may affect the outcome of insurance decisions. Including what they need to do to change their behaviour and run fewer risks with the additional benefit of reducing their premium.

Proposal for the right to game the automated decision-making system

Arguments are therefore put forward for an explanation of the GDPR that provides a so-called '*right to gaming the system'*. The idea is that providing a better understanding of how data is used including how their behaviour influences decisions for example in the context of insurance will actually enable them to make more informed decisions as well as empower them to change their behaviour in such a way that reduces their risk and with that innovations become beneficial for everyone not just for companies seeking to optimize profit.

Article 15(1) of the GDPR states that people shall have access to personal data and information including On the purpose of the processing; on the categories of data and in the case of automated decisions, meaningful information what this means in terms of logic and significance and what the possible consequences are for the citizen. Article 15(3) further provides that, in addition to the right of access, one has the right to receive a copy of personal data being processed, unless disclosure would adversely affect the rights and freedoms of others. How this balancing of interests should be done in practice has not been explained in more detail. In view of the debate that has arisen on the scope of protection and exception based on which a request for information can be refused, it is still unclear what exactly is required under Article 15.[38]

We would argue in favour of the following in the context of insurance: In order to be able to make an informed choice, the consumer is entitled to be provided with information on what personal data is used and how this is analyzed to come to a decision about the consumer.

Recital 63 states that A data subject should have the right of access to personal data which have been collected concerning him or her, [...], in order to be aware of and verify, the lawfulness of the processing. Following the example given for a patient record, a consumer should be given an insight into the insurer's assessment and the risk score and other considerations.

[36] The General Data Protection Regulation (2016/679, "GDPR")
[37] GDPR recital 63
[38] Van den Boom, F (2020)

In practice, we would argue that an insurer who, for example, uses external sources, must not only indicate in general terms that this is the case but upon request must also indicate from which sources exactly what personal data has been received and used so that the consumer is actually able to asses and if necessary correct their personal information as well as to consider whether they agree and consent with this data being used for this particular purpose. An example would be the use of credit scores in determining car insurance premiums. The insurer will have to indicate, on request, which companies they have consulted and what personal data they have on the consumer.

Article 15 gives a right to information on request, but not the obligation for insurers to provide this comprehensive information when motivating any decision. However, this could still be required not on the grounds of the GDPR but as a requirement under an industry code of conduct to which the insurer is held.

Proposal for insurance information to be provided

On the basis of the above, what is advocated here is that more information should be provided to citizens than is currently the case. Consumers should be able to understand at least

- What personal data is used and the origin of this data? This means that not only categories such as credit history but also, the source of the score and what data they hold so that citizens can actually check the correctness of this data.
- Factors and extent(weight) to which these contribute to the decision. In the case of car insurance, there are several factors that insurers consider relevant for estimating the risk of an accident, among other things. It is not only in the interest of the insurer that no damage is caused, but above all in the interest of the citizen that he or she knows what risks he or she is exposed to and how they can possibly change this, for example improving certain driving behaviour or even locate to another area where there is less theft.
- Based on the knowledge and experience of insurers, an explanation of the (possible) consequences of a decision and the behaviour that led to it. An insurer is given their experience and expertise more able to estimate what consequences possible decisions may have for the consumer. As they are able to monitor their acceptance and claims policy and analyse their data. They also have expertise in how the insurance industry works. An example in the context of telematics is that an insurer can cancel the insurance if the conditions are not met, for example, to maintain a safe driving score. This can have a negative impact if citizens try to take out new insurance somewhere else. It is important that this information is shared and how to avoid any negative consequences.
- Any relevant additional information including statistics and research may not be known to the consumer. For example, the influence of certain factors on the risk of damage and theft or insight into human behaviour can also benefit the consumer and enable him to make a better-informed decision about whether or not to take out an insurance policy and under what conditions.

Legitimate reasons for insurers to refuse access

Legitimate interests for insurers include the protection and preservation of trade secrets, the prevention of fraud, whether or not by consumers and/or competitors, and the effort (cost, time, expertise) required to provide personal data and insights.

Given the importance of being able to assess risks as accurately as possible and the factors involved, insurers will be reluctant to provide access to more detailed information.[39] It depends if in practice insurer's interest in not providing specific information outweighs the consumer's interest in receiving such information. We argue that only in exceptional cases it is justified to limit the right for consumers to obtain the information they need to become informed given the importance of insurance in society. Think for example about the impact it has on my people's ability to go to work if they were no longer able to afford car insurance. Being able to challenge a refusal which may be unlawful becomes vital for a person's livelihood.

Closing remarks: A positive outlook on AI in insurance

We would like to end this report with a positive look towards the future. It is sometimes lost in the discussions on how to regulate AI and mitigate its harms, that innovations can be beneficial and help to improve people's lives. We feel therefore that it is important that the positive effects of the use of innovations are also taken into account in any critical assessment of the use of Big Data and AI.

In particular, algorithms can contribute to detecting, reducing and preventing discrimination and inequality in society. In addition, data innovations contribute to the possibility for insurers to offer more adequate premiums on the basis of a better assessment of the risks. It is also expected that as more and more data becomes available and the ability to analyze it, risks for which little or no data was previously available, can now be adequately analyzed so that insurers can reach their aim of making insurance possible and affordable for as many people as possible. This will help contribute to our societies becoming safer and fairer for people.

[39] With regard to providing a copy of the data, the GDPR mentions, among other things that the right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, it indicates that this consideration cannot lead to a situation where no data at all is provided. GDPR recital 63

References

Akerlof GA (1970) The market for “Lemons”: Quality uncertainty and the market mechanism. The Quarterly Journal of Economics 84(3): 488–500.

BBC News “Black boxes: Can you trust them to lower your car insurance? “, 9 November 2016. Accessed online March 2019 at > https://www.bbc.com/news/uk-england-37910773 <

Baker T and Simon J (eds) (2002) Embracing Risk: The Changing Culture of Insurance and Responsibility. Chicago: University of Chicago Press.

Barocas S and Selbst AD (2016) Big Data’s disparate impact.California Law Review 104(3): 671–732.

Barry L and Charpentier A (2020) Personalization as a promise: Can Big Data change the practice of insurance? Big Data & Society. DOI: 10.1177/2053951720935143

Casey et. al, Rethinking Explainable Machines: The GDPR's 'Right to Explanation' Debate and the Rise of Algorithmic Audits in Enterprise (2018). Berkeley Technology Law Journal.

Drechsler L and Benito Sanchez JC (2018) The price is (not) right: Data protection and discrimination in the age of pricing algorithms. European Journal of Law and Technology 9(3): 1–31.

European Insurance and Occupational Pensions
Authority (2019) Big Data analytics in motor and health insurance: A thematic review.

FCA (Financial Conduct Authority) (2019) General insurance pricing practices. Market Study, MS18/1.2.

Gellert R, Vries KD, de Hert P, et al. (2013) A comparative analysis of anti-discrimination and data protection legislations. In Custers B, Calders T, Zarsky T, et al. (eds) Discrimination and Privacy in the Information Society.Berlin: Springer, pp.61–90.

Heath, J. (2007). Reasonable restrictions on underwriting. In P. Flanagan, P. Primeaux, & W.Ferguson (eds.), Insurance ethics for a more ethical world (Research in Ethical Issues in Organizations, Volume 7), (pp.127-159). Emerald Group Publishing Limited,Bingley.

Kahneman, D. (2011). Thinking, fast and slow. Penguin, New York.

Landes, X. (2015). How Fair Is Actuarial Fairness? Journal of Business Ethics , Vol. 128 (no. 3), 519-533.

Lehtonen T-K and Liukko J (2012) The forms and limits ofinsurance solidarity. Journal of Business Ethics 103(S1):33–44.

McFall L and Moor L (2018) Who, or what, is insurtech personalizing?: Persons, prices and the historical classifications of risk. Distinktion: Journal of Social Theory 19(2):193–213.

Meyers G and Van Hoyweghen I (2018) Enacting actuarial fairness in insurance: From fair discrimination to behaviour-based fairness. Science as Culture 27(4): 413–438.

Minty D (2016) Price optimisation for insurance optimizing price destroying value? Thinkpiece Chartered Insurance Institute. Retrieved from https://www.cii.co.uk/learning/ learning-content-hub (accessed 20 July 2020).

Mittelstadt BD, Allo P, Taddeo M, et al. (2016) The ethics of algorithms: Mapping the debate. Big Data & Society. DOI: 10.1177/2053951716679679;

OECD (2020) The impact of big data and artificial intelligence (AI) in the insurance sector. Available at: www. oecd.org/finance/Impact-Big-Data-AI-in-the-Insurance-Sector.htm (accessed 20 August 2020).

O'Neil C (2016) Weapons of Math Destruction. New York:Crown.

Pasquale, F. (2015). The black Box society: The secret algorithms that control money and information. Cambridge, MA: Harvard University Press.

Prince AER and Schwarcz D (2020) Proxy discrimination in the age of artificial intelligence and Big Data. Iowa LawReview 105(1257): 1257–1318.

Rebert L and Van Hoyweghen I (2015) The right to underwrite gender. The Goods & Services Directive and the politics of insurance pricing. Tijdschrift Voor Genderstudies 18(4): 413–431.

Swedloff R (2014) Risk classification's Big Data (r)evolution. Connecticut Insurance Law Journal 21(1): 339–374.

Van den Boom, F. (2020). Regulating Telematics Insurance. in P. Marano, K. Noussia (eds.), Insurance Distribution Directive, AIDA Europe Research Series on Insurance Law and Regulation 3, https://doi.org/10.1007/978-3-030-52738-9_12

Zarsky, T. (2016). The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. Science, Technology, & Human Values, 41(1), 118–132. doi:10.1177/0162243915605575

Zuboff S (2019) The Age of Surveillance Capitalism. London:Profile Books, Kindle Edition.

Zuiderveen Borgesius, F. (2015) Improving privacy protection in the area of behavioural targeting, Alphen aan den Rijn: Kluwer Law International.

Annex I User Feedback

On question A Use

- Insurer X had some trouble with getting the tool to function properly. They noted that they couldn't get the code working with their own data and that the Shiny tool didn't work with the latest R version.
- Insurer Y noted that using a dataset with all available variables (99 variables), the tool will eventually crash after about 2 hours of running. This could be a memory issue, however, by making the choice for which variable is included, they were concerned that this is at the expense of the objectivity of the results. Ideally, you want to use the same dataset that was used for the model itself.
- The Shiny tool distinguishes between direct and indirect differentiation. There are enough perspectives to look at the results in detail. This is one insurer noted: [*..]makes us think and inspires us to further investigate (especially indirect differentiation) in our models.*
- The calculation of indirect differentiation assumes that the correlation between variables also has a causal relationship. In practice, however, this is not always the case.
- There were questions about whether it is possible to test multiple models at the same time on multiple datasets at the same time. If so how would the results be interpreted?
- Insurer X was unsure whether the tool would work with models such as an XGBoost, Neural network and survival analysis, and various model frameworks in R such as tidy models and caret. because it seemed to be made to work only on classic models like GLM.

On Question B Adoption

- Again there are technical problems which according to Insurer X relate to the question of whether to make the tool available to members by hosting it online. This would help solve technical issues but increases the risk of data sharing. Sharing data with third parties is often not allowed due to security and privacy regulations. Instead, the tool could be made available to insurers to be used on-premise. This raises other issues including package and version management within R.
- Insurer X recommended rewriting the R-code to improve readability. They found the data structure illogical and unclear. Although they assumed the tool worked, debugging and understanding what was happening is far from easy. (Note: reading ch. 10 in Description differentiation detection device.pdf does help with this). The confusion arises partly due to the use of the package data, and table and partly due to the design of the code. Recommendations include looking at the use of the tidyverse in combination with nested data frames to better structure the data and code. And to develop our own package for the differentiation detection tool.
- The tool was not considered to be very intuitive according to insurer X, making it necessary to read through the accompanying documentation carefully to understand what you are looking at. This ensures that the use is limited to specialists in the field of the model, such as an actuary or data scientist. Insurer X would therefore not recommend that the tool be used by employees with insufficient modelling knowledge. The data scientist/actuary must make a translation himself in order to communicate the results in an understandable way and to discuss them with the business. An info button in the tool could help with this. Insurer Y on the other hand did not have problems installing the tool although they did need to make some minor adjustments. They also encountered some errors but found a way to work around them.
- About the Shiny App, Insurer C commented on the design and recommendations to pay attention to: What does the user want to know and is the right information shown at the right time? (UI follows the function principle). A clearer flow/order in the tool would make it easier to find relevant information. And an export function was recommended because capturing the results is now only possible by taking screenshots.

On Question C Relevance

- Insurer X found that the tool provides interesting insights that can help develop fair and transparent models.
- Within their current modelling process, Insurer X noted that they already pay attention to, among other things. Fair AI and explainability. Zooming in further on indirect differentiation is considered to be an addition to this.
- Insurer X noted they prefer to develop a methodology themselves to be tailored to their own needs and the differentiation detection tool can be a useful example of how to do this.
- Although Insurer X said they would not be adopting the tool it does give them a reason to improve their own modelling process. Insurer Y said they would be interested to adopt the tool but to work together to improve its useability.

In addition, insurer X mentioned they would monitor the market developments by providing tools such as the fair model's package in R with which various fairness checks can be carried out.

Annex II SIVI

Transparency & explainability

Transparency refers to how understandable the inner workings of the application are and how well its individual components and their interrelationships can be viewed. By **explainability**, we mean the way an application conveys to the user (from customer to expert) in a human-understandable way, how the input leads to the results (i.e. output), which in turn can lead to novel or confirming insights about the model and its dataset. Both aspects ensure that the application as a whole becomes more robust. It can help increase the accuracy of your model, justify its functionality, prevent unwanted biases, uncover new knowledge and help prevent and correct errors.

#	Subject	Check	Answer type	Elucidation
1	Transparency	Elaborate whether the application itself is already transparent to the user or whether external techniques are needed to increase transparency and explainability?	Open answer	We can divide AI algorithms into two categories. Namely, *transparent algorithms* and *non-transparent algorithms*. Transparent algorithms include *rule-based systems*, *linear/logistic regression*, *decision trees*, *k-nearest neighbours*, *rule-based learners*, *general additive models*, *bayesian models*. Non-transparent algorithms include *tree ensembles*, *support vector machines* and *neural networks*. It is also possible that your application falls somewhere between the two extremes.
2	Purpose	Which of these reasons is most important with regard to the application's explainability. You can choose more than one option.	Multiple choice	The need to explain a model mainly stems from one of four reasons: 1) increasing the model's accuracy. 2) discovering or confirming causality. 3) verifying and justifying the model's fairness & robustness. 4) Checking the model for errors and removing bugs.
3	Development	Explain how the right balance has been found between accuracy and explainability in the development process and where priority has been placed.	Open answer	AI applications are often developed with the highest possible accuracy as the main priority. In such a case, a compromise is usually made between accuracy on the one hand and transparency on the other. In practice, explainability and transparency are often just as important and sometimes even more important for the use of an application.
4	Impact	1) Explain what the results of the application are used for (i.e. what role do the results fulfil). Are the results advisory to the user or does the application make autonomous decisions based on these results, or something in	Open answer	The application's impact on the organization is influenced by the role that the application's result must fulfil. If the application plays an advisory role, the result's impact on the end customer is relatively less than if the results play a decisive role (meaning the application makes autonomous decisions based on the results). If the application makes autonomous choices and implements incorrect logic, incorrect decisions can go unnoticed. After which they can only be corrected
5	Ante-hoc methods	1) What external ante-hoc techniques have been used to improve explainability? 2) Explain how these techniques increase the explainability of the application.	Open answers	There are external techniques that can increase the transparency and explainability of AI applications. Some of these techniques involve baking in explainability from the beginning. This has to do, for example, with paying extra attention to input processing or training on the dataset. These techniques are called *ante-hoc techniques*. Think of techniques such as *Reversed Time Attention Model (RETAIN)*, *Bayesian deep learning (BDL)*, etc.
6	Post-hoc methods	1) What external post-hoc techniques have been used to improve explainability? 2) Explain how these techniques increase the application's explainability.	Open answer	There are external techniques that increase an application's explainability after or during the model run-time. These techniques are called *post-hoc techniques*. There are post-hoc techniques that are universal for all algorithm types, but also for specific algorithms types. Think of techniques such as *Local Interpretable Model-Agnostic Explanations (LIME)*, *SHapley Additive exPlanations (SHAP)*, *Layer-wise Relevance Propagation (LRP)*, etc.
7	Explanation output type	Explain what type of explanation is outputted?	Open answer	The outputted explanation of the application can be given in several types. The possible options are *textual, numeric, categorical, pictorial, time series or rule-based.*
8	Stakeholders	For each relevant stakeholder, explain... 1) what demands they have concerning the application's explainability and 2) how these interests are fulfilled by the application's explanations:	Open answer	Different people in and outside the organization have different demands concerning the application's explainability. The organisation must take these different demands into account, so that an application is as robust and usable as possible, while also considering ethical concerns and legislation. (stakeholders: *creator, examiner, operator, executor, decision-subject, data-subject*) (demands: *informativeness, causality, trustworthiness, confidence, accessibility, interactivity, transferability, privacy-awareness*)
9	Redress (Question and complaint handling)	Explain how the consumer receives more information when he or she has an in-depth question or complaint regarding	Open answer	If the application affects consumers, they may sometimes need further explanation about their situation. In such a case, the application's results may have to be explained to this person. This can happen, for instance, if the application makes a mistake, or the consumer has a
10	Bias	Explain... 1) which end-customer groups may be unfairly disadvantaged by the application and how this is prevented. 2) whether the explanations provided reveal (unwanted) biases (in the data or the algorithm)?	Open answer	Undesired biases can arise in applications that are trained on data. These unwanted biases must be actively prevented. Explainability and transparency can be a means to that end. Example of bias: Statistically, red cars take more damage. However, does this mean that red car owners have to pay a higher premium?
11	Expertise	Explain whether new expertise concerning explainability and ethics were needed in the company, since the implementation of the application?	Open answer	Some companies hire 'explanation experts' and 'ethics experts' to support their AI projects.