THE GENEVA ASSOCIATION

INSURANCE FOR A BETTER WORLD

Promoting Responsible Artificial Intelligence in Insurance

January 2020

The Geneva Association

The Geneva Association was created in 1973 and is the only global association of insurance companies; our members are insurance and reinsurance Chief Executive Officers (CEOs). Based on rigorous research conducted in collaboration with our members, academic institutions and multilateral organisations, our mission is to identify and investigate key trends that are likely to shape or impact the insurance industry in the future, highlighting what is at stake for the industry; develop recommendations for the industry and for policymakers; provide a platform to our members, policymakers, academics, multilateral and non-governmental organisations to discuss these trends and recommendations; reach out to global opinion leaders and influential organisations to highlight the positive contributions of insurance to better understand risks and to building resilient and prosperous economies and societies, and thus a more sustainable world.

The Geneva Association—International Association for the Study of Insurance Economics
Talstrasse 70, CH-8001 Zurich
Email: secretariat@genevaassociation.org | Tel: +41 44 200 49 00 | Fax: +41 44 200 49 99

Photo credits:
Cover page—Carlos Castilla, Shutterstock.

January 2020
Promoting Responsible Artificial Intelligence in Insurance

Published by The Geneva Association—International Association for the Study of Insurance Economics, Zurich.

Promoting Responsible Artificial Intelligence in Insurance

Benno Keller

Contents

Foreword

1. Management summary 4

2. Benefits of AI in insurance 6

3. Responsible AI in insurance 9
   3.1 Transparency and explainability 10
   3.2. Fairness 12

4. Recommendations 15

5. Glossary 17

6. References 19

Acknowledgements

This publication is a product of the New Technologies & Data work stream of The Geneva Association, sponsored by Thomas Buberl, Group CEO of AXA.

We are very much indebted to the members of the Working Group established in support of our New Technologies & Data research activities, namely: Paul DiPaola and Evan Hughes (AIG), Henning Schult (Allianz), Ashley Howell (Aviva), Patricia Plas, Chaouki Boutharouite and Dora Elamri (AXA), Atsushi Izu (Dai-ichi Life), Bruno Scaroni (Generali), Chris Reid (Intact Financial), Hiroki Hayashi (Nippon Life) and Lutz Wilhelmy (Swiss Re).

We would also like to thank the following experts who made themselves available for a dedicated workshop on the topic of this report: Ahed Abdelky (Allianz), Marcin Detyniecki and Sarah El Marjani (AXA), Alberto Branchesi (Generali), Bo Gong (Ping An Technology Research Institute) and Achraf Louitri (Intact Financial). Finally, we would like to thank Michele Loi of the University of Zurich for his very helpful insights and inputs.

Foreword

Artificial intelligence (AI) already shapes our everyday lives, from controlling the news and advertisements we see online to selecting the GPS routes we take. Entire sectors, such as education, medicine, law and finance are being shaped by its influence.

In insurance, AI can potentially lower costs, smooth claims processes and make customer interactions more efficient. Products like AI-powered parametric insurance can help to insure more people. AI can also enhance risk prevention, management and mitigation abilities and diminish problems long endemic to the industry, like moral hazard and fraud.

However, the benefits of AI will be diluted if customers do not trust that insurers use technology and data responsibly.

This Geneva Association report explores the principles we see as most relevant for achieving the necessary levels of customer trust: transparency, explainability and fairness. We hope this analysis—and our corresponding recommendations for insurers—contribute to the responsible use of AI in insurance and the actualisation of its full benefits for society.

Jad Ariss
Managing Director
The Geneva Association

1. Management summary

Since the end of the 'AI Winter', the period from the 1970s to the end of the 1990s characterised by setbacks and disappointment, artificial intelligence (AI) has made remarkable progress. Today, AI systems are commercially used in a growing field of applications.

Many insurers are rolling out intelligent systems that automate routine tasks or assist human decision-making along the entire insurance value chain. Such systems combine new types of learning algorithms with the analysis of data from new types of data sources, such as online media data and Internet of Things (IoT) data (The Geneva Association 2018).[1] In the future, intelligent systems will autonomously take standardised decisions in a growing number of areas.

The use of AI in insurance has the potential to yield economic and societal benefits that go beyond insurers and their customers by improving risk pooling and enhancing risk reduction, mitigation and prevention.

In order to foster the adoption of AI systems and realise these benefits, insurers need to earn the trust of their customers by using the new technology responsibly.

1 Therefore, intelligent systems in the broad sense are technologies that assist or replace human decision-making (Monetary Authority of Singapore 2018 and BaFin 2018).

In recent years, an intense debate has developed on what the responsible use of AI entails, and there has been a proliferation of ethics guidelines issued by governmental and non-governmental organisations over the past 12 to 24 months. The question of what the responsible use of AI means and how it should be ensured is the subject of an ongoing debate.

An analysis of guidelines for the ethical use of AI suggests that there is a global convergence towards five core principles: (1) *transparency and explainability*, (2) *fairness*, (3) *safety*, (4) *accountability* and (5) *privacy* (Jobin et. al. 2019).[2] There are, however, critical differences in how these principles are interpreted as well as what requirements are considered necessary for their realisation. In addition, considerable uncertainty remains regarding how ethical principles and guidelines should be implemented in a specific context (Jobin et. al. 2019).

A definitive answer to these questions will remain elusive. This report, however, aims to contribute to identifying and discussing key trade-offs that arise when implementing core principles for responsible AI in insurance. For example, enhancing the interpretability of complex models often comes at the cost of reduced benefits, and overly complex models may have limited additional benefit.

While all five core principles for responsible AI are critical, this report focuses on the two principles that raise particularly complex issues in insurance: 1) transparency and explainability and 2) fairness.[3]

Implementing the principle of fairness in insurance requires trade-offs that do not typically arise in other industries. Mitigating bias and discrimination is particularly challenging in insurance, with different, mutually exclusive standards of fairness.

Ensuring a balanced assessment of the benefits and risks of specific uses of AI requires a clear assignment of roles and responsibilities within an organisation as well as expertise and experience in data science, actuarial science, risk management and data protection.

Based on our research, the report concludes with three recommendations for insurers to promote the responsible use of AI within their organisations: 1) establishing internal guidelines and policies, 2) adopting an appropriate governance structure to address related risks and 3) developing and rolling out comprehensive training programmes for employees and agents. Although there are many different governance models, insurers can draw on existing risk management and actuarial frameworks.

2 Jobin et. al. (2019) use the terms 'transparency', 'justice and fairness', 'non-maleficence', 'responsibility and accountability' and 'privacy' for the five core principles.

3 The focus on transparency and explainability and on fairness does not imply that the principles of safety, accountability and privacy are of lesser importance. As a matter of fact, the safety of AI systems and the preservation of the privacy of customers may be seen as a foundation for the fair use of AI. For a discussion of privacy issues arising with big data analytics in insurance, see The Geneva Association 2018.

2. Benefits of AI in insurance

Today, intelligent systems can perform tasks that are particularly useful in insurance. For instance, progress in natural language processing allows intelligent systems to 'talk' and interact with humans. Insurers are increasingly using conversational agents (e.g. chatbots) that can identify and respond to complex customer queries and are available 24/7 (see Box 1).

Box 1: Conversational agents in insurance

Conversational agents are deployed in various lines of business and different parts of the value chain. They allow customers to interact with their insurer 24/7 via online chat, a channel preferred by younger customers in particular.

Using deep learning techniques to classify input in natural language, conversational agents can identify and respond to complex customer queries, e.g. related to health and car insurance. As a result, they are able to considerably enhance customer experience as well as increase the insurer's efficiency by processing large amounts of customer requests. Furthermore, conversational agents allow insurers to engage in new types of communication with their customers.

Conversational agents may perform different types of roles, from guiding users to the information they need and coaching users through insurance-related procedures (e.g. submitting a claim) to actually processing business transactions. They are beneficial not only to customers but also to service agents, who are able to focus on tasks where human judgement is key.

To provide personal answers to queries submitted by users, conversational agents may use internal product information as well as non-personal, third-party information, e.g. related health expertise.

Insurers have designed conversational agents to ensure reliability and user privacy.

Furthermore, intelligent systems can 'view' and recognise objects in pictures and extract related information. Such computer vision allows insurers to automate manual and cognitive routine tasks, e.g. to extract data from written documents and pictures for use in the underwriting or claims process (see Box 2).

Box 2: Computer vision in insurance

Several insurers use computer vision to automate routine tasks in underwriting and claims management by extracting information from documents and pictures using machine learning and deep learning techniques.

For example, computer vision is used on existing documents to validate and verify information provided by the customer during the underwriting process. This includes the verification of pictures provided by customers in car insurance applications, e.g. indicating the type of car and identifying customers through their licence plates. Computer vision can thus help to ensure the appropriate level of coverage as well as to identify insurance fraud.

In claims, computer vision is used to validate the authenticity of images provided by customers and to extract information from documents such as accident forms as a basis for claims triage and to automate claims processes.

For example, one insurer uses machine and deep learning to understand the document type and to extract information from medical documents (including bills, prescriptions and other documents) photographed and submitted by policyholders of health insurance plans. The system identifies the medical treatment and the diagnosis, extracts all the medical bill data (amount, date, VAT number, fiscal code, receipt number) and within seconds, matches the information with the applicable insurance cover of the policyholder.

Computer vision is also used on new types of data. For example, one insurer uses deep learning to recognise and measure the size of plantations of different crops such as wheat, corn and rice from satellite images, using this as a basis to offer crop insurance. Computer vision is also used to improve disaster response by identifying and locating customers affected by natural disasters.

Intelligent systems excel in detecting patterns and correlations in complex data in ways that are very difficult, if not impossible, for humans. The identified patterns are the basis for analytical tasks such as classification, regression and clustering that play an important role in the insurance business model. Compared to traditional modelling approaches used in insurance (such as generalised linear models), intelligent systems have the potential to provide much more accurate predictions, because they can learn complex non-linear relationships between variables. Today, these capabilities of intelligent systems are used to assist human decision-making (see Box 3).

Box 3: The use of AI to assist decision-making

AI is used to assist sales agents in decision-making and enable them to offer more personalised customer service by deriving insights from customers' internal data, such as product and claims data and location, as well as from customer interaction. Agents receive recommendations for cross- and upselling opportunities and related product offerings, while the decision to approach customers with specific product offerings remains with the agent. Such tools have proven very effective in enhancing the efficiency of the sales channel.

One insurer uses machine learning to predict claims as a basis for determining the optimal premium rates for existing and new customers in auto insurance. The algorithms are applied to traditional data used in underwriting auto insurance and provided by customers, including type and make of car, age and claims history. Compared to traditional pricing models, the enhanced algorithms considerably increase predictive accuracy.

Insurers use various measures to avoid bias and ensure fairness in their applications that assist human decision-making (see section Fairness).

Although intelligent systems are not yet widely used by insurers to fully automate decision-making, further progress in learning algorithms may in the future enable these systems to automate standardised decision-making in a growing number of areas under human supervision. For example, AI could be used to automate the underwriting approach of standardised and homogeneous areas of risk (SCOR 2018). McKinsey expects manual underwriting to be extinct by 2030 for most personal and small-business products across life, property and casualty insurance (McKinsey 2018).

The potential benefits of such uses of AI go well beyond insurers and their customers (see Figure 1). For instance, applications of AI can help to extend insurance cover to new and previously uninsured or underinsured customer groups or to expand the range of risks for which insurance cover is available. In doing so, these uses of AI allow the expansion of the scope of risk pooling, which lies at the core of the economic and societal role of insurance. At the same time, however, there is an important concern that the increased personalisation of insurance enabled by AI could lead to the exclusion of specific groups of customers, for example those considered high risk (see section 'Fairness' p. 12, and The Geneva Association 2018).

The use of AI may also reduce the cost of risk pooling by automating specific tasks, increasing the accuracy of risk assessments or reducing moral hazard and adverse selection.[4]

Furthermore, the use of AI could lead to novel risk insights that may help to mitigate and prevent risks. The use of AI could encourage risk reduction by better aligning premiums and risk. Moreover, enhanced data would facilitate the establishment of advanced risk management and early warning systems that allow for timely interventions to reduce losses. The use of AI may thus help to extend the role of insurance from pure risk protection towards 'predicting and preventing' (The Geneva Association 2018).

Figure 1: Socio-economic benefits of AI for Insurance

Expand the scope of risk pooling
• Extend insurance cover to new and previously uninsured customer segments by facilitating access to personalised products (e.g. life insurance for individuals with pre-existing conditions) • Expand the range of risks for which insurance cover is available through improved risk insights (e.g. cyber risks)
Reduce the cost of risk pooling
• More cost-efficient insurance through the automation of specific tasks, better risk assessments and reduction of moral hazard and adverse selection
Mitigate and prevent risks
• Novel risk insights that help mitigate and prevent risks • Early warning systems that enable the reduction of losses

Source: The Geneva Association

Intelligent systems do, however, raise several challenges. For instance, they need large amounts of data to learn, and their learning is only as good as the data used to train them, so any potential biases in the data will be learned by the system.[5] Moreover, due to their complexity, it may be difficult to understand why a system has reached a particular decision, since such systems are difficult to interpret. Much like humans, intelligent systems can make errors even when the data is not biased. Known problems in computer science that can lead to errors are *overfitting* or the *curse of dimensionality*, for example.[6] In such situations, patterns learned from the data cannot be generalised.

Human decision-making is not free from bias and errors. However, in contrast to decentralised human decision-making, the use of AI for autonomous decision-making at scale implies that even a small systematic error may have far-reaching consequences (Ruf et al. 2019a).

Concerns about unintended consequences and about the malicious use of AI have triggered an intense debate on its responsible use. The next section discusses key principles for responsible AI that have emerged from this debate and identifies important considerations for the implementation of those principles in insurance.

4 Moral hazard occurs when individuals increase their exposure to risk because they know they do not bear the cost of those risks. Adverse selection occurs when individuals have better information than their insurer about their risks. In this case, high-risk individuals tend to buy insurance, while low-risk individuals do not. This can trigger an adverse market dynamic with increasing premiums and declining cover.

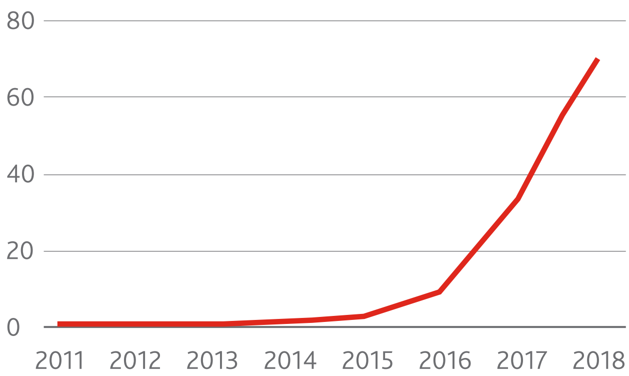
5 An illustrative instance of this effect is Tay, a chatbot by Microsoft that quickly turned racist and sexist (Metz 2016).

6 One examples of overfitting was the failure of Google Flu, a system to predict the occurrence of flu based on peoples' searches on Google (Lazer and Kennedy 2015).

3. Responsible AI in insurance

Over the past two years there has been a proliferation of guidelines for responsible AI issued by governments, international organisations, regulatory authorities, academic institutions, industry bodies and companies (see Figure 2). AlgorithmWatch has catalogued over 80 such guidelines in its global inventory of AI ethics guidelines.[7]

Figure 2: Number of AI ethics guidelines



Source: Jobin et. al. 2019

About one third of these guidelines have been published in the U.S., another third in the EU (excluding the U.K.), and about one fifth each in the U.K. and Asia (in particular in Japan).

As voluntary commitments, such guidelines postulate general principles for the responsible use of AI.[8] While there are substantial differences between the guidelines, there seems to be a global convergence towards five core principles: safety, accountability, privacy, transparency and fairness (Jobin et. al. 2019).[9]

Existing guidelines typically do not discuss how to address trade-offs that arise when these principles are applied in practice.[10] It is also worth mentioning that published guidelines primarily focus on how to preserve key values, rather than on how AI could contribute to the advancement of these values (Jobin et. al. 2019).

7 https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/.

8 Guidelines also vary considerably in their degree of specification. For example, the Microsoft AI Principles consists of 51 words, while the guidelines of the IEEE Standards Association comprise more than 260 pages.

9 Jobin et. al. (2019) use the terms 'transparency', 'justice and fairness', 'non-maleficence', 'responsibility and accountability' and 'privacy' for the five core principles.

10 With the exception of the guidelines issued by the EU High-Level Expert Group (European Commission 2019), guidelines in general do not acknowledge the existence of trade-offs when implementing principles for responsible AI.

In insurance these core principles are not new and have long played an important role. It goes without saying that insurance products should be safe, that they should protect customer data and that insurers remain accountable to their customers. As a matter of fact, various laws and regulations—including insurance law, privacy and data protection laws, anti-discrimination laws, and supervisory requirements—govern the fair, transparent and accountable behaviour of insurers as well as the protection of privacy.

Nevertheless, the use of AI raises some intricate questions. Which trade-offs arise with the implementation of core principles? How can insurers foster and demonstrate adherence to such principles, and what changes, if any, are necessary to existing governance mechanisms and risk management frameworks for this purpose? Insurance regulators are also increasingly asking such questions (see e.g. BaFin 2018).

In what follows, the principle of transparency and explainability and the principle of fairness are discussed in more detail, based on an analysis of eight guidelines issued by various governmental and non-governmental actors (see Box 4). The focus on transparency and explainability and fairness does not imply that safety, accountability and privacy are of lesser importance. However, as we will discuss below, the principle of transparency and explainability and the principle of fairness raise issues that are specific to insurance.[11]

Box 4: Eight guidelines

- **Ethics Guidelines for Trustworthy AI** by the High-Level Expert Group on AI set up by the European Commission (European Commission 2019)
- **Everyday Ethics for Artificial Intelligence** by IBM (IBM 2019)
- **Ethically Aligned Design** by the Institute of Electrical and Electronics Engineers (IEEE 2019)
- **Responsible Machine Learning Principles** by the Institute for Ethical ML (The Institute for Ethical ML 2019)
- **AI Principles** by Microsoft (Microsoft 2019)
- **OECD Council Recommendation on Artificial Intelligence**, adopted on May 22, 2019 (OECD 2019)
- **Principles to Promote Fairness, Ethics, Accountability and Transparency (FEAT) in the Use of Artificial Intelligence and Data Analytics in Singapore's Financial Sector** by the Monetary Authority of Singapore (Monetary Authority of Singapore 2018)
- **How to Prevent Discriminatory Outcomes in Machine Learning** by the World Economic Forum (World Economic Forum 2018)

3.1. Transparency and explainability

Transparency and explainability is a key and core principle in all the guidelines analysed.[12] They are considered important to building trust with customers and other stakeholders. Guidelines usually demand that individuals should be empowered to understand the reasons behind decisions and the consequences that affect them. Some guidelines also mention the importance of transparency and explainability to enable individuals to seek redress against decisions affecting them (European Commission 2019).[13] Providing an explanation is particularly important when a decision has a significant impact on the affected individual. Therefore, the degree to which explainability is needed is highly dependent on the context and the severity of the consequences when an output is erroneous or otherwise inaccurate (European Commission 2019).

Interpretability of algorithmic outcomes—understood as the 'ability to explain or to provide the meaning in understandable terms to a human' (Doshi-Velez and Kim 2017)—is not only important to provide meaningful explanation to affected individuals; it is also indispensable for assessing the performance of AI systems and for their continuous improvement, and thus for sound data science. Moreover, interpretability of algorithmic outcomes helps to build confidence in the system's ability to make accurate predictions. Insurers should therefore strive to enhance the interpretability of their AI systems, particularly if these have a significant impact on individuals.

Published guidelines do not specify in detail what should be disclosed to whom, and there are substantive differences between the guidelines as to what transparency entails. Generally, however, ex ante disclosures and ex post explanations can be distinguished.

11 The safety of AI systems and the preservation of customer privacy may be seen as the foundation for a fair use of AI. For a discussion of privacy issues arising with the use of big data analytics in insurance see The Geneva Association 2018.
12 Guidelines also use the terms 'interpretability', 'explicability' or 'understanding'.
13 See section 'Fairness' p. 12.

All the guidelines analysed suggest some form of ex ante disclosure to users. For example, several guidelines emphasise that individuals should be made aware that they are interacting with an AI system such as a chatbot (e.g. IBM 2019, OECD 2019, Monetary Authority of Singapore 2018). Several guidelines demand that the use of AI in the decision-making process be declared to users (e.g. OECD 2019, Monetary Authority of Singapore 2018 and World Economic Forum 2018). To build trust with users, there are guidelines that further suggest that such disclosures should include a description of the capabilities and purpose of AI systems (European Commission 2019) or that organisations commit to fostering a general understanding of AI systems (OECD 2019).

Furthermore, the guidelines analysed require that decisions made by AI systems be explained in understandable terms to those affected (OECD 2019, World Economic Forum 2018). Some guidelines state that such ex post explanations should be provided to affected individuals on request (Monetary Authority of Singapore 2018). To enable individuals to understand decisions that affect them, some guidelines require that the logic involved in decision-making be explained (OECD 2019, World Economic Forum 2018).[14] This may include clear explanations on which data are used, how the data affect the decision and the consequences of the decision (Monetary Authority of Singapore 2018, European Commission 2019).

Several guidelines emphasise that the information provided to individuals has to be meaningful to them. For example, providing a complex explanation of the algorithm is hardly meaningful to individuals (European Commission 2019). To be meaningful, explanations provided to users must serve a clear purpose. Wachter et al. 2017 identify the following aims of meaningful explanations:

1. To inform and help the subject understand why a particular decision was reached
2. To provide grounds to contest adverse decisions
3. To understand what could be changed to receive a desired result in the future.

Providing meaningful explanations is a challenge, as some 'black box' algorithms are by nature complex—the price to pay for better accuracy—and therefore difficult to interpret and explain. In such systems it is usually not possible to interpret and explain the role of the different variables in general.[15]

In recent years, considerable efforts have been undertaken in computer science to overcome challenges of interpreting and explaining 'black box' algorithms. Reverse engineering approaches, for example, consist of building interpretable algorithmic surrogates[16]—a recent technique which needs to be better understood (Ruf et al. 2019a). Design approaches rely on imposing certain constraints on the predictions (Guidotti et al. 2018 and Hall et al. 2017).[17]

When it is not possible to provide an explanation on the role of different variables in an individual decision, other types of explanations may be used (see Figure 3).

Figure 3: Options for ex post explanations depending on significance of impact

Significance of impact on users

Type of explanation	Description	Example
General logic of decision-making	Providing a qualitative understanding of the relationship between the input variables and the model's prediction.	«The car insurance premium is based on age, type of car and a risk score calculated based on speed, acceleration and breaking severity»
Logic of individual decision	Providing a qualitative understanding of the key factors that drove the decision.	«Your premium has increased because your risk score based on speed, acceleration and breaking severity has increased»
Counterfactual explanation	Explanation of the form «If X had not occurred, Y would not have occurred»	«Your home insurance would not have been denied if you had installed storm shutters»
Certification / independent audit	Certification by an independent body	«Our rating model has been certified to be accurate and fair»
Disclosing results of independent audits of algorithms to the public	Full disclosure of report of independent audit	

Source: The Geneva Association

14 In the EU, to provide 'meaningful information about the logic involved' in automated decisions is a legal requirement (General Data Protection Regulation).
15 The EU General Data Protection Regulation (GDPR) acknowledges these challenges by stating that an explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) may not always be possible. According to the GDPR, other explicability measures (e.g. traceability, auditability and transparent communication on system capabilities) may be required in those circumstances, provided that the system as a whole respects fundamental rights.
16 Algorithmic surrogates can be understood as approximation models that mimic the output of the algorithm and are easily interpretable.
17 For example, humans easily understand relationships between variables that only change in one direction (risk increasing with age, for example) (Hall et. al. 2017). Traditional modelling techniques used in insurance, such as generalised linear models, typically produce such relationships and are therefore relatively easy to interpret. Imposing constraints on an AI system to produce so-called monotonous relationships may therefore enhance its interpretability at the cost of reduced accuracy.

For example, counterfactual explanations have been proposed to address the information rights of individuals under the GDPR (Wachter et al. 2017). A sample counterfactual explanation is, 'You were denied a loan because your annual income was £30,000. If your income had been £45,000, you would have been offered a loan.' Counterfactual explanations of complex 'black box' decisions, however, may not always be reliable and should satisfy certain statistical criteria (Laugel et al. 2019). The quality of ex post counterfactual explanations should therefore be monitored.

Conclusions and discussion

Interpretability of algorithmic outcomes is important to reinforce customer trust and understanding. Perfect interpretability, however, is often difficult to achieve, and enhancing the interpretability of complex models may come at the cost of reducing their accuracy. In such cases, the challenge becomes how to create economic benefits without undermining customers' trust.

The implementation of interpretable models should be encouraged, in particular if their outcomes have a significant impact on customers. When used for risk selection and pricing, trust in AI systems can be fostered by using data sources that are related to the insured risk in a way which is intuitively understandable to customers (Christen et al. 2019).

Where it is difficult to explain algorithmic outcomes in an understandable way to consumers, there are other measures that can foster customer trust, such as traceability, auditability and transparent communication about a system's capabilities (European Commission 2019). The benefits of overly complex models may not always justify a reduction in interpretability.

Insurers should develop and implement respective internal guidelines and policies to ensure a consistent approach to the transparency and explainability of algorithmic outcomes.

3.2. Fairness

Fairness is another core principle and of utmost importance in guidelines for responsible AI. Fairness is associated with many different values such as freedom, dignity, autonomy, privacy, non-discrimination, equality and diversity, among others. These values often need to be interpreted in context, including the cultural context. It is therefore impossible to provide a universal standard of fairness.

At a general level, a procedural and a substantive dimension of fairness can be distinguished (see e.g. European Commission 2019).

Fair process: the procedural dimension implies that consumers are treated fairly throughout the entire process. An important aspect of fair treatment is the ability for customers to contest and seek effective redress against decisions affecting them (see e.g. European Commission 2019).[18] In insurance there are market conduct requirements to ensure fair treatment of customers irrespective of the technology used.[19]

Fair decisions: decisions should be fair in the sense that they do not unfairly discriminate and disadvantage individuals or groups of individuals (OECD 2019, European Commission 2019, Monetary Authority of Singapore 2018). Most guidelines therefore emphasise the absence or minimisation of unfair bias and discrimination of AI-driven decisions as a key element of fairness. Some guidelines also mention equal and just distribution of both benefits and costs as a feature of fair decisions (e.g. European Commission 2019).

Fairness in computer science

In computer science, bias refers to a systematic error that places certain groups at a systematic advantage and others at a disadvantage. Humans are not free from bias, and the use of AI may actually enhance fairness of decisions in certain circumstances. However, with machine learning algorithms having the potential to be deployed at scale, even a minimal systematic bias can affect a large number of individuals (Ruf et al. 2019a). For data scientists, the challenge is therefore to identify, measure and mitigate potential bias that could put certain groups at a systematic disadvantage.

Bias can enter algorithmic decision-making via data at several levels (Barocas and Selbst 2016, Ruf et al. 2019b). For example, data used in training algorithms may be the result of biased data collection. Bias can also result when an algorithm is used on new data that is very different from the data on which it was trained. Finally, bias may result from human judgement in the labelling of training data.

Moreover, discrimination may also occur with accurate and unbiased data. For example, even if a protected attribute such as gender, ethnicity or the like is not explicitly taken into account, such attributes can enter decision-making via proxies or a complex combination of them that correlate with this attribute (indirect discrimination).

18 The procedural dimension of fairness is thus closely related to the principle of transparency and explainability discussed above.

19 Insurance Core Principle 19 (ICP19) of the International Association of Insurance Supervisors (IAIS) states that 'The supervisor requires that insurers and intermediaries, in their conduct of insurance business, treat customers fairly, both before a contract is entered into and through to the point at which all obligations under a contract have been satisfied.'

In computer science, different approaches have recently been developed to mitigate bias. Input-based approaches rely on a better sampling of the data, while output-based approaches seek to eliminate discrimination in an algorithm's output.

In order to identify, measure and mitigate discrimination and bias of AI systems, concepts of fairness need to be mathematically defined. To date, there is no consensus on the mathematical formulation of fairness (Fiedler et al. 2016), and there are a multitude of different and mutually incompatible definitions.[20] However, different definitions may produce entirely different outcomes (Bellamy et al. 2018), and it is impossible to simultaneously satisfy all definitions of fairness (Kleinberg et al. 2017).

Fairness in insurance

In insurance, how to ensure fair decisions is particularly intricate and complex in comparison to other industries. In fact, as we will argue below, perfect non-discrimination is impossible to achieve, in particular with respect to risk selection and pricing decisions. Rather, insurers must choose how to discriminate (see e.g. Loi and Christen 2019).

Minimising discrimination involves balancing trade-offs between different concepts of fairness and raises the question of which data can be used in the underwriting process and as a basis for risk selection and rate setting. Such questions have been discussed for many years, and insurers are subject to substantive fairness requirements under existing law that differ between jurisdictions (see Box 5). Such fairness requirements typically govern what kind of information an insurer can use in decision-making, for example with respect to the use of genetic data in life and health insurance in some jurisdictions (see The Geneva Association 2017).

The following fairness concepts are particularly relevant in insurance:

Actuarial fairness demands that similar risks are treated similarly, so that the premium an individual pays corresponds to the actual risk.

Non-discrimination implies that the premium an individual pays is not based on irrelevant factors that the individual cannot influence, in particular if these factors relate to socially protected groups. Non-discrimination relates to the notions of group fairness (the goal of groups defined by protected attributes such as gender, ethnicity, sexual orientation, etc. receiving similar treatments or outcomes) and disparate treatment decisions are (partly) based on a sensitive attribute. Group fairness is largely consistent with the notion of disparate impact, i.e. outcomes that disproportionally hurt or benefit people with certain sensitive attributes.

Solidarity or mutualisation is often mentioned as a key feature of insurance. Increasing individualisation of insurance, driven by AI systems and data analytics, may disadvantage certain groups, e.g. by charging unaffordable premiums or being denied cover altogether (The Geneva Association 2018). While such considerations are not specific to the use of intelligent systems, they may become more accentuated.

It is often not enough to eliminate sensitive attributes from the data to ensure non-discrimination ('fairness through unawareness'), as such attributes can easily be picked up in proxies that correlate with these attributes (Pedreschi et al. 2008). The choice of appropriate metrics of non-discrimination is particularly intricate, because there are a multitude of related definitions that cannot be simultaneously satisfied.[21]

Box 5: Fairness in the law

The freedom of insurers to use certain features in risk selection and pricing is typically governed by a combination of anti-discrimination legislation, privacy/data protection legislation and insurance law.

- Anti-discrimination laws prohibit the unfair treatment of people based on sensitive attributes in many jurisdictions, e.g. in the U.S., federal laws prohibit discrimination based on race, colour, religion, nationality, sex, marital status, age and pregnancy in many circumstances. In the EU, the Gender Directive prohibits unequal treatment based on gender.
- Privacy law imposes restrictions or prohibits the processing of certain types of sensitive information, for example relating to genetic data.
- Insurance law may allow the use of sensitive attributes if actuarially justified. For example, in some countries (including the U.S.), gender is an admitted underwriting factor. Insurance law may also impose restrictions on certain uses of personal information (e.g. prohibition to use personal data for price optimisation in some jurisdictions).

20 Narayanan 2018 provides 21 mathematical definitions of fairness from the literature.

21 Concrete metrics of non-discrimination being discussed include equalised odds, equal opportunity, demographic (statistical) parity and predictive rate parity, among others (see e.g. Ruf et al. 2019b).

Although up to now there has not been a broad debate around possible metrics to be used to measure affordability and exclusion, regulators do in some cases impose limits on the individualisation of insurance, e.g. by limiting the allowed ratio between the highest and lowest premium. The Dutch Insurance Association has developed a 'solidarity monitor' to assess the development of the spread of insurance premiums and individual insurability over time.

The relative importance of different fairness concepts varies between different types of insurance and jurisdictions. For example, in many jurisdictions, solidarity is considered an essential feature of health insurance. Depending on the application, a set of different metrics may be selected and continuously monitored. For example, in order to limit adverse impacts of enhanced pricing algorithms on customers, such as bias and implications for affordability, one insurer implemented a comprehensive post-monitoring system. Models are scrutinised by diverse teams from different functions, and system output is tested using nine different tests to validate that the models are in line with expectations of the modelling team, business partners and regulators. Where regulatory approval of rates is required, regulators have been provided with dedicated tools to track the performance of the system.

Conclusions and discussion

Fairness requires a focus on mitigating discrimination and bias. Fairness is therefore closely related to the principle of transparency and explainability in its emphasis on enabling affected individuals to understand and contest algorithmic outcomes. In insurance, mitigating bias and discrimination is particularly challenging, as there are different and mutually exclusive concepts and metrics of non-discrimination.

To monitor and mitigate bias requires the quantification of fairness. While academic research on appropriate fairness metrics is still evolving and should be encouraged, insurers need to identify context-specific fairness definitions for each use of AI. As fairness is not a new concept in insurance, existing frameworks and norms (such as actuarial ethics) should be leveraged. Roles and responsibilities with respect to monitoring and mitigating bias should be clearly defined.

As bias can enter decision-making at various stages, it is important to raise awareness at different management levels through appropriate educational and training programmes.

4. Recommendations

If insurance customers are to reap the potential benefits of AI, it is critical that they trust these systems. The insurance industry can facilitate this trust by affirming core principles for responsible use of AI, such as safety and privacy, transparency and explainability, and fairness and promoting the responsible use of AI through the actions described below.

1. Establish internal guidelines and policies for the use of AI

Internal guidelines and policies play an important role in raising the awareness of the benefit–risk trade-offs in the use of AI in insurance. Insurers should therefore develop and adopt respective guidelines and policies that include principles for dealing with issues of transparency and explainability and fairness. In particular, guidelines should help to clarify how the benefits and risks of using AI should be assessed on a case-by-case basis. Actuaries, risk managers, data scientists and data protection officers should closely cooperate in the development and implementation of such guidelines and policies.

In doing so, insurers may adopt a risk-based approach to the governance of AI, implying a special focus on the uses of AI systems that may have a significant impact on individuals. The significance of impact refers to the consequences of decisions to affected individuals and depends on the specific circumstances in which AI is used.

For instance, uses of AI that automate specific tasks but do not change the logic of decision-making in any way (such as extracting relevant information from documents via computer vision) are likely to exhibit low significance of impact. In contrast, any use of AI that changes the logic of decision-making (i.e. that applies a new model to existing data) may exhibit higher significance. The highest level of significance of impact may be in uses that change the logic of decision-making based on new data sources.

Similarly, applications in customer engagement may have lower significance than applications that are used to determine payouts to policyholders. Applications in underwriting/pricing may exhibit the highest levels of significance, in particular if they could lead to the exclusion of customers.[22]

Figure 4 provides an illustrative classification of use cases based on their nature and position in the value chain. In practice, each use case has to be assessed on its individual merits.

22 The EU Guidelines on automated individual decision-making and profiling mention, as an example of a significant impact, the automatic refusal of an online credit application. According to these guidelines, decisions to present targeted advertising based on profiling will, in general, not have a similarly significant effect on individuals. Differential pricing based on personal data could have a significant effect if it leads to unaffordability.

Figure 4: Possible classification of AI applications and their significance (for illustration)

Change in logic and new data sources	• Robo-advice using external data sources	• Price optimisation using lifestyle data • Pricing allgorithms using lifestyle data	
Change in logic of decision-making	• Customer segmentation and targeted advertising	• Pricing algorithms using traditional data	• Fraud detection • Automated claims triage
No change in logic or data	• Conversational agent (chatbot)	• Computer vision to extract information from documents	• Computer vision to extract information from documents
	Customer engagement	**Underwriting / pricing**	**Claims**

High significance Medium significance Low significance

2. Adopt appropriate governance structures

Ensuring a responsible use of AI requires an appropriate governance structure that assigns clear accountabilities to individuals, committees or departments that have the necessary decision-making competencies, the necessary skills and expertise as well as the associated processes in place, including triggers and escalation.

There are many different governance models, each with their own advantages and disadvantages. The choice of an organisational model will depend on a company's existing structure and culture.

The following are some important aspects to consider in choosing an appropriate organisational model:

- ***Centralised vs decentralised model***

The responsibility for the fair use of AI may be assigned to a central team (e.g. under the lead of a data ethics officer) or delegated to the local business. While a centralised model may have the advantage of facilitating a consistent approach across the entire organisation, there is a risk that such a team may be perceived as too remote from business and customer needs.

- ***Reliance on existing governance structures vs creation of new structures***

The oversight of fair and responsible use of AI systems could be delegated to an existing framework such as the risk management framework, IT governance framework or compliance framework. Alternatively, a new and dedicated oversight mechanism may be established that is closely embedded in the existing risk management framework.

- ***Appropriate expertise and experience***

Ensuring appropriate levels and diversity of skills, expertise and experience (including data science, actuarial, legal and regulatory expertise) in key decisions on the use of AI systems is crucial.

3. Develop and roll out internal training programmes

Finally, ensuring responsible use of AI requires awareness of related benefits and risks across different functions and managerial levels. In order to raise awareness, insurers should consider developing and rolling out comprehensive training programmes on the benefits and risks of AI as well as the respective internal guidelines and policies. Such training programmes would ideally target employees across management levels and decision-making functions, including agents and other customer-facing employees.

Glossary

Artificial intelligence: A branch of computer science dealing with the computer simulation of intelligent behaviour. More commonly, the term is used to refer to the capability of a machine to imitate intelligent (human) behaviour (https://www.merriam-webster.com/dictionary/artificial%20intelligence).

Bias: A systematic error that places certain groups at a systematic advantage and others at a systematic disadvantage.

Big data: A high-volume, high-velocity and high-variety information asset that demands cost-effective, innovative forms of information processing for enhanced insight and decision-making. Big data may be assessed through five 'V' parameters: volume, velocity, variety, veracity and variability (Swan 2015). Some commentators have added visualisation and value to those parameters (Devan 2016). Other definitions emphasise the complexity of big data. The National Institute of Standards and Things (NIST) defines big data as data that exceed the capacity and capability of current methods and systems.

Computer vision: The field of study surrounding how computers see and understand digital images and videos. Computer vision spans all tasks performed by biological vision systems, including 'seeing' or sensing a visual stimulus, understanding what is being seen and extracting complex information into a form that can be used in other processes. This interdisciplinary field simulates and automates these elements of human vision systems using sensors, computers and machine learning algorithms. Computer vision is the theory underlying AI systems' ability to see and understand their surrounding environment (https://deepai.org/machine-learning-glossary-and-terms/computer-vision).

Curse of dimensionality: Statistical phenomena that occur when classifying, organising and analysing high dimensional data (with hundreds or thousands of variables) that do not occur in low dimensional spaces, specifically the issue of data sparsity and 'closeness' of data.[23]

Data: Facts and statistics collected for reference or analysis (Swan 2015).

Data protection: Technically speaking, the process of safeguarding important information from corruption and/or loss.[24] European jurisprudence tends to treat data protection as an expression of the right to privacy. While there are overlaps in the concepts of data protection and privacy, there are also differences in their scope, as the scope of data protection is broader than the scope of privacy (Kokott, J. and Sobotta, C. 2013).

Data science: The extraction of knowledge from data. Data science employs techniques and theories from mathematics, statistics, computing and information technology—for example, machine learning—to uncover patterns in data from which predictive models can be developed (Swan 2015).

Deep Learning: A machine learning technique that constructs artificial neural networks to mimic the structure and function of the human brain (https://deepai.org/machine-learning-glossary-and-terms/deep-learning).

Disparate impact: The incident of outcomes disproportionally hurting or benefiting people with a certain sensitive attribute.

Disparate treatment: Decisions based (partly) on a sensitive attribute.

Explainability: See interpretability.

23 https://deepai.org/machine-learning-glossary-and-terms/curse-of-dimensionality
24 http://searchstorage.techtarget.com/definition/data-protection

Information: Facts provided by or learned about something or someone. Data and information may both be used as a basis for reasoning or calculation. While there used to be more of a distinction between data as underlying facts and statistics, and information as knowledge gleaned from these facts and statistics, the definitions have become quite close and are often used synonymously (Swan 2015).

Informational privacy: The interest of individuals in exercising control over access to information about themselves (Stanford Encyclopedia of Philosophy 2014). See section 'Transparency and explainability'.

Internet of Things (IoT): Defined by the International Telecommunication Union (ITU) as 'a global infrastructure for the information society, enabling advanced services by interconnecting (physical and virtual) things based on existing and evolving interoperable information and communication technologies' (IoT-GSI 2015).

Interpretability: The ability to explain or to provide the meaning in understandable terms to a human (Doshi-Velez and Kim 2017).

Machine learning: A technique or subfield of AI that provides systems with the ability to automatically learn, and from experience or examples, improve without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it to learn for themselves.[25]

Overfitting: In statistics, an analysis or model which corresponds too closely or exactly to a particular set of data and may therefore fail to fit additional data or predict future observations reliably.[26]

Personal information or data: Information or data that are linked or can be linked to individual persons (Stanford 2014). In the European General Data Protection Regulation, personal data is defined as 'any information relating to a person who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that person'. This means that in many cases, online identifiers, such as IP address and cookies will now be regarded as personal data if they can be (or are capable of being) linked back to the data subject without undue effort.

Privacy: There is no universally accepted definition of the concept of privacy. In this report, we define privacy as the 'appropriate use of personal data'.

25 http://www.expertsystem.com/machine-learning-definition/
26 https://www.lexico.com/en/definition/overfitting

References

AXA. 2019. Authors: Ruf, B., M. Hirot. 2019a. *Regulating machine learning: Where do we stand? State of the art and challenges ahead.* https://axa-rev-research.github.io/static/AXA_WhitePaper_RegulatingML.pdf

AXA. 2019. Authors: Ruf, B., and V. Grari. 2019b. *Understanding and mitigating bias*. https://axa-rev-research.github.io/static/AXA_Booklet_Bias.pdf

Barocas, S., and A. D. Selbst. 2016. Big data's disparate impact. *California Law Review* 104 (3). http://dx.doi.org/10.15779/Z38BG31

Bellamy, R.K.E. et al. 2018. *AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias*. https://arxiv.org/abs/1810.01943

BaFin. 2018. *Big data meets artificial intelligence—Challenges and implications for the supervision and regulation of financial services.* Federal Financial Supervisory Authority. July 2018. https://www.bafin.de/SharedDocs/Downloads/EN/dl_bdai_studie en.html;jsessionid=1DE4EFB708FA602B6829FC41B0FB7432.1_cid372?nn=9866146

Christen, M. et al. 2019. *Big data ethics recommendations for the insurance industry. A consolidation report outlining the results of the NRP 75 project: Between solidarity and personalisation—Dealing with ethical and legal big data challenges in the insurance industry.* www.nfp75.ch/SiteCollectionDocuments/NFP75-Website-Publikation-Christen.pdf

DeVan, A. 2016. *The 7 V's of Big Data.* https://www.impactradius.com/blog/7-vs-big-data/

Doshi-Velez, F., and B. Kim. 2017. *Towards a rigorous science of interpretable machine learning.* Cornell University research paper. https://arxiv.org/abs/1702.08608

European Commission. 2019. *Ethics guidelines for trustworthy AI.* High-Level Expert Group on Artificial Intelligence. April 2019.

Everyday ethics for artificial intelligence. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf

Fiedler, S. A., C. Scheidegger, and S. Venkatasubramanian. 2016). *On the (im)possibility of fairness.* Cornell University research paper. https://arxiv.org/abs/1609.07236

Guidotti, R., A. Monreale, S. Ruggieri, F. Turini, F. Gianotti , and D. Pedreschi. 2018. *A survey of methods for explaining black box models.* ACM Computing Surveys. February 2018. https://dl.acm.org/citation.cfm?doid=3271482.3236009

Hall, P., S. Ambati, and W. Phan. 2017. *Ideas on interpreting machine learning.* https://www.oreilly.com/radar/ideas-on-interpreting-machine-learning/

International Association of Insurance Supervisors (IAIS). 2019. *Insurance Core Principles and Common Framework for the Supervision of Internationally Active Insurance Groups 2019 (ICP19).*

Internet of Things Global Standards Initiative (IoT-GSI). 2015. http://www.itu.int/en/ITU-T/gsi/iot/Pages/default.aspx

IEEE. 2019. *Ethically aligned design.* First Edition. https://ethicsinaction.ieee.org/

The Institute for Ethical ML . 2019. *The Responsible Machine Learning Principles. A practical framework to design AI responsibly.* https://ethical.institute/principles.html

Jobin, A., M. Iencya, and E. Vayena. 2019. *Artificial intelligence: The global landscape of ethics guidelines.* Preprint version. Health Ethics and Policy Lab, ETH Zurich.

Kleinberg, J., S. Mullainathan, and M. Raghavan. 2017. Inherent trade-offs in the fair determination of risk scores. In *Innovations in Theoretical Computer Science.* ACM.

Kokott, J. and C. Sobotta. 2013. The distinction between privacy and data protection in the jurisprudence of the CJEU and the ECtHR. In *International Data Privacy Law,* 3(4). https://academic.oup.com/idpl/article/3/4/222/727206/The-distinction-between-privacy-and-data

Laugel, T., M. Lesot, C. Marala, and M. Detyniecki. 2019. Issues with post-hoc counterfactual explanations: A discussion. *ICML Workshop on Human in the Loop Learning* (HILL 2019).

Lazer, D., and R. Kennedy. 2015. What we can learn from the epic failure of Google Flu trends. *Wired Opinion*. 10 January 2015. https://www.wired.com/2015/10/can-learn-epic-failure-google-flu-trends/

Loi, M., and M. Christen. 2019. *Insurance discrimination and fairness in machine learning: An ethical analysis.* 17 August 2019. SSRN paper. http://dx.doi.org/10.2139/ssrn.3438823.

McKinsey & Company. 2018. *Insurance 2030 The impact of AI on the future of insurance.*

Metz, R. 2019. Why Microsoft accidentally unleashed a neo-nazi sexbot. *MIT Technology Review*, 24 March 2016. https://www.technologyreview.com/s/601111/why-microsoft-accidentally-unleashed-a-neo-nazi-sexbot/

Microsoft. 2019. *Microsoft AI Principles.* https://www.microsoft.com/en-us/ai/our-approach-to-ai

Monetary Authority of Singapore. 2018. *Principles to promote fairness, ethics, accountability and transparency (FEAT) in the use of artificial intelligence and data analytics in Singapore's financial sector.* 12 November 2018. https://www.mas.gov.sg/~/media/MAS/News%20and%20Publications/Monographs%20and%20Information%20Papers/FEAT%20Principles%20Final.pdf

Narayanan, A. 2018. Tutorial: 21 fairness definitions and their politics. *Conference on Fairness, Accountability, and Transparency*, February 2018.

OECD. 2019. *Recommendations of the Council on Artificial Intelligence*, adopted on May 22, 2019.

Pedreschi, D., S. Ruggieri,and F. Turini. 2008. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. Las Vegas, 2008.

SCOR. 2018. *The Impact of Artificial Intelligence on the (Re) Insurance Sector*. SCOR, 2018.

Stanford Encyclopedia of Philosophy. 2014. Privacy and Information Technology. November 2014.

Swan, M. 2015. Philosophy of big data: Expanding the human-data relation with big data science services. *2015 IEEE First International Conference on Big Data Computing Service and Applications*.

The Geneva Association. 2017. *Genetics and life insurance—A view into the microscope of regulation.* Author: Ronald Klein. June 2017.

The Geneva Association. 2018. *Big data and insurance: Implications for innovation, competition and privacy*. Author: Benno Keller. March 2018.

The Geneva Association. 2019. *The role of trust in narrowing protection gaps.* Author: Kai-Uwe Schanz. November 2019.

Wachter, S., B. Mittelstadt, and C. Russell. 2017. *Counterfactual explanations without opening the black box: Automated decisions and the GDPR.* Cornell University. https://arxiv.org/abs/1711.00399

World Economic Forum. 2018. *How to prevent discriminatory outcomes in machine learning.* White Paper. Global Future Council on Human Rights 2016-2018. March 2018. http://www3.weforum.org/docs/WEF_40065_White_Paper_How_to_Prevent_Discriminatory_Outcomes_in_Machine_Learning.pdf

The use of artificial intelligence (AI) in insurance can potentially bring economic and societal benefits by lowering insurance costs and helping insure more people. For this report, The Geneva Association analysed two of the five core principles identified for the responsible use of AI—1) transparency and explainability and 2) fairness—that raise particularly complex issues in insurance. With this analysis and a set of recommendations for insurers, we aim to contribute to the responsible use of AI in insurance and the realisation of its benefits for society.

The Geneva Association

International Association for the Study of Insurance Economics
Talstrasse 70, CH-8001 Zurich
Tel: +41 44 200 49 00 | Fax: +41 44 200 49 99

secretariat@genevaassociation.org