MILLIMAN RESEARCH REPORT

Artificial Intelligence

The ethical use of AI in the life insurance sector

November 2020

David Burston, FIA
Charlie Howell
Andrew Gilchrist, FIA
Philip Simpson, FIA

Table of Contents

1. Introduction and scope

There isn’t a day that goes by where you don’t engage with artificial intelligence (AI) in some way.

We are aware that this is a sweeping generalisation, but it is a reasonable assumption to make in 2020. We constantly engage with technologies that are embedded with algorithms, tools and techniques that seek to replicate human intelligence: from the personalised recommendations you receive on Spotify or Netflix, to voice assistants like Amazon’s Alexa and Apple’s Siri that understand and process your words, to ride-sharing apps like Uber and Lyft that predict demand for taxis. We even work to improve and train AI systems, in many cases without even knowing it (Malhan, 2020).

The list of use cases of technology powered by AI continues to rapidly expand. When used properly, AI’s unique capabilities can better society by, for instance, improving healthcare, protecting the environment, or enhancing how humans interact with each other and the world around them. Consequently, most developed economies have recognised the game-changing nature of AI and are investing increasing amounts of money and research into a race for global dominance in the field, with the U.S. and China at the forefront in terms of both the level of investment and research output.

Despite the aforementioned examples, AI is not some science fiction that is cooked up in Silicon Valley and only developed by, and for, the Big Tech giants[1]. Life insurance has a data-fuelled business model that makes it an industry that appears, on the surface, to be well primed for the technological advances that AI can offer in addressing the challenges it faces in an increasingly digital world. This is summarised well in ‘Mind Your ABCs’, an article by Milliman Principal Pat Renzi:

> ‘Emerging technology offers a solution for many of the concerns today’s insurers face as it creates a foundation of opportunity and lays the groundwork to not only navigate disruption, but to deliver true innovation and value.’

This can range from changing the way in which products are sold and operate so that they are more personalised to the individual, to driving internal changes within organisations so that they can operate in a more cost-efficient manner, with such cost savings eventually passed on to customers through cheaper premiums, fuelling competitive advantage.

Therefore an undeniable global boom in investment in insurance technology (InsurTech), has occurred including solutions embedded with AI technologies as insurers develop and try emerging solutions to old problems. Indeed, movements to reposition traditional insurers into digital, data and technology-driven business models are well underway, with insurers seeking to embrace InsurTech and fundamentally transform how they do business.

Innovation in AI is important but cannot take place in a social vacuum.

The worst-case fears of some is that unconstrained technological advances that narrowly focus on the bottom line are creeping us ever closer to a ‘Minority Report’[2] society where our behaviours, interactions, and movements are continuously tracked such that our individual or collective decision making is rendered predictable and so can be exploited by private and public bodies.

Others fear a society where AI systems have the ability to make decisions that impact the prospects and life course of an individual, but where those systems are too complex or opaque for users to understand how such conclusions were determined. Where the systems are discriminatory and yet it is not clear on how the system creates such biases. Where we are not clear on what data is being captured about us and how it is used. Where no one is willing to accept accountability for the decisions that are being made by the system. Interestingly, as highlighted by Stephen Cave for the Ada Lovelace Institute (Cave, 2019), this dystopian vision of the future was actually offered to us over 100 years ago in 1915 in Frank Kafka’s novel *The Trial* but remains relevant to the fears and concerns around the implementation of AI in public and private institutions.

[1] Commonly referred to as FAANG: Facebook, Apple, Amazon, Netflix, and Google.

[2] The 2002 dystopian film *Minority Report*.

2020 has been a year of significant global change that has increasingly driven consideration and deliberation on the ethical use of AI into the mainstream, rather than just being the concerns of academia, think tanks, industry bodies and governments, including:

- COVID-19 track-and-trace systems introduced in certain countries have forced us to weigh our willingness to lose some of our personal freedoms for the greater societal good (Ada Lovelace Institute, 2020a).
- In light of the killings of George Floyd, Breonna Taylor, Ahmaud Arbery, and many other Black American, communities around the world are protesting, confronting and calling for an end to systematic racial discrimination, injustice and inequality. As discussed in this paper, AI systems can be vulnerable to bias and discrimination—society must ensure that they do not become another system that directly or indirectly allows or reinforces systemic biases.
- A mass movement to home schooling and home working has included a cancellation of the formal examinations for students in the U.K. such that exam results were initially[3] predicted by an algorithm based on students' mock examinations and the school's track record for results. Given the effects that exam results can have on an individual's potential job prospects and access to further or higher education, the fact that an algorithm was allowed to make such an important decision on someone's life sparked outrage. It raised questions of transparency, and of fairness and discrimination, as the algorithm was thought to favour students from fee-paying schools, leaving high achievers from free state schools disproportionately affected (BBC, 2020).

At the same time, councils in the U.K. have chosen to drop algorithms that were being used to determine whether someone was likely to abuse the welfare system by committing benefit fraud or to identify children at risk of abuse, due to concerns about bias and a lack of public consultation, among other issues (Marsh, 2020). This highlights a significant risk to a firm of unethical AI development beyond the potential harm to its customers; when things go wrong there can be a breakdown of trust in the organisation, leading to a significant reputational risk and a loss of trust in the technology, reversing any public desire for its use.

Consequently, ethical considerations must guide innovation to act in the public good, to drive the development, deployment and use of AI systems that can be trusted.

This paper is intended to be exploratory, not exhaustive. We aim to provide an introduction to the use of AI systems in the life insurance industry, ethical questions raised and some insights on the external environment that influences, or provides guidance on, ethics in the context of AI:

- Section 2 outlines some of the AI techniques and concepts that will be discussed elsewhere in the paper and provides some background on the current uptake of AI in the life insurance sector, and how it is being used.
- Section 3 provides a high-level introduction and discussion of some of the broad ethical and professional considerations that stakeholders involved in AI in the life insurance sector may come across.
- Section 4 provides a review of relevant literature on frameworks, guidelines and reports published by relevant bodies to the U.K. life insurance market that insurers can lean on to guide an ethical approach to the use of AI.
- Appendix 1 describes the methodology of a number of machine learning (ML) techniques as applied to the application of spam email detection, to give some understanding of model features that give rise to ethical issues (e.g., complexity and difficulty to explain the inner workings).
- Appendix 2 provides the results of a review of the key ethical principles raised in 18 AI ethics guidelines.
- Appendix 3 provides a detailed review of the European Commission's 'AI Ethics Guidelines for Trustworthy AI,' which is key guidance to the ethical use of AI that is relevant in Europe.

Given the authors are all actuaries working in the life insurance sector in the U.K., the report focuses more heavily on the U.K. market.

[3] Following public outrage, the U.K. government decided to instead rely on teachers' predictions (which the government was initially concerned would result in inflated grades).

2. AI and its use in the life insurance sector

2.1 WHAT IS ARTIFICIAL INTELLIGENCE?

There are numerous definitions and views out there for what artificial intelligence is. Various authors draw the line on what does and does not count as AI at different points of sophistication or *intelligence* demonstrated by a system.

The authors of this report take a more general view: AI is used in this report as an umbrella term for the branch of computer science whose aim is to create computer systems that behave intelligently and independently, mimicking (to different degrees of sophistication) the nature and behaviour of humans.

The human qualities that the computer system aims to mimic may be that of learning from experience (and adjusting responses based on new input), seeking patterns, problem-solving, decision making, classifying, and predicting, or the system may aim to replicate our ability to see and process images, to hear and discern meaning, or to speak and construct language in a meaningful and understandable way.

This report considers the following subfields of AI:

- **Machine learning:** To learn patterns and relationships in data
- **Natural language processing (NLP):** To read and understand human language
- **Computer vision:** To see and understand images
- **Speech recognition:** To hear and understand audio

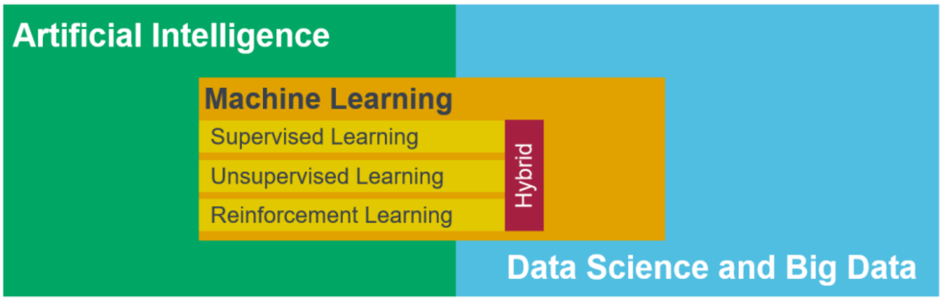
The above list is not exhaustive, and it should be noted that there is much overlap between these fields. For example, ML algorithms can help train NLP tools to perform better.

In this report we aim to make it clear, when relevant, the type of AI system we are describing, although generally speaking, the use of AI will be referencing machine learning.

2.2 WHAT DISTINGUISHES MACHINE LEARNING FROM AI?

Whilst AI describes a broad area of computer science that incorporates *any* computer system aiming to replicate tasks that would traditionally require human intelligence, ML predominantly focuses on the development of models that are trained (i.e., calibrated, or fit to) on vast sets of data in order to discern or recognise patterns (i.e., to learn) in that data that would not be initially obvious. It is a field that sits at the intersection of two technological fields: AI and big data[4]. This is summarised below in Figure 1.

FIGURE 1: MACHINE LEARNING SITS AT THE INTERSECTION OF AI AND BIG DATA



Such models can then be incorporated into AI systems in algorithms that have the purpose of solving a particular problem[5] without human input (or with limited input). In particular, the system will be fed on sample data and, based on its understanding of the key relationships and patterns in past data, used to predict outcomes or make decisions.

[4] Big data is a field that describes the methodologies for analysing and systematically extracting information from large volumes of data—both structured and unstructured.

[5] This is a form of 'weak' or 'narrow' AI whereby the system is focussed on a single narrow task that it is programmed to perform. The AI systems we discuss in his paper are all such forms. 'Strong' AI systems where the system can learn, solve and perform any task presented, without any human intervention, is a subject of academic research.

This differs from traditional AI systems that aim to learn, predict, and make decisions where the thought process, and thus *intelligence*, had to be explicitly algorithmically programmed and coded—known as knowledge-based systems. In this sense, the programmer had to be able to describe the complex interdependencies, relationships and patterns between variables in rules-based lines of code. This has its limits in terms of the ability to maintain the code, the ability of humans to accurately identify the relationships and patterns, and the ability to capture nonlinear and complex relationships between variables. Therefore, models that review data and discern relevant relationships for themselves to come up with the rules, without the need for explicit code, are machine learning models.

2.3 WHAT TYPES OF MACHINE LEARNING ARE THERE?

There are several categories of ML models, and they are commonly categorised in terms of the level of human input that is required in order to train the model (i.e., the type of data that will be provided to the model). These include:

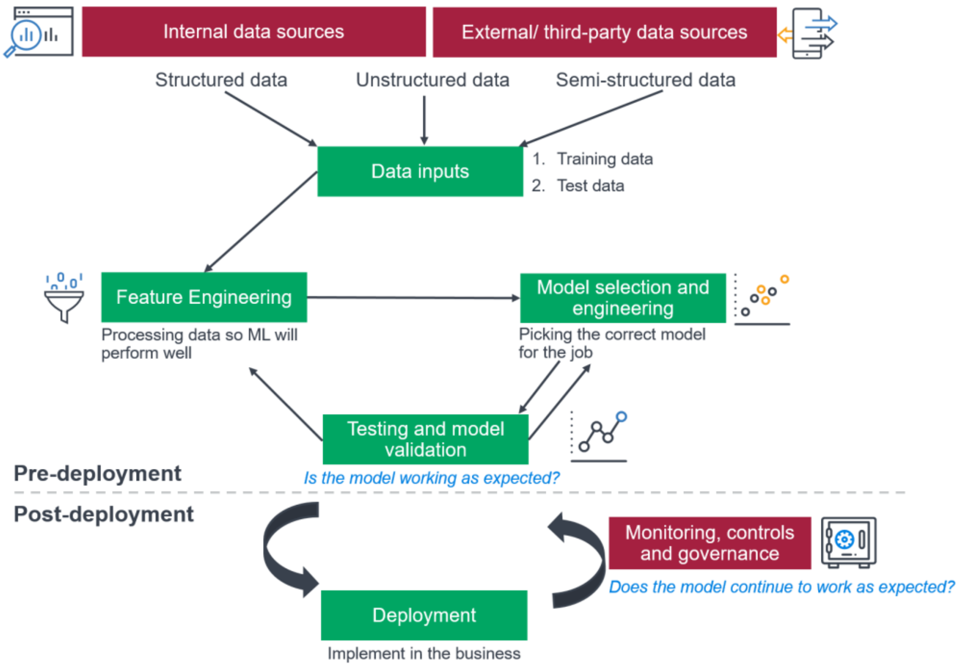
- **Supervised learning systems:** The model is trained using a **labelled data set** that maps the inputs to the output target variables. That is, the model is trained that given data item *X*, it should map to target variable *Y*. In this sense, the model trains a function *F* such that *F(X)* = *Y*. Take, for example, a set of electronic messages that have been labelled either as bona fide emails directly meant for the recipient or as generalised (potentially malignant) spam. Trained on this data, the model will then *learn* a set of rules for classifying and labelling (as email or spam) messages outside of the training data set.
- **Unsupervised learning systems:** The model is trained using an **unlabelled data set** and is tasked with discovering specific patterns, structures and regularities in the data without being told where to look or how to go about the task (i.e., it is unsupervised). In this case the model does not know what the data item *X* is in order to look for a function *F*—it must determine these for itself. There is a range of methods for doing this, including clustering, association rules and principal component analysis (PCA). In our example, clustering would involve the model splitting a set of emails into clusters with similar characteristics (e.g., lots of '£,' misspelling, aggressive wording).
- **Reinforcement learning systems:** Such models are neither strictly supervised nor unsupervised, and they instead somewhat bridge the two. The model is trained on an **unlabelled data set** and uses this to determine its outcome or action for each data point. In our example, it would have its first attempt of classifying messages into emails or spam. The system then receives feedback by way of a reward (from the system or perhaps from a human) that allows the algorithm to learn from experience in a trial-and-error fashion. Over time, the model will take the path that maximises its reward and will continue to adjust and improve in light of new data.

Since various supervised and unsupervised learning models have their own pros and cons, **hybrid systems** that employ a mix of learning models are often used with the aim that the flaws of one approach are addressed by another.

2.4 HOW DO MACHINE LEARNING AI SYSTEMS WORK?

Machine learning is an involved process that is broken down into multiple distinct stages that are summarised in Figure 2. We will discuss each of these in turn.

FIGURE 2: HOW TO CREATE AN ML SYSTEM: AN OVERVIEW


Internal data sources
External/ third-party data sources
Structured data
Unstructured data
Semi-structured data
Data inputs
1. Training data
2. Test data
Feature Engineering
Processing data so ML will perform well
Model selection and engineering
Picking the correct model for the job
Testing and model validation
Is the model working as expected?
Pre-deployment
Post-deployment
Monitoring, controls and governance
Does the model continue to work as expected?
Deployment
Implement in the business

As discussed, for an ML model to learn, it requires data from which to learn. There are three common types of data structures (Enterprise Big Data Framework, 2019) that ML is capable of processing and discerning insight from:

- **Structured data:** This refers to highly organised data in a tabular format that adheres to a predefined data model—a model of how data can be stored, processed and accessed. The data has structured rows and columns, each data item has a fixed, narrow, unambiguous meaning, and can be easily analysed. Examples include data held on policy administration systems or finance data kept in ledgers.
- **Unstructured data:** Sometimes referred to as 'nontraditional data sources' or 'alternative data sources,' this type refers to any data that does not conform to a predefined data model or that is not organised in any predefined way. Techniques are required to transform unstructured data into something meaningful that can be analysed. Examples include textual data, audio, video files and images.
- **Semi-structured data:** This is any data that is not organised into the formal structure of data models but does contain some tags or other markers that allow for the analysis of some of the relationships with the data within the fields. Examples of semi-structured data include the languages HTML used on websites and the XBRL format required for the submission of Solvency II Quantitative Reporting Templates.

In particular, ML's ability to interpret meaning from unstructured data has opened up new opportunities for firms to discern meaningful insights from data that they were unable to achieve using traditional forms of analysis on structured data. That said, as discussed later in this paper, it also creates the risk that these new data sources are used without taking into account ethical considerations and without due regard to data privacy.

In order for an ML model to perform well for the particular problem at hand, the data needs to undergo a number of preprocessing steps. This includes data cleansing (e.g., removing unwanted [duplicate, or irrelevant] data, fixing structural issues like typos, treatment of missing data and/or outliers), feature extraction (i.e., reducing the dimensionality of data sets that are too large), feature selection (i.e., choosing the most relevant variables for the problem), and feature engineering.[6]

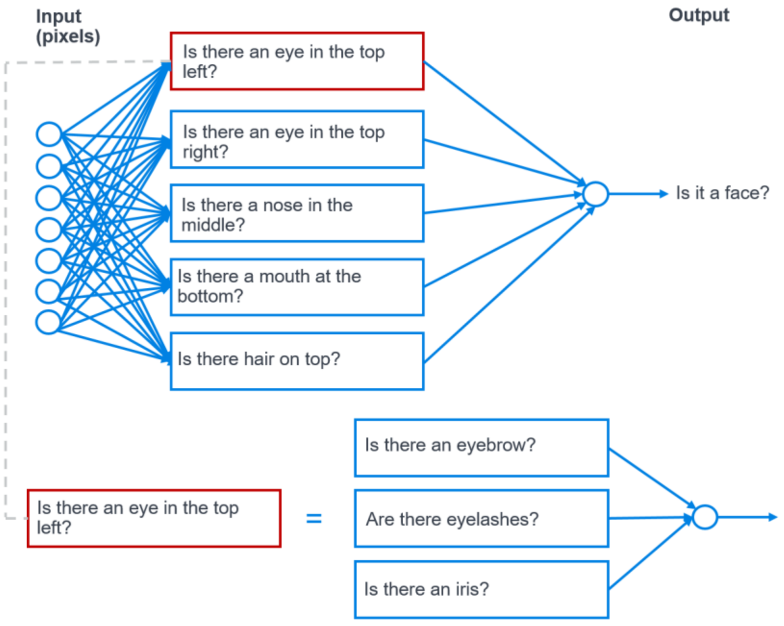
Feature engineering is a key area for data scientists to improve the functioning of the ML algorithm and involves the creation of new variables, or features that may be of relevance to discerning relationships in the data. As a simple example, in an ML model to predict lapses from a policyholder data set, you may have the fields *issue_date* (when the policy started) and *cancellation_date* (when the policy ended), but you could create a new field *duration_to_lapse* that is the calculated duration in-force before the policy lapses. This composite feature will have a greater predictive ability for lapses than that of its parts.

Such feature engineering is crucial, but time consuming and challenging, and requires prior expert knowledge and judgement. Crucially, it highlights a weakness of ML processes—that they cannot learn to extract the key features from the data by themselves, requiring humans to compensate for this weakness (Goodfellow, Bengio, & Courville, 2016).

This has led to the field of representation learning (i.e., feature learning) whereby the ML model automatically discovers the representations needed for feature detection or classification from the raw input data alone. The most ground-breaking technique has been deep learning that structures machine learning algorithms to learn hierarchal features of data by combining multiple layers of simpler features. An example of such is provided in (Nielson, 2015) and reproduced in Figure 3. The example is of an ML model that is attempting to identify a face in an image built of pixels.

The challenging problem can be broken down into layers, each of which aims to discern certain features from the data. Layers can be broken down further and further until you have a deep network where the complex problem has been broken down into simpler sub-problems. The individual layers will be based on the presence or absence of very simple shapes at particular points of the image.

FIGURE 3: DEEP LEARNING EXAMPLE (NIELSON, 2015)


Input (pixels)
Output
Is there an eye in the top left?
Is there an eye in the top right?
Is there a nose in the middle?
Is there a mouth at the bottom?
Is there hair on top?
Is it a face?
Is there an eye in the top left?
=
Is there an eyebrow?
Are there eyelashes?
Is there an iris?

Some common machine learning models are listed in Table 1 below[7]. In order to better understand the inner workings of ML models, and therefore to identify some of the ethical implications that the use of such models may introduce (e.g., around transparency, accountability, fairness), Appendix 1 provides an introductory overview of a number of these models.

[6] ML methods can be used in these processes.

[7] The list is designed to be illustrative and not exhaustive in terms of the algorithms included or the common applications listed.

TABLE 1: COMMON MACHINE LEARNING APPROACHES

LEARNING ALGORITHM FAMILY	LEARNING ALGORITHM EXAMPLES	TYPE OF LEARNING	TYPE OF PROBLEM SOLVED	TYPES OF DATA COMMONLY APPLIED TO	COMMON APPLICATIONS
Regression models	▪ Linear ▪ Polynomial ▪ Logistic ▪ Piecewise ▪ Splines ▪ Regularised	Supervised	Both regression and classification	Structured	Forecasting, prediction
Decision trees	▪ Iterative Dichotomiser 3 (ID3) ▪ C4.5 ▪ CART ▪ MARS	Supervised	Both regression and classification	Structured	Prediction, data manipulation, variable selection
Random forests	N/A	Supervised (most common) or unsupervised.	Both regression and classification	Structured	Prediction, variable selection
Support vector machine (SVM)	▪ Radial kernel ▪ Linear kernel ▪ Polynomial kernel	Supervised (unsupervised if dealing with unstructured data)	Both regression and classification	Structured (an extension of SVM can be used for unstructured data)	Prediction, image, text classification
Neural networks	▪ Artificial neural networks ▪ Long short-term memory ▪ Convolutional neural networks	Can be supervised (most common), unsupervised, or used in reinforcement learning	Both regression and classification	Unstructured and semi-structured	Pattern recognition, image identification, system identification, data mining and visualisation, machine translation, filtering, medical diagnoses, finance
K-nearest neighbour	N/A	Supervised	Both regression and classification	Structured	Forecasting, recommendation engines, fraud detection
Boosting	▪ Adaptive Boosting ▪ GBM ▪ XGBoost ▪ LightGBM ▪ CatBoost	Supervised	Both regression and classification	Structured	Prediction
Naive Bayes	N/A	Supervised	Classification	Structured	Prediction, classification
Clustering	▪ Naive K-means clustering ▪ K-means ++ ▪ Jenks natural breaks optimisation ▪ Hierarchical (dendograms) ▪ Fuzzy ▪ Density Based	Unsupervised	Clustering	Structured or unstructured	Feature learning, image segmentation, customer segmentation, pattern recognition, recommendation engines

The choice of ML model will depend on the data set available and on the task that the algorithm is required to perform. At first it may not be clear what ML model should be used for a particular problem, and so a process of model engineering will be deployed whereby different models are chosen, trained, tested and compared to one another using a range of model validation metrics[8] that will vary based on the type of data being used, the type of problem the ML model is aiming to solve and the type of validation that is sought.

This is not simply a task of determining which ML model has the greatest predictive power. Instead, as part of the model validation, it is important to ensure the model is robust, performs well on different data sets and is flexible, with some of the core questions that need to be asked, tested and validated being:

[8] We do not describe model validation metrics in this paper. For a good introduction to validation methods for ML models, please see (Grootendorst, 2019), which also provides a link to some model validation code on GitHub (using Python).

- **Is the model accurate in its predictions?**
  The performance metric will differ based on the type of problem being assessed. For example, a binary classification problem of 'policy will lapse' or 'policy will not lapse' will be assessed by looking at the number of true positives (i.e., correctly predicted to lapse), true negatives (i.e., correctly predicted as non-lapsed), false positives (incorrectly predicted to lapse) and false negatives (incorrectly predicted as non-lapsed). The approach of assessing model accuracy for a predictive ML algorithm based on a regression approach would consider model fit metrics[9] to compare observed outputs to predicted results.

- **When the model gets things wrong, how wrong is it and what are the implications?**
  In some cases, certain incorrect predictions are worse than other incorrect predictions for the particular problem being solved. In the above example it may be preferable to have a model that produces fewer false negatives (policies incorrectly predicted not to lapse) than false positives, as mislabelled policies that end up lapsing may result in financial losses (e.g., unrecouped acquisition costs and lower contributions to ongoing expenses). Consequently, it is important to validate not just accuracy but the implications when the model is inaccurate.

- **Is the model stable?**
  It is necessary to have a model that performs to a high level of accuracy (i.e., the performance metric) when running the model on different data sets, not just for a single set of test data. To do this efficiently, a common approach is to split the training data set into training data and validation data in numerous ways in order to maximise the amount of data that can be used to validate that the model is stable.[10]

- **Do I understand the bias in the model?**
  Predictive ML models are biased and discriminatory by nature; it is how they operate. They classify, categorise and look for observable patterns and discriminant features from a data set. However, as part of model validation, it is necessary to understand what those biases are, to document them and to ensure that those responsible and accountable for the ML model are comfortable with them from an ethical viewpoint and sign off on them. This falls into the realms of the explainability and interpretability of black box ML models amongst other ethical areas discussed further in Sections 3.5 to 3.8.

- **Is the model too sensitive?**
  The ML model needs to still perform well, in terms of making sensible predictions or decisions, in light of unseen scenarios or extreme scenarios that did not form part of the initial training data set, or that the model was not initially tested on. For example, an ML model based on investment market data may work reasonably well when markets are not volatile, but when markets move in extreme ways, it may start to react unpredictably or make decisions that a human would not.

Crucially, ML is not a 'one and done' model validation exercise in the pre-deployment stage of the model; certain ML models continue to learn over time in response to new data (i.e., reinforcement learning) and others need recalibrating over time to ensure they do not lose their predictive power.[11] Consequently, model validation is an ongoing process and must be considered at all stages of the AI system's life cycle that must feed into the model governance frameworks of a firm. We discuss this concept further in Section 3.7.

Finally, firms must put in place controls and governance structures that ensure that the risks associated with the use of ML models are identified, reduced, mitigated and monitored, as would be the case for other risks facing a firm. Although many of the concerns addressed by traditional model risk frameworks are relevant, there are a number of new risks that are unique to ML models owing to their complexity, continuous life cycle (and thus dynamic calibration) and ethical risks (e.g., bias and discrimination). See (Babel, Buehler, Pivonka, Richardson, & Waldron, 2019). Further discussion on this, in the context of ethical risks, can be found in Section 3.7.

[9] For example: R-squared, Root Mean Squared Error, Residual Standard Error, Mean Absolute Error, Akaike's Information Criteria.

[10] This is a process called k-fold cross-validation and is one of a number of different ways of testing stability.

[11] If you were to repeat the training of the ML model in one year's time, different characteristics may then be important in determining predictions.

2.5 NATURAL LANGUAGE PROCESSING, COMPUTER VISION AND SPEECH RECOGNITION?

NLP

We pass on information to others using human language through the words we say and how we say them. It is a task that humans find incredibly easy to do but that is much more challenging for a computer system. Some of the rules for passing on information can be extremely nuanced and abstract:

‘I’m really looking forward to this piece of work.’

Discerning whether the sentiment in this sentence is genuine or sarcastic would be reasonably straightforward for a human, but the passing of information through sarcasm would be more challenging for a computer system to discern.

NLP AI systems are machines that perform the intelligent process of understanding and processing human language. NLP systems aim to understanding both the meaning of words and phrases, and how the concepts are connected and structured to deliver an intended message. They use unstructured data sources (text, audio) and work to:

- **Identify and extract natural language rules** so that the unstructured language data can be converted into a format that a computer can understand and analyse.
- **Use various approaches to interpret the natural language** including statistical methods, rules-based systems and ML (for example, see (Saravia, 2018) for a clear overview of deep learning applications for NLP). It should therefore be noted that not all NLP systems use ML techniques.

Many modern NLP approaches will work by having multiple ML techniques working in tandem. For example, this could involve the use of various supervised or unsupervised ML algorithms for the following subfunctions:

- **Tokenisation:** Breaking up a text document or audio clip into pieces, known as tokens, that a machine can understand (e.g., words)
- **Part of speech tagging:** Identifying each token’s role in the part of the text (e.g., noun, adjective, adverb) and tagging it as such
- **Named entity recognition:** Identifying people, places, hashtags, email addresses and other named entities of interest in text or speech
- **Sentiment analysis:** Identifying the mood and opinions towards each of the named entities (positive, negative or neutral)
- **Latent semantic indexing:** Identifying words and phrases that commonly occur together

The text can then be processed in order to solve common NLP problems, such as document classification, text summarisation, machine translation (into other languages) and to answer questions (Sigmoidal, n.d.).

The final extension of NLP is natural language generation (NLG) and speech generation whereby the rules that govern natural languages can be used to generate original text or speech. For a comprehensive review of how deep learning algorithms can tackle this problem, see (Karpathy, 2015), and to be ‘terrified’ by this technology in action, see the *Guardian* article ‘A robot wrote this entire article. Are you scared yet, human’ that was written entirely by an AI system (GPT-3, 2020).

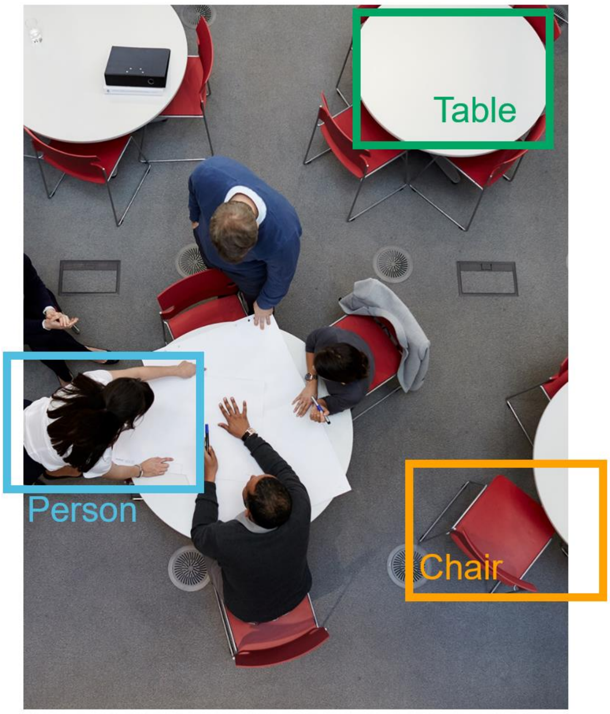
COMPUTER VISION

This is a subsection of AI that focuses on developing techniques that let machines capture and understand various digital images and video content and extract meaningful, contextual information from those images.

Computer vision is a broad field and includes:

- **Image classification:** Assigning a label to an image (e.g., handwritten digits)
- **Object detection:** Detecting objects of a certain class in an image (e.g., chairs, people, tables in Figure 4)
- **Facial recognition:** Identifying or verifying a person from a digital image (or a video frame)
- **Action and activity recognition:** Determining what activities or actions are taking place from a series or range of observations of the actions of people
- **Human pose estimation:** Determining the position of human joints from images, video or skeleton data

FIGURE 4: ILLUSTRATION OF OBJECT RECOGNITION



An effective tool for computer vision is the use of a deep neural network (an ML algorithm) that has been designed to process, correlate and understand the data in high-resolution images. See (Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018) for a comprehensive overview.

SPEECH RECOGNITION

Speech recognition is the technology behind understanding and interpreting spoken words from audio data.

Such AI systems work by identifying and interpreting words and phrases in the spoken word and converting these into texts where NLP can be applied to discern meaning from those words. NLP and speech recognition are therefore used together in hybrid systems such as voice assistants (Alexa and Siri). Other features, such as frequency and pitch, can also be captured.

There are a range of algorithms that can be used for speech recognition, but recent advances in the field involve the use of deep neural networks—for a comprehensive review see (Yu & Deng, 2015).

2.6 WHAT IS THE AI SYSTEM LIFE CYCLE?

AI systems are based on models and algorithms that require data for their training, validation and operation. In some ML applications, the system will be designed to continuously evolve over time: learning and recalibrating in order to make better predictions or make better decisions. Owing to these characteristics that differentiate the AI systems from more general system development, the concept of the AI system life cycle is common in many academic papers, frameworks and guidelines.

The AI Group of Experts at the Organisation for Economic Cooperation and Development (OECD) defines AI systems as having four distinct stages (OECD, 2019), as shown in Table 2. Each stage is a cross-collaborative effort involving a number of key stakeholders that will be involved in, or impacted by (either directly or indirectly), the AI system life cycle at each of the various stages.

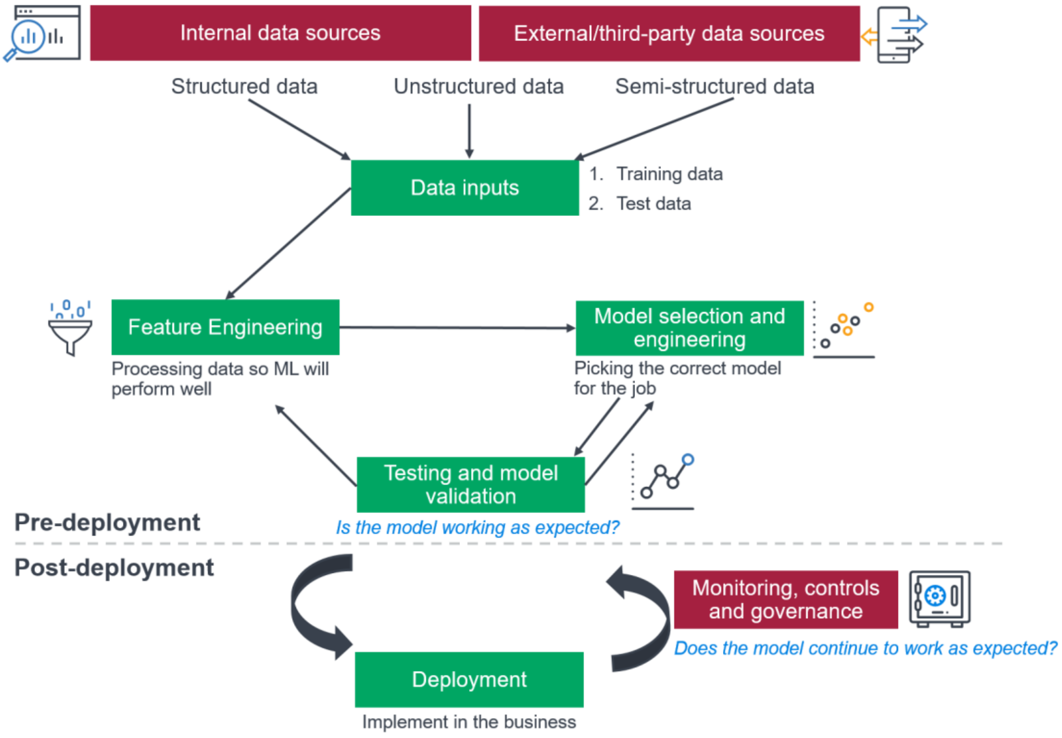
TABLE 2: STAGES OF THE AI SYSTEM LIFE CYCLE AND THOSE INVOLVED

STAGE	STEP	SOME KEY STAKEHOLDERS
1	Planning and design, data collection and processing, model building	Business activity leaders, end users (internal, external), developers (data scientists, data engineers, actuaries), third-party data providers, general public,[12] ethicists, domain experts
2	Testing, verification and validation	Developers (data scientists, data engineers, actuaries), business activity leaders
3	Deployment	Business activity leaders, IT
4	Operation and monitoring	Business activity leaders, end users (internal, external), risk managers, general public, IT, regulatory bodies, internal and external auditors of system, and wider oversight bodies

In Figure 5 we overlay the stages onto the phase diagram set out in Figure 2 to highlight these four stages.

As discussed earlier, these phases often take place in an iterative manner, are not necessarily sequential and form part of a ‘develop, use, analyse, redesign’ feedback loop. An AI system can be retired from use at any time once it is in the operation and monitoring stage—ending its life cycle.

FIGURE 5: THE STAGES OF THE AI SYSTEM LIFE CYCLE


Internal data sources
External/third-party data sources
Structured data
Unstructured data
Semi-structured data
Data inputs
1. Training data
2. Test data
Feature Engineering
Processing data so ML will perform well
Model selection and engineering
Picking the correct model for the job
Testing and model validation
Is the model working as expected?
Pre-deployment
Post-deployment
Monitoring, controls and governance
Does the model continue to work as expected?
Deployment
Implement in the business

[12] Other than those who will be potential end users of the AI system but that may be indirectly be impacted by it.

2.7 WHAT IS THE UPTAKE OF AI IN THE LIFE INSURANCE SECTOR?

There is no doubt: AI systems are increasingly being designed, developed and implemented for use in the insurance industry.

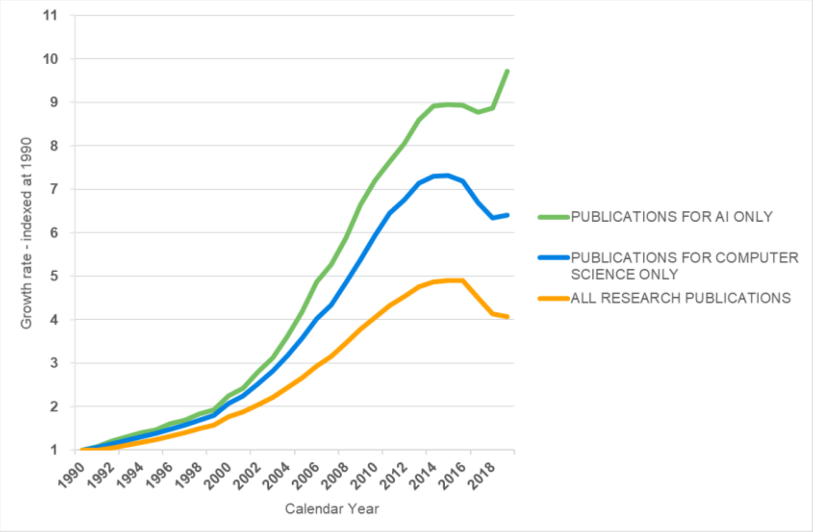
The U.K. insurers that responded to a 2019 survey by the Bank of England (BoE) and Financial Conduct Authority (FCA), referred to in this report as 'The 2019 BoE and FCA Study', (BoE and FCA, 2019) all reported that they have live machine learning applications in use, with the median respondent having 7.5 live applications.[13] Respondents expected this to triple to 21.5 applications over the next three years.

A 2020 European Commission (EC) study ('The 2020 EC Study') found that when studying firms across the EU member states, the finance and insurance sector had the highest reported proportion of firms planning to adopt AI in the next two years, at 27%[14] (European Commission, 2020b). The study concluded that the uptake of AI across Europe was on the rise.

Such trends are perhaps not surprising—they follow overall trends of increasing financial investment in AI research and development, and consequently a growth in research publications in the field of AI, as shown in Figure 6.

In 2019, the median **U.K. INSURER** had **7.5 LIVE ML APPLICATIONS, expected to rise** to **21.5 BY 2022**

FIGURE 6: GROWTH RATE IN AI PUBLICATIONS, COMPARED TO COMPUTER SCIENCE PUBLICATIONS AND ALL RESEARCH PUBLICATIONS AS BENCHMARKS



Source: Data from OECD.AI (2020), version of 02/07/2020, accessed on 22 September 2020, www.oecd.ai

[13] The use of machine learning for underwriting, investment, compliance, etc., would each be counted separately.

[14] The sector also reported an adoption rate that was broadly consistent with the average across all business sectors.

To maximise the benefit AI can bring to society and ensure they are not left behind, national governments are providing increasing amounts of funding for research. The U.S. and China are at the forefront of AI development, both in terms of the level of investment and research output, with the European Union (EU) aiming to catch up—see (Castro, McLaughlin, & Chivot, 2019). For example, for non-defence AI research and development, the latest figures[15] suggest:

- **U.S.:** Plans to spend $1 billion in 2020 according to 2020 fiscal budget (Bloomberg, 2019).[16]
- **China:** Spent between $1.7 billion and $5.7 billion in 2018[17] (Center for Security and Emerging Technology, 2019).
- **EU:** The EC planned to invest on average $0.6 billion[18] each year between 2018 and 2020, with a further $2.9 billion[19] to be invested between 2021 and 2027. See (European Commission, 2020d) and (European Commission, 2018e). This will be supplemented by national government spending in each of the member states.

However, although the uptake of AI in insurance, or more widely, financial services, may be on an upward trajectory, the uptake of this technology in the life insurance sector, rather than in general insurance or health insurance, is more challenging to determine.

It is the view of the authors that the uptake of ML has been slower in the life insurance sector, particularly in the U.K. Although the exact percentages are not clear, the 2019 BoE and FCA Study confirms this: In the U.K., general insurers had a higher proportion of live ML applications than life insurers and a greater skew towards more mature implementations (i.e., beyond the proof-of-concept stage and more embedded in terms of the proportion of the business that uses the technology).

Indeed, both private financing and mergers and acquisitions involving technology firms in the insurance industry reflect that the appetite for ML technology is currently focussed on general insurance applications. Research by FT Partners showed that whilst there has been a dramatic increase in financing for general insurance technology[20] (from $130 million in 2013 to $3.14 billion in 2019), life insurance investment has not seen such an exponential rise, with investment not exceeding $278 million per annum over the same period (FT Partners Research, 2020).

So, why is the uptake so much lower in life insurance when compared to the rest of the insurance sector?

To answer this, we need to review the structural factors that may have, to date, acted as barriers to the adoption in life insurance when compared to, for example, general insurance:

- **The problems that are being solved are inherently different, and thus so are the models.** Given the short-term nature of many general insurance products, pricing and reserving models are typically predictive models (e.g., Generalised Linear Models [GLM] in pricing (Yao, 2013), calculating loss reserves and Incurred But Not Reported [IBNR] reserves (Taylor & McGuire, 2004)), whereas life insurance pricing and reserving modelling involves cash flow projections. Owing to their nature, predictive models based on regression techniques like GLMs can more easily be extended to ML models, and some ML approaches are not a natural fit to life insurance modelling.

[15] Some of the targets and budgets may be revised in light of the COVID-19 pandemic, with funds instead moved to funding healthcare, and protecting businesses and the economy.

[16] It should be noted that the U.S. also committed to spending $4 billion to defence-related AI spending in 2020 (Bloomberg, 2019). Such investment could result in innovation that will cross over into domestic use in the future.

[17] It is more challenging to get a clear view on budgeted spend in China for AI research, but the study by the Center for Security and Emerging Technology provides estimates of the upper and lower bounds of China's AI investments for 2018 by cross-referencing its ministry of finance's national expenditure report with other government agencies' funding calls and abstracts from researchers describing agency-funded projects.

[18] €1.5 billion over three years (2018–2020) converted to U.S. dollars using the exchange rate as at 7 October 2020.

[19] €2.5 billion across the 2021–2027 budget, converted to U.S. dollars using the exchange rate as at 7 October 2020

[20] This study reviews investment in InsurTech—this will include technology that relies on ML and AI.

- **General insurance products are often shorter-term, life insurance products are long-term.** Given predictive models require data, in abundance, for training, the long-term nature of life insurance products may make them less amenable to certain ML modelling. For example, a life insurance predictive model that aims to determine if a policy will make a claim or lapse may take 10 to 20 years to confirm if the model is correct (i.e., whether the predicted event occurs). A comparative model for short-term general insurance would take just a year or so. By the time you know if your model has been working correctly, the calibration of the model will likely be out of date. This may restrict the possible use cases in life insurance to those that can be measured over a shorter time horizon.
- **Historical life insurance data may be more challenging to prepare for use in ML models.** Owing to the long-term nature of life insurance, data that may be of relevance to ML models (e.g., policyholder data, underwriting data) may be challenging to collect and collate as it may be from unstructured data sources (e.g., handwritten documents), in nonstandard formats (e.g., the format of a health assessment form may have changed a number of times over the last 20 years) or may be stored on multiple legacy systems that are not integrated. This would mean substantial work would be required to collate this data into a meaningful format, and even if possible, there could be an increased risk of data quality issues such as missing data or missing key data fields (e.g., the date a policy changed from in-force to paid-up was not directly captured by the system; it would need to be deduced from the available records).
- **Life financial reporting is typically at the aggregate level.** While it may be possible to model life insurance policy cash flows more accurately by adopting more granular assumptions about the policyholders, it may not be possible to use these assumptions in reserving and thus for the firm's financial reporting. For example, under certain regulatory regimes, the calculation of insurance liabilities may require the use of prescribed assumptions that are set by regulators. Even where this is not the case (e.g., Solvency II in Europe) insurers will be reviewing the performance of a particular block of business or product determined on an aggregate level, rather than at a policy level. When using aggregate assumptions, rather than more granular assumptions, the assumptions used to calculate the reserves for certain policies will be less accurate at a policy level, but over the portfolio as a whole these inaccuracies should net off. Thus, senior management may have limited incentive to change valuation model assumptions to be more granular. One of the main benefits of applying AI to big data can be to help differentiate individuals; when the difference cannot be shown in financial reporting, there is reduced incentive to determine more granular assumptions.
- **Firms and regulators may still be satisfying themselves that pricing or making business decisions on a granular basis is good and ethical.** Over the last 10 years, pricing in the general insurance market has become increasingly granular, almost to the individual level in certain markets. This has allowed conduct regulators, such as the FCA in the U.K., time to improve their understanding of the data and techniques used, and to determine how to regulate the practice through industry consultation (FCA, 2016). Generally, we are aware that there has been an openness from regulators to the idea of granular pricing in general insurance.

  For life insurance, however, the industry may be less ready to adopt more granular pricing. Despite more granularity that may result in more accurate pricing being possible[21] few life insurers appear willing to make such a change to their pricing practices. This may be in part due to the potential risk of discrimination when using alternative data sources to get more granular information on an individual (albeit this problem would also exist for general insurance), but also uncertainty on how conduct regulators may react to such changes in pricing practices.

  In addition, when it comes to in-force management, life insurers may be quite reluctant to analyse the in-force business at a granular level because they are not comfortable making business decisions that may impact individuals differently.

[21] For example, through the use of third-party data on an individual's financial situation, purchase behaviour or mortgage information, to model policyholder behaviour much more accurately at a more granular level. This will allow further risk factors to be used in pricing rather than just age, smoker status, sum assured, etc.

The application of data and advanced ML applications is currently more limited in life insurance than in general insurance. There is innovation in less complex areas such as marketing analytics and underwriting, but less uptake in other business areas where it is currently used in general insurance.

Some of these barriers are set to change: Traditional structured data (e.g., policyholder data, underwriting data) are becoming increasingly digitalised, and legacy systems are being replaced, insurers are capturing and storing more data on their policyholders from their interactions with them, and new structured and unstructured data sources are becoming available to life insurers (e.g., data from wearables, social media, individual finance data) to use in ML models.

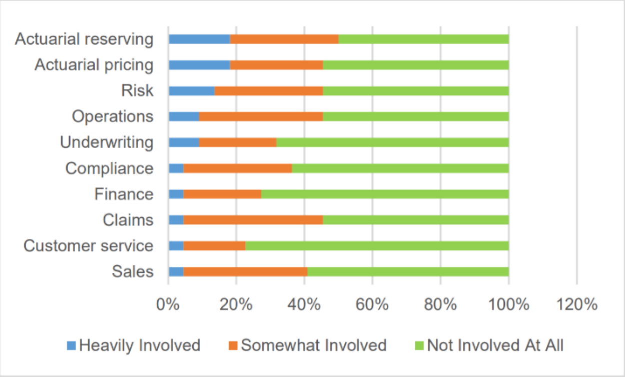
2.8 HOW IS AI USED IN THE LIFE INSURANCE SECTOR?

With much of the online literature on AI use in the insurance sector focusing on *transformation* (Vartak & Jain, 2019), on *potential* use cases (Deloitte, 2017), and hypothesising *what could the insurance world be like?* (Balasubramanian, Libarikian, & McElhaney, 2018). In identifying use cases of AI in life insurance it is challenging to discern fact from fiction, and the purely academic from the widely implemented and adopted business use cases.

Studies like the 2019 BoE and FCA Study, the 2020 EC Study and the 2019 Milliman Data Science Survey (Phelan, Murray, Tucker, & Comerford, 2019)[22] play a crucial role in cutting through this noise. Some key takeaways from these were:

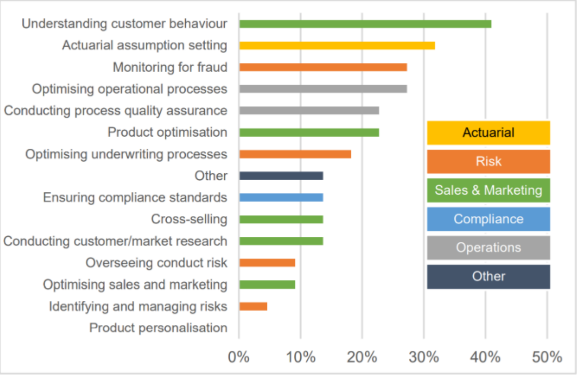
- **ML and AI are used in a range of business functions in life insurers.** The Milliman Data Science Survey identified that actuarial and risk roles are most heavily involved in data science applications (see Figure 7). The 2019 BoE and FCA Study also highlights that the most mature deployments in the insurance sector, outside of general insurance and life insurance use cases,[23] are in risk management and compliance, customer engagement and asset management.
- **ML tools and techniques are used to solve a number of different business problems.** These include better understanding policyholder behaviour, setting of actuarial assumptions (in pricing and reserving models) and fraud prevention (see Figure 8), based on the Milliman Data Science Survey. The 2020 EC Study is reasonably consistent with this; for the finance and insurance sector the survey found the focus of AI use in the industry was on anomaly detection (e.g., for fraud detection) (18% already using, 23% planning to use), forecasting, NLP, process automation, process optimisation and recommendation engines (12%–15% using, 13%–17% planning to use).

FIGURE 7: TO WHAT DEGREE ARE THE FOLLOWING AREAS OF THE BUSINESS INVOLVED WITH DATA SCIENCE APPLICATIONS?



Source: Milliman Data Science Survey (Phelan, Murray, Tucker, & Comerford, 2019)

FIGURE 8: FOR WHAT BUSINESS DECISIONS OR APPLICATIONS IS DATA SCIENCE CURRENTLY USED IN YOUR COMPANY?



Source: Milliman Data Science Survey (Phelan, Murray, Tucker, & Comerford, 2019)

Although there is research to suggest that in light of data-driven technologies such as AI, the insurance value chain will shift as a result of AI to a data-driven business model more akin to Big Tech companies like Facebook, Google, and Amazon (Zarifis, Holland, & Milne, 2019), a view strengthened by affirmative

[22] This study was conducted on 22 Republic of Ireland–based life and health insurers and reinsurers and focussed predominantly on the current and planned future applications of data science, including ML and AI.

[23] The exact definition of 'general insurance and life insurance use cases' is not clear in the research. It appears likely to include pricing, underwriting and reserving, and to exclude supporting functions.

comments from industry leaders, in this report we continue to look at use cases through the current traditional insurance value chain structure that many firms still operate under, or could be transitioning away from. This report therefore provides examples of use cases in the following business areas:

- **Product Design and Development (2.8.1)**
- **Sales, Marketing and Distribution (2.8.2)**
- **Pricing and Underwriting (2.8.3)**
- **Policy Administration and Customer Servicing (2.8.4)**
- **Investments (2.8.5)**
- **Claims management (2.8.6)**

The use cases we consider here are a combination of implementations and some that are more academic. We do not aim to discern the difference between the two; when looking through the lens of the ethical use of AI, it is important to be both reflective (‘*Are the current designed or developed AI systems adhering to ethical principles?*’) and preemptive in considering the implications for potential use cases (‘*If I were to design or develop an AI system to solve problem X, what are the ethical issues I need to consider?*’).

The use cases are intended to be illustrative, rather than exhaustive.

2.8.1 Product design and development

CREATING MORE PERSONALISED PRODUCTS

Akin to the structures of the gig economy, where you are paid for what you work, products offering usage-based insurance or real-time pricing are entering the insurance market whereby a policyholder only pays for what they use and the real-time risk they pose to the insurer.

In general insurance, pay-as-you-drive and pay-as-you-mile payment structures allow for a more tailored product that takes into account, depending on your plan, where you drive, how far you drive and, in certain cases using telematics, how you drive. This data is provided to the insurers in real time and allows them to assign a more accurate risk rating (and thus price) to the driver. This data flow provides the insurer with an opportunity to reward safer drivers with cheaper insurance, and provides riskier drivers with more transparent information on how they could achieve a cheaper premium.

Some examples of pay-as-you-live products in the U.K. life insurance market include:

- **Vitality:** Discounted life insurance premiums up-front if you complete the health review and commit to tracking your activity, and continued discounts that change year-on-year depending on your health and lifestyle.
- **Reviti (underwritten by Scottish Friendly):** Uses the SideKick Health app to capture data on a policyholder’s lifestyle and well-being. Reviti aims to incentivise positive lifestyle changes by offering policyholders a discount to their premiums if they give up smoking or move to alternatives such as heated tobacco or e-cigarettes.
- **YuLife (underwritten by AIG Life Ltd.):** Provides group life insurance, critical illness and income protection to companies, aiming to inspire employees and their families to build healthy habits by rewarding employees with YuCoins when their fitness tracker or apps show they have achieved high step counts or practised mindfulness.

The usage-based insurance model can also apply to life insurance through pay-as-you-live products. Some insurers actively track and collect data on the activities and lifestyles of their policyholders (e.g., through the use of wearable technology or fitness apps) such that when a customer shows lifestyle characteristics that will result in better health outcomes (e.g., regular exercise, high step count, practicing mindfulness[24]) the product is designed to reward the policyholder either through premium discounts or other benefits (e.g., cinema tickets, vouchers). These products are in the early stages of development and so any discounts may function more as a data acquisition strategy than a truly personalised pricing approach at present. However, the collection of such data will permit, in a few years, more accurate and granular pricing approaches.

ML algorithms can be used to analyse and make predictions on the risk profile of an individual in light of their lifestyle decisions, in order to allocate rewards to lower-risk policyholders. Examples of classification of policyholders using telematics for general insurance have been provided in (Gao & Wüthrich, 2019) and (Dong, et al., 2016), and (Janković, Savić, Novičić, & Popović, 2018) provides an overview of how deep learning methods applied to wearables data could be applied to the identification and classification of various health issues or characteristics.

Providing an incentive to live a healthier lifestyle is limited to products that have a significant element of mortality or morbidity risk for the insurer (e.g., term assurance); insurers that instead aim to take on longevity risk through annuities will not benefit from encouraging people to be healthier. This misalignment of incentives makes it very complicated for an insurer to offer this kind of product in the annuity market.

CREATING LIFE INSURANCE PRODUCTS WITH FEWER TOUCHPOINTS

High-touch sales models involve multiple interactions and are focussed on the human connection in sales and customer service, whereby firms aim to develop a close relationship with their customers. This is at the cost of speed and efficiency. The increased digitisation in all industries has created demand for the creation of quick-to-access, fewer touch-point products.

Such models are well suited to certain life insurance products that are reasonably standardised and simple (e.g., term assurances or critical illness) where increased contact may be seen as an unnecessary. AI systems can be used to create more efficient, fewer touch-point life insurance products by:

- **Speeding up the sales process:** For example, by using image and facial recognition to allow users to confirm their identity using a photo taken on their smartphone or for reviewing key documents (e.g., driving licence, proof of address) using text analytics and NLP.
- **Reducing (i.e., making less intrusive), or eliminating, the underwriting process:** See Section 2.8.3 on underwriting.
- **Providing instant guidance and answers to customer queries:** See Section 2.8.4 on customer servicing.
- **Quicker claims processing:** For example, by automating the processing of claims documentation (e.g., using NLP to review documents) and fraud detection—see Section 2.8.6.

2.8.2 Sales, marketing and distribution

EXISTING CUSTOMERS: IDENTIFYING CROSS-SELLING AND UPSELLING OPPORTUNITIES

Which products are suited to an individual will depend on their circumstances, and will develop over time. Some insurers envision a world where they use ML, applied to internal data and supplementary external data, to identify patterns in their policyholders' behaviour in order to predict the optimal time to provide a cross-sell or upsell opportunity. For example, after purchases from childcare stores appear in your financial data and your social media posts mention a 'baby shower,' a life insurer could determine you are at a suitable life-stage to consider purchasing a life insurance policy, and send you a targeted email or letter on your potential options.

[24] There are a number of mindfulness apps that provide guided mediation sessions. These apps can share data with an insurer on the duration of use, allowing mindfulness time to be quantified and rewarded.

This is through the use of systems commonly known as recommendation engines or recommendation systems and should be reasonably familiar to most: It is the technology used to recommend films, music, books, food, and items in online stores and streaming services, based on a variety of factors, such as the history of the consumer, but also the behaviour of other customers considered to have similar characteristics. There is a range of recommendation system algorithms currently use, which vary in complexity: For an explanation of some deep learning applications, see (Le, 2019).

Recommendation engines are beginning to be used in the life insurance sector; as part of Aviva's multifaceted algorithmic decision agent (ADA), built using the Dataiku data science platform, ML is used to determine which products Aviva's consumers are most likely to buy based on the data they can obtain on their customers (Dataiku, 2019).

POTENTIAL CUSTOMERS: MORE EFFECTIVE MARKETING

One of the biggest challenges when marketing life insurance is how to personalise the messaging to potential policyholders in a way that will resonate with them. ML makes it possible to compile and process the huge amounts of data coming from multiple sources (e.g., purchase behaviour, mobile app usage and responses to previous marketing campaigns) required to predict what marketing offers and incentives will be most effective for each individual customer and thus to provide timely, focussed marketing content (Segal, 2020).

POTENTIAL CUSTOMERS: ROBO-ADVISERS?

Robo-advisers, which have been implemented successfully in wealth management (see (Vanguard, 2020) (Fidelity, 2020)), are considered an opportunity to automate the process of guiding potential policyholders on the most appropriate product option(s) (e.g., type of policy, benefits), taking into account details on the policyholder and their unique needs (e.g., family, home, income and finances). Ultimately it aims to remove the need, and cost, for this guidance to come from financial advisers.

Examples, for simpler and more comparable product structures (e.g., term assurances, critical illness and income protection) are provided by brokers like PolicyGenius and Anorak Technologies. However, there may be more difficulty and resistance to the use of robo-advisers for more complex insurance products or those that have a significant impact on the customer's lifestyle (e.g., retirement options), where the face-to-face interaction and expertise of a professional financial adviser would be valued over guidance provided by an AI system.

2.8.3 Underwriting and pricing

IMPROVING THE SPEED AND ACCURACY OF UNDERWRITING USING EXISTING DATA

ML used to train models of mortality (survival models) can improve the speed and accuracy of risk assessments in underwriting whilst using the same historical internal insurance data held by insurers.

A model implemented by MassMutual showed an ML approach outperformed traditional underwriting by better subdividing lives by risk, such that the healthiest risk pool had a 6% reduction in deaths after 15 years compared with traditional underwriting (Maier, Carlotto, Sanchez, Balogun, & Merritt, 2019). This model is used across all of MassMutual's life products to assess 90% of 'lower risk' applicants,[25] bringing operational benefits of reduced time to policy issuance as well as the gain from better modelling. Similarly, predictive models for risk assessment in underwriting using Prudential Financial Inc.'s life insurance data for a range of ML models (multiple linear regression models, decision trees, random forests and artificial neural networks) were explored in (Boodhun & Jayabalan, 2018).

[25] Algorithmic underwriting is not used for policies with a benefit in excess of $3 million, where the applicant is age 60 or above, or where the health risk class is below standard.

INDIVIDUALISED RISK ASSESSMENT IN UNDERWRITING AND PRICING

Life insurance underwriting describes the process of assessing the health and financial circumstances of a potential policyholder to determine whether they should be provided with life insurance coverage, and if so, the appropriate price for that individual to pay for such coverage.

The focus of underwriting is primarily on medical underwriting designed to determine the health or medical status of a potential policyholder. Medical underwriting is carried out where there is mortality or morbidity cover in an insurance contract and is used to determine whether the health of the potential policyholder is good enough for insurance acceptance on standard terms and, if not, how poor the individual's health is relative to the benchmark set for the standard terms.

- In 2017 **MetLife** announced it was partnering with Digital Fineprint, a U.K.-based InsurTech whose technology would allow insurers (with customer permission) to look through data on social media platforms like Facebook and Instagram in order to identify and flag information that can be used to determine risk factors (e.g., an individual's physical shape or smoking habits) (Digital FinePrint, 2017).

Broadly speaking, potential policyholders are placed into homogeneous groups based on similar medical and health conditions, such that they can be treated consistently in terms of whether a policy can be offered (and if so, on what terms). Traditional pricing and underwriting techniques limit this to relatively few pricing bands (e.g., based on distribution channel, age, benefit size, occupation). Insurers can use ML on existing or new data sets (including unstructured data) to develop a greater range of risk factors and refine its current pricing bands. A greater number of risk pricing bands should result in fairer individual premiums by taking into account the diversity of human life in determining a particular individual's own mortality/morbidity risk.

ALGORITHMIC UNDERWRITING

More bespoke forms of medical underwriting (e.g., medical exams, testing) are costly and seen as a potential barrier to entry as they are time consuming and intrusive. AI systems have been used by some life insurers to reduce the need for invasive medical information by using alternative data sources to identify risk indicators.

In particular, ML can look at the wealth of health data held by insurers from historical medical assessment forms, and the existence or absence of subsequent claims, to identify patterns in the information that may not have been identified. Such internal data is supplemented by third-party data on individuals such as medical prescription histories, driving records and credit history in order to determine the underwriting decision. In any cases where the AI system cannot make a decision (or is not empowered to do so) mechanisms can be incorporated to notify a human underwriter.

- **Haven Life Insurance Agency**, a subsidiary of MassMutual, applies ML to MassMutual's internal structured and unstructured data, and to third-party data such as medical prescription histories, driving records and credit history in order to offer real-time algorithmic underwriting, allowing some customers to buy life insurance online in just minutes without a medical exam (Dignan, 2017).
- Other companies offering algorithmic underwriting include **Bestow** (with policies issued by the North American Company for Life and Health Insurance), **Ladder** (with policies issued by Fidelity Security Life Insurance Co. and Allianz Life Insurance Company of New York**)**, **Ethos** (with policies issued by a range of insurers including Legal & General America) and **Manulife**. See (Bestow, 2020)**,** (Ladder Life, 2020)**,** (Ethos Life, 2020) and (Manulife, 2018).

AUGMENTED UNDERWRITING

For firms that do not use a fully automated algorithmic underwriting approach, AI is still being used in parts of the underwriting process, augmenting the traditional underwriting process. One of the main challenges of life insurance underwriting is that underwriters must process large amounts of unstructured data from various sources—images, emails, call centre data and documents. An individual's medical history may, for example, contain hundreds of pages of information that an underwriter would need to sift through in order to identify any items relevant for their decision making.

To speed up this process, AI techniques such as NLP and ML can be used in the processing and assessing of unstructured text information from any handwritten application forms, doctors' notes and health assessments; computer vision can be used to process images; and speech recognition can be used to assess call centre data, to draw out key information to assist the underwriter, and accelerate the underwriting process.

UPDATING PRICING ASSUMPTIONS

ML can be used to derive the assumptions for use in traditional pricing models.[26] One example is in the forecasting of mortality; (Hainaut, 2018) uses neural networks to forecast mortality rates in France between 2001 to 2014 using a model trained on general population mortality rates between 1946 and 2000. The model showed that the mortality forecasts based on neural network models had a higher predictive power when compared with a range of Lee-Carter models, the industry standard approach. Such an approach was applied to U.K. mortality data in (Richman, 2020b).

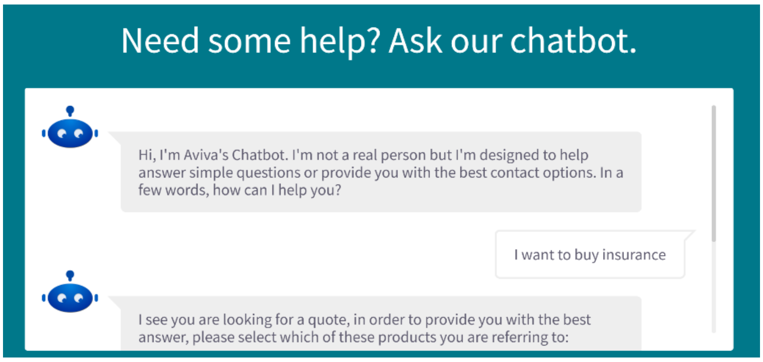
Richman also explains that ML can be used to create a company, or product, specific mortality table since it has been shown (see (Tomas & Planchet, 2014)) that the creation of such tables can be transformed into a regression problem that ML models are well placed to tackle (Richman, 2020a).

2.8.4 Policy administration and customer servicing

SELF-SERVICED CUSTOMER SERVICE

After sale, AI is being used by insurers to improve the customer experience at various touchpoints in the lifetime of their policy. Widespread implementation of AI includes virtual assistants and chatbots, and in call centres where NLP, NLG, speech recognition and speech generation technology is being used to understand questions and responses from policyholders and to provide responses back to individuals in conversational text or audio. These systems need to retain information throughout the conversation, such that anything relevant mentioned earlier can be used to help answer subsequent questions.

FIGURE 9: CHATBOT EXAMPLE: AVIVA—CONTACT US, SEPTEMBER 2020



Need some help? Ask our chatbot.

Hi, I'm Aviva's Chatbot. I'm not a real person but I'm designed to help answer simple questions or provide you with the best contact options. In a few words, how can I help you?

I want to buy insurance

I see you are looking for a quote, in order to provide you with the best answer, please select which of these products you are referring to:

When this technology works well, it serves customers by allowing instant self-servicing of queries regarding their insurance policy that is not constrained by the working hours of a call centre team.

Increasingly such tools aim to make it clear when a user is interacting with a machine and the limitations of what it can do, see Figure 9 taken from the Aviva website from September 2020. The chatbot makes it clear at the onset of the conversation that it is 'not a real person' and is there to answer 'simple questions or provide you with the best contact options' rather than give advice.

Alternatively, insurers are looking for ways to benefit from AI systems developed by third parties to improve customer engagement, allow for greater self-servicing, and to create easier interactions between themselves and their policyholders. For example, Aviva has recently provided the option to pair its app with Amazon's Alexa to allow policyholders to find out their latest pension value without needing to log in to the MyAviva website or app (Aviva, 2020).

[26] The assumptions can also be considered when determining the reserving assumptions.

POLICYHOLDER BEHAVIOUR AND RETENTION

AI can be used to determine patterns in policyholder behaviour, including the likelihood of exercising options available to them on their life insurance contract (e.g., conversions, guaranteed annuity option take-up, surrenders, lapses).

Particularly relevant is lapse risk, the risk of policyholders lapsing any time before the maturity of their policy such that the insurance company incurs a loss.[27] Therefore, predictive models for lapses ('lapse prediction' or 'churn prediction' in various academic texts) have been developed, aiming to identify the policy-specific factors, company-specific factors or macroeconomic factors that drive lapses, in order to better predict and identify any policyholders who may be at a heightened risk of lapsing. This may allow insurers to act proactively in order to retain individual or groups of policyholders through the creation of early warning indicators that can trigger communications to at-risk policyholders.

TABLE 3: SOME EXAMPLES OF THE USE OF AI FOR LAPSE PREDICTION IN INSURANCE

AI APPROACH	AUTHORS
ML (SVM, DECISION TREES) AND OTHERS	(Morik & Köpcke, 2004)
ML (LOGISTIC REGRESSION, CART)	(Milhaud, Loisel, & Maume-Deschamps, 2010)
ML (LOGISTIC REGRESSION, DECISION TREES, NEURAL NETWORKS, SVM)	(Siemes, 2016)
ML (RANDOM FORESTS, LOGISTIC REGRESSION, NAIVE BAYES)	(Babaoglu, Ahmad, Durrani, & Bener, 2017)
ML (XGBOOST, CART, SVM, LOGISTIC REGRESSION)	(Loisel, Pierrick, & Tsai, 2019)

There is a range of academic studies outlining various approaches to creating predictive ML models for lapses (e.g., logistic regression, decision trees, neural networks, SVM applied to insurance data) (see Table 3). Predictive lapse models that incorporate economic factors for lapse rate prediction (i.e., different lapse rates for different scenario simulations) can be used to evaluate the time value of options and guarantees for reserving. See (Aleandri, 2019).

Although the use of ML models for customer churn prediction is widely researched and widespread for other industries (e.g., the telecommunications industry (Amina, et al., 2019), e-retailers (Subramanya & Somani, 2017) and music streaming services (Gregory, 2018)), there are a number of issues that arise when applying such techniques to life insurance customer data sets. These arise from policyholder data being skewed to non-lapses (i.e., only a few policies actually lapse for the model to use in training, testing and validating models), highly dimensional (i.e., there are a lot of attributes—sales channel, age, sum assured, date of issue) and time-stamped (i.e., they are point in time snapshots of policyholder data, whereby one data set is related to the next).

Even where data is available to predict policyholder behaviour at a granular level, the extent to which insurers will use such analytics will depend on the incentives they have to do so. For example, where this data cannot be reflected in the assumptions used in financial reporting, or its use to make business decisions on an individual policy level is not desirable,[28] firms may see little reason to perform such analysis.

Rather than assessing each individual's risk of lapse, an alternative use of AI could be to assess broader macroeconomic trends correlated with lapses, and use this to better gauge lapse risk at a portfolio level. AI could be used both in pattern detection and to assess a broader set of data sources (e.g., performing sentiment analysis on social media).

[27] This may be due to the company not having sufficient time to recover any acquisition expenses and commissions on the contract, or that the reserve released on lapse is less than any surrender payout made to the policyholder, thus the lapse results in a loss to the firm. The loss of a policy also means a reduction in the future policy charges that would otherwise be available to meet future expenses, and the lost potential for selling riders or additional policies.

[28] For example, ML applied to big data can be used to predict lapses at a granular level which can be used in turn to estimate a customer lifetime value. This could be a useful tool in determining how to manage its in-force business with the aim of maximising its value to the firm. However, life insurers may be quite reluctant to analyse the in-force business at a granular level because they are not comfortable making business decisions that may impact individuals differently.

RESERVE APPROXIMATION, CAPITAL MODELLING AND DAILY SOLVENCY MONITORING

The majority of life insurance contracts can be valued in a reasonably straightforward way using discounted cash flow models built using a range of software solutions, from Microsoft Excel to proprietary asset liability modelling software. However, in certain cases, alternative techniques to estimate the output of a more complex cash flow projection model may be appropriate, and these could potentially utilise AI. Such techniques may be considered in cases where:

- Stochastic simulations considering a large number of possible future scenarios are required (e.g., for with-profits contracts and other contracts with options and guarantees)
- Nested simulations are required (e.g., in the calculation of the Solvency II capital requirements) that may necessitate, or otherwise increase, the number of model runs and simulations required
- Decision makers in insurers require an understanding of the solvency position more frequently than monthly or quarterly valuations would allow (i.e., daily solvency monitoring) in light of changes in the economic environment

Proxy modelling has been in use in the U.K. and Europe for several years, principally to support the calculation of the solvency capital requirement (SCR) in Solvency II internal models. Such modelling has also enabled insurers to achieve daily solvency monitoring, to approximate a value of their liabilities and assets on a daily basis to get more up-to-date approximations of their solvency position, allowing them to be more reactive to changes in the market. The overall goal of proxy modelling is to create a relatively simple function to estimate the output of a more complex cash flow projection model, such that a loss function (the fit of the model to the data) is minimised. In order to fit the model, algorithms can be used that remove the need for expert judgement as to the ‘best’ model structure. Although, a Least Square Monte Carlo (LSMC) approach to proxy modelling is considered by many as the present state-of-the-art methodology for the insurance industry, ML alternatives (including random forests and neural networks) to LSMC are presented by (Kopczyk, 2019) and (Castellani, et al., 2018).

Alternatively, there are certain products where an accurate valuation of the policy is required regularly—variable annuity products are typically hedged daily, for which a daily valuation is required. A novel solution to this problem that uses a neural network to derive the Greeks (i.e., the effect of small changes in market variables required to dynamically hedge the portfolio) for a specific contract is presented in (Hejazi & Jackson, 2016) and (Doyle & Groendyke, 2019).

2.8.5 Investments

Many life insurers either have wealth and asset management subsidiaries, investment teams or outsourced investment activities that are focussed on managing the assets of their policyholders in line with an agreed trading objective. Strong investment returns are important in attracting and retaining savings business, generating returns that allow for financial guarantees, and to fund bonuses on with-profits contracts. They are also a key component in the generation of capital for the shareholders, or owners, of a company and are a key driver in the profitability of the firm.

In 2019, the CFA Institute, a global association for investment professionals, provided a comprehensive review of the current state of implementation of ML and AI in the investment management sector (the 2019 CFA Report). See (Cao, Preece, & Baker, 2019). The key applications of AI for investment management that were identified in the 2019 CFA Report were:

- **To generate new data:** To extend the data that can be analysed in order to support investment decisions, by using NLP, computer vision and speech recognition to process text, images and audio data respectively
- **To support systematic investment strategies:** Using machine learning to improve the algorithms that are used by quantitative investment managers in their processes
- **To support active fund management decision making:** To get new investment insights for active fund managers by using AI techniques to process big data, including the use of alternative and unstructured data (e.g., satellite images, earnings conference call recordings and transcripts, social media postings, consumer credit and debit card data, and e-commerce transactions)

Some examples of the use of AI and ML in investment processes are provided below.

To enhance trading strategies

ML algorithms for pattern recognition based on macro features (i.e., patterns in economic data like interest rate trends, changes in government policies), fundamentals (i.e., focusing on company-specific events) or market inputs are used by some in order to determine strategies on what to buy and sell, and at what time, or trade signals (i.e., triggers that indicate whether it is a good time to buy or sell). Such algorithms may include constraints (e.g., allowable trades in line with strategy, diversification from other investment strategies).

For example, reinforcement unsupervised deep learning can be used to find what are considered to be optimal trading strategies, and these can be incorporated into algorithms that trade with no, or limited, human oversight—'algorithmic trading.' See (Zhang, Zohren, & Roberts, 2019) and (Théate & Ernst, 2020).

Alternatively, algorithms may take the information harvested from various sources (using NLP, computer vision, speech recognition) and apply machine learning algorithms to generate signals, or alerts for human analysts and portfolio managers, to inform the investment decision-making process.

An example of this is provided in the 2019 CFA Paper from Bloomberg, which, via the Bloomberg Terminal, provides a time series of sentiment towards most asset classes using a wide range of data sources.

To make investment decisions quickly

High-frequency trading systems aim to make automated investment decisions (based on price movements and market conditions) and execute trades in fractions of a second.

Since the decision making is made by a computer system, it can be considered an application of AI. ML algorithms are also being considered for such systems, see (Borovkova & Tsiamas, 2019) and (Murillo & Ricardo, 2016).

Predicting uncertain events

AI can also be used to produce forecasts that can inform traders' investment decisions.

For example, the results of the 2016 European Union referendum vote could have been predicted through the use of sentiment analysis on tweets, as shown in the Milliman paper Emerging risk analytics—Application of advanced analytics to the understanding of emerging risk (Cantle, 2017).

This would allow investment managers to adjust their portfolios in light of this sentiment data, to improve their resilience to changes that the uncertain event may have on markets (e.g., in this case, movements in exchange rates, changes in U.K. stock market performance).

2.8.6 Claims management

Insurers can automate and improve operational efficiency of their claims processes using AI in order to reduce the time between notification and payment of a life insurance claim by:

- Using NLP to read and process handwritten documents in any claims documentation
- Using NLP and computer vision to process images and visual data (e.g., an image of the insured's death certificate and validating against death certificates on government websites)
- Using NLP and NLG in chatbots and virtual assistants to resolve simple questions and queries on claims

However, the most widely adopted use of AI in the insurance claims management process is in enhancing the detection of fraud. Life insurers face the risk of paying out money on fraudulent claims. There is a range of types of fraud that life insurers are susceptible to: On the one hand, opportunistic underwriting fraud involves misinformation on health assessment forms or other application documents that result in incorrect underwriting or lower premiums than had a truthful declaration been made (e.g., not disclosing that the insured smokes or is diabetic). Fraud is costly for the insurance sector—according to the Association of British Insurers (ABI), U.K. insurers found 107,000 fraudulent insurance claims in 2019, worth £1.2 billion (ABI, 2020).[29] The ABI estimates insurers invest at least £200 million as an industry each year in fraud prevention (ABI, n.d.).

[29] Albeit the majority were general insurance claims.

On the other hand, life insurers face the risk of professional claims fraud (such as money laundering) whereby criminals attempt to buy savings insurance contracts with money generated from criminal activities, before surrendering their policy to create the impression their assets are from a legitimate source. In the U.K., various legislation requires insurers to have strict controls in order to identify unusual or suspicious transactions which they believe could be connected to criminal activity.[30]

An issue for fraud prevention is that it can be a time-consuming process, involving collecting and aggregating structured and unstructured data from several areas of the company. Traditionally, standard statistical methods such as regression and discrete choice models have been employed (Artís, Ayuso, & Guillén, 2002). Increasingly, supervised and unsupervised ML approaches are being developed and used by insurers—some examples from academic literature are presented in Table 4.

AI systems working in fraud detection have different goals they wish to achieve, for example to identify:

- **Unknown fraud:** Unsupervised ML techniques can be used with the goal of identifying anomalies, or outliers in data, that suggest unusual behaviours that may be indicators of fraud, and notify human claims experts that they deserve further investigation.
- **Suspected fraud:** For claims that have been received, such algorithms aim to answer the question of *'Could this be a fraudulent claim?'* based on known types of fraud. Supervised ML algorithms can be trained on large data sets of claims known to be fraudulent or legitimate. This way, the AI system learns what indicators of fraud look like in the labelled fraudulent claims. It can then flag future claims with similar characteristics as suspicious. Alternatively the classification of future claims can be carried out by knowledge-based systems based on expert advice or features that were previously determined by ML that classify policies following a set of rules. Rather than simply presenting a binary classification, an AI system may aim to provide the likelihood or probability of a claim being fraudulent, based on its analysis of past insurance claims.

TABLE 4: EXAMPLES OF THE USE OF AI IN INSURANCE FRAUD DETECTION

AI APPROACH	AUTHORS
KNOWLEDGE-BASED SYSTEM (FUZZY LOGIC)	(Stefano & Gisella, 2001)
ML (SVM)	(Kirlidog & Asuk, 2012)
ML (MULTIPLE)	(Palacio, 2019)
ML (NEURAL NETWORK)	(Johnson & Khoshgoftaar, 2019)
ML (NEURAL NETWORK)	(Xu, Wang, Zhang, & Yang, 2011)

Such techniques ultimately will output either a suspicion score and/or visual anomalies on the claims data it is reviewing. Alternatively, it may provide the insurer with a set of rules or patterns that it can use to better understand the signs of fraud in various sets of data.

Recently insurers have also considered the use of alternative and emerging data sources to help reduce fraud. In March 2020, Zurich announced a partnership with Carpe Data purchasing its ClaimsX solution in order to:

> 'Leverage publicly available web data for real-time assessment and automated decision making, from first notice of loss to settlement.' (Zurich, 2020)

Such data includes a claimant's web presence, social media data, email data, interests, cell phone details and professional skills. In a life insurance setting, tools like this can use the claimants' personal details such as names, emails and birthdays to look for information on the Internet that might show whether they lied on their policy application (e.g., about smoking or drug use).

[30] For example, the Money Laundering Regulations; Proceeds of Crime Act; Criminal Finances Act; and Money Laundering, Terrorist Financing, and Transfer of Funds Regulations.

2.9 KEY TAKEAWAYS

- Research and investment into AI and ML and related business applications are growing and accelerating, with increases forecast for the financial services sector.
- Features unique to ML models and AI systems create new risks to life insurers.
- ML models aim to discriminate—that is how they have predictive power. The features that such models use to determine predictions will raise ethical questions around bias, fairness and discrimination against groups and individuals.
- AI systems are increasingly allowing value to be extracted from unstructured data sources that were previously inaccessible, including some data sources that individuals might not choose to share with insurers (e.g., social media data). This raises ethical questions around data privacy and data security.
- ML models can be extremely complex (i.e., neural networks) and so it is hard to understand how they are making decisions (i.e., they are black boxes). When decisions by such models can have adverse impacts on policyholders, it raises the question of whether the lack of explainability and transparency in decision making is ethical.
- In many cases AI systems are now making the decisions that humans would have taken, with limited or no oversight. This raises the question: When things go wrong, who is accountable?
- Unlike other modelling approaches, or other systems, AI is often a living system that needs regular validation and calibration to ensure the model is continuing to work in line with expectations. This increases the burden on model risk management and necessitates the creation of dynamic controls, checks and governance structures that review and re-review the AI system throughout its life cycle.

With more innovation anticipated from life insurers that want to capitalise on the benefits of AI, ethical considerations should be at the forefront of the minds of business activity leaders, digital transformation teams, developers and, ultimately, risk managers.

3. Ethics and the life insurance industry

This section provides a reasonably high-level introduction and discussion of some of the broad ethical and professional considerations that stakeholders involved in AI in the life insurance sector may come across, including:

- **3.1**—What is ethics?
- **3.2**—Why is it important that life insurers care about ethics in AI?
- **3.3**—What are the main ethical concerns associated with AI systems?
- **3.4**—Should life insurers care about ethics in all AI use cases?
- **3.5**—Key ethical consideration: Transparency
- **3.6**—Key ethical consideration: Fairness, bias and discrimination
- **3.7**—Key ethical consideration: Accountability
- **3.8**—Key ethical consideration: Explainability
- **3.9**—Key ethical consideration: Data privacy and data security
- **3.10**—Public understanding and engagement
- **3.11**—Competence and professionalism

We conclude with some key takeaways.

3.1 WHAT IS ETHICS?

Ethics is defined by the Oxford Dictionary as:

> ‘Moral principles that govern a person’s behaviour or the conducting of an activity.’

Simply put, it is a code of conduct, a set of principles or guidelines that aim to provide an individual (or group of individuals) clarity on what is morally right and what is morally wrong as they make decisions and lead their lives. However, ethics has a historical, cultural and religious context (Britannica, 2020), with no single, concrete definition that applies to all people, everywhere. Modern Western philosophy on ethics tends to fall into one of three main areas:

1. **Meta-ethics:** This focuses on the high-level questions on the nature and origins of moral judgement and ethical principles (i.e., whether ethical judgments are truths about the world or actually reflections of those who create the principles).

2. **Normative ethics:** This involves the setting of standards or criteria, norms or codes of conduct for moral judgments (e.g., saying if an action is right or wrong, that a person is good or bad, or that a situation is just or unjust). This is concerned with debates on what one ought to do in a given situation. For example, should an activity be judged right or wrong solely on the basis of its consequences and the outcomes of the activity, or by whether the person followed a certain rule? It raises into question whether you should do what is good even when it is not in your self-interest.

3. **Applied ethics:** This focuses on the application of normative ethical theories to practical problems (e.g., war, animal rights, capital punishment and the use of AI) taking into account different views on how these theories should be applied to particular use cases.

The ethics of AI bridges the latter two areas; it addresses questions on normative ethics such as ‘*Should I refrain from using social media data in an AI system because it is not right to infringe on personal privacy, even if it would reduce fraud, cut cost and boost the bottom line?’* but also aims to ensure a common, systematic application of those principles to practical problems through the creation of guidelines, standards, frameworks or shared industry standards and norms.

Applied ethics in AI is much debated, is highly contextual (i.e., to a particular use case), may be informed by the cultural norms of a given country and may be temporal and change over time in line with public opinion. For these reasons, it is a significant challenge for firms to get right; this is made easier when companies, industry bodies, regulators and governments, supported by public debate, come together to determine what standards to apply for a particular use case.

Questions on AI ethics span both immediate ethical issues that firms can address (e.g., ensuring currently used or in-development AI models do not lead to unfair discrimination), medium-term concerns (e.g., the impact of the loss of jobs in insurance due to AI-driven automation) and quite long-term debates (e.g., the possibility of creating AI systems that match or surpass human intelligence). The focus of our discussions are on the immediate applied ethical issues facing life insurers, as firms look to ensure that they balance innovation with the consideration of ethical standards and norms.

3.2 WHY IS IT IMPORTANT THAT LIFE INSURERS CARE ABOUT ETHICS IN AI?

The straightforward answer is that since ethics aims to determine what is morally right and results in the betterment of society, considering ethics is thus the morally right thing for firms to do. In particular, the challenges presented in the key takeaways in Section 2 are relevant to life insurance use cases—there is a real potential to cause harm to policyholders and to infringe on their personal freedoms. In a democratic system, insurers need to engage in the debate of what is morally right, and what is not, to ensure the use of AI does not create or exacerbate any problems for society. Insurers therefore have a moral responsibility to their policyholders and to wider society to consider ethics.

However, it is not simply a corporate social responsibility issue for life insurers. The benefits of using AI in life insurance (e.g., improved risk pooling, better risk management and prevention of fraud), both for policyholders and for the firms that implement them, will not be achieved unless the public has trust in the firms that are using AI, in how their data is used and in the decisions made by AI systems.

Gaining public trust is particularly relevant to the insurance sector with its long-standing public trust issue; in 2019 most U.K. policyholders were looking for another provider and felt that insurers try to avoid paying out legitimate claims (YouGov, 2019), and an earlier poll suggested that 32% of people in the U.K. would trust insurance companies with their data (much below banks [57%] and healthcare providers [64%]) (Open Data Institute, 2018).

In recent years, there have been headlines over ethics and the potential dangers of AI technology, from both Big Tech firms and insurers, all acting to undermine public trust, including:

- The fatal crash of Tesla's autonomous vehicle (BBC, 2018)
- The Facebook and Cambridge Analytica scandal (The Guardian, 2018)
- The use of Facebook data for pricing car insurance (The Telegraph, 2016)
- Racial discrimination in pricing algorithms (The Sun, 2018)

There is a risk that the failure to develop and deploy AI in line with the public's ethical expectations will transpire into a lack of public trust that will not only damage a firm's reputation and thus its ability to continue to sell products (i.e., a reputational risk) but also have a systematic impact on the sector. Indeed, following the Facebook and Cambridge Analytica scandal, a study published by *The Atlantic* showed that trust was not only broken in Facebook, but around 26% of people also changed their behaviour on other social media platforms (e.g., by being more careful in what is shared, deleting accounts) (The Atlantic, 2018).

We discuss how insurers may engage the public in order to develop public trust further in Section 3.10.

Finally, ethics can be seen as a differentiator, a unique selling point and a means for competitive advantage. In the U.K., ethical consumer spending was at a record high in 2019. See (The Guardian, 2019), with consumers increasingly determining how to use their purchasing power to shape markets. Indeed, it is possible to find websites that rank U.K. insurance companies on their ethical credentials, ranking firms on their commitment to the environment, people and whether they have an ethical investment policy (Good Shopping Guide, n.d.). If such trends continue, and the public becomes more educated on the use of AI generally, consumers may opt to purchase life insurance from companies that can evidence the ethical use of AI (e.g., certification, meeting standards).

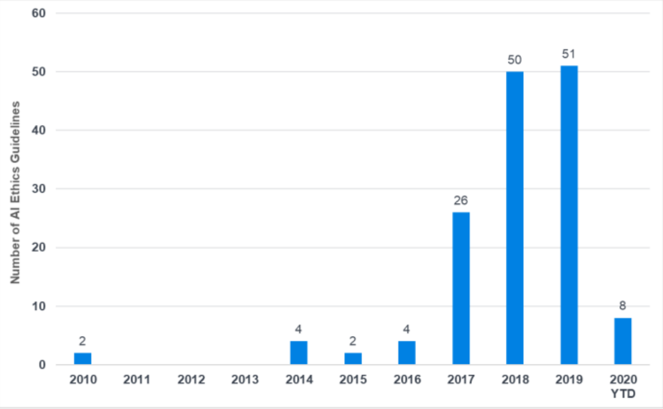
3.3 WHAT ARE THE MAIN ETHICAL CONCERNS ASSOCIATED WITH AI SYSTEMS?

A subject of ongoing debate concerns what ethical AI looks like, and how an AI system can be developed, deployed and used in a responsible way, worthy of public trust.

Consequently, over recent years there has been a proliferation of guidelines from academia (e.g., (University of Montreal, 2018)), the private sector (e.g., (Google, n.d.) (Microsoft, n.d. a)), government bodies (e.g., (Monetary Authority of Singapore, 2018)), industry bodies and think tanks (e.g., (Insurance Europe, 2020) that aim to provide guidance on how to develop AI in an ethical, moral and responsible way (see Figure 10 and Figure 11).

The challenge for life insurers that use, or are considering the use of, AI may have therefore shifted from a struggle to find guidance on what best practices look like, to one of how to deal with the saturation of information that is now out there—*‘So what guidelines should we follow exactly?’*

FIGURE 10: NUMBER OF AI ETHICS GUIDELINES PUBLISHED OVER TIME



Source: Data collected from (Algorithmwatch, 2020). Only 147 papers included in the chart out of 167 in the database at the date of download. This is because certain guidelines are not dated and so could not be included.

FIGURE 11: ETHICS GUIDELINES BY SECTOR OF PUBLISHER

Source: Data collected from (Algorithmwatch, 2020). Based on 167 publications.

However, based on a literature review we have conducted on a subset of the available AI ethics guidelines (see Appendix 2), despite this saturation, it appears that there has been a convergence to a number of core principles (see Table 5).

In particular, most guidelines focus on the need for **transparency, fairness, accountability, explainability, data privacy and security (data and prevention of harm and misuse).**

Given these topics are applicable to the use of AI in the context of life insurance, we discuss these core concepts further in this paper.

These core areas are reflected in the names of groups that are focusing on researching AI ethics, for example FAT (fairness, accountability, transparency) and FATE (fairness, transparency and ethics in AI). See (FAT ML, n.d.) and (Microsoft, n.d. b), respectively.

TABLE 5: PERCENTAGE OF GUIDELINES REVIEWED THAT COVERED ETHICAL TOPICS

Transparency, openness, communication	94%	Robustness (reproducibility, accuracy, reliability)	50%
Fairness, nondiscrimination, diversity	83%	Human rights, human autonomy, etc.	39%
Accountability (auditability, reporting negative impacts, redress)	83%	Human oversight and control	33%
Explainability, interpretability	72%	Public awareness, education on AI and its risks	33%
Data privacy	72%	Impact on employment	28%
Data security and cybersecurity	67%	Potential use in weapons	22%
Societal and environmental well-being	61%	Cultural considerations (e.g., different groups and cultures may have different ethical views)	22%
Prevention of harm/misuse of AI	56%	Competence and skills of those using AI	17%

Many of the frequently mentioned ethical topics are those that are already researched, measurable and thus practicable for insurers that are considering the use of AI. As we will discuss later, there is a growing wealth of expertise, research and development into identifying and mitigating discrimination and bias in the use of AI systems, measures and mechanisms for accountability, and tools that can help with the explainability of models. Furthermore, data security and privacy are issues independent of the specific use in AI and so are widely researched and embedded in data laws and regulations.[31]

Thus, what is perhaps more interesting is what is not included, or what ethical areas featured less in our review. In particular:

- **Technical guidance:** Although guidelines, frameworks and recommendations provide a high-level vision towards which firms can start to develop their own internal policies and practices, those that we reviewed provide little to no guidance on how to implement these practically. There was the use of high-level examples (IBM, 2019) and some reference to tools and techniques that could be considered to meet the principles (European Commission, 2019), but generally this level of guidance was narrow or thin, leaving firms to work this out for themselves. This is likely due to the variety of audiences targeted by some of these guidelines that therefore requires generalisation, may be due to a desire to not be prescriptive, or reflect the ongoing debate about what best practice looks like around these areas. This leaves the onus on life insurers to translate these principles into practice themselves.
- **Ambiguity:** For a given ethical principle, there is ambiguity around the definition of key principles. The guidelines reviewed do not provide the authors' definition of the various key concepts, again putting the onus on life insurers to determine the definitions for themselves. For example, an insurer may aim to be fair in its underwriting processes that use AI, but by which measure of fairness should you use—do you aim to treat everyone equally, or in an equitable way in your particular use case?[32]
- **Trade-offs:** There were limited acknowledgement and guidance on the substantial challenge of dealing with trade-offs between the various ethical concepts. A challenge beyond simple adherence to guidelines and requirements is to understand, deliberate and document any trade-offs between the various ethical principles for a particular use case. The guidelines were not always clear on what you should do if you cannot find an ethically acceptable trade-off in the development stage—do you proceed anyway or stop development?
- **Public deliberation:** Although many of the guidelines acknowledged the need for the public to trust in AI, few of the guidelines reviewed actually provided any guidelines or requirement to include public deliberation in the design, development and ongoing review of AI systems.
- **Skills, competence:** Generally, the guidelines did not aim to establish the need for those making decisions on the development or deployment of AI to have undergone any training requirements or to meet any professional requirements. Given the complexity of AI models themselves and the ethical considerations when using AI, coupled with the potential for harm to consumers when things go wrong, it is not unreasonable to require minimum standards for those looking to develop or deploy AI in a business. In the absence of guidance, such decisions on the requisite skills and professional accreditations and general training in a firm on AI and ethics will fall to the insurer.
- **Less on governance:** Although there was a heavy focus on accountability throughout the guidelines reviewed, they did not always provide extensive, or clear-cut, guidance on how a company should actually change its organisational structure such that there are clear mechanisms in place for decision making on the ethical development and deployment of AI—who exactly should be made accountable and in what circumstances, and what governance structures could result in effective risk management for ethical risks faced by firms that are using AI.

[31] Such as the General Data Protection Regulation in Europe (European Union, 2016c) and the California Consumer Privacy Act, which applies to California consumers (incorporating Silicon Valley firms) (State of California Department of Justice, 2018).

[32] Or one of a range of other fairness measures. See Section 3.6.

- **Time frame:** The focus of the ethical guidelines was on current applications, the more technical 'What can we do now?' rather than aiming to set in place guidelines for potential medium to longer-term ethical concerns (e.g., the impact of AI and automation on job markets, AI superseding human intelligence).
- **Cultural aspects:** Ethical values are subjective and vary amongst different groups and across cultures. In the guidelines reviewed there was limited acknowledgement or guidance on how to develop AI that accounts for the cultural norms in the various markets in which a firm may operate.

In considering the usefulness of such guidelines for life insurers, we conclude that there is a need for greater development of use-case specific guidance and norms on what is morally right and what is morally wrong (in line with public expectations) and the development of industry and use-case specific established norms for tools, mechanisms and governance structures used to ensure the ethical use of AI. In such case, domain-specific expertise from insurance industry bodies (e.g., the ABI in the U.K.), insurance regulators (e.g., the Prudential Regulation Authority [PRA] and FCA), in conversation with the public that they serve (to ensure cultural aspects are considered and to ensure trust is built), is needed. In light of this, we provide an update on the current guidance provided by various groups that are relevant to U.K. life insurers in Section 4.

3.4 SHOULD LIFE INSURERS CARE ABOUT ETHICS IN ALL AI USE CASES?

Insurers in Europe may be well-versed in the need for proportionality; the Solvency II insurance regulation that governs the industry has been criticised by some as not fairly applying proportionality across the industry, resulting in onerous and costly compliance[33] (Insurance Europe and AMICE, 2019).

Similarly, a one-size-fits-all approach to ethical considerations for all AI use cases is not necessarily a correct approach to take. Instead, a case-by-case, risk-based approach is required for the application of controls and governance in the development, deployment and ongoing monitoring of AI systems, with efforts focussed on those use cases that can cause the greatest harm to policyholders, or to society as a whole.

The challenge then for life insurers, in light of limited guidance from regulation or industry best practices, is to determine the objective criteria to judge the risk presented by the use of an AI system in a firm. That is, they need to conduct some form of an ethics-based risk assessment.

In line with the work of (The Geneva Association, 2020) and extending to the use cases presented in Section 2 of this report, we present in Table 6 an illustrative risk-based approach to allow proportionality in the governance of ethical issues. In this example, one could consider the impact of the introduction of an AI system in terms of whether that AI system:

- Creates new logic for making decisions that humans could not have determined in coming to a conclusion or producing its outcome—where issues of fairness, transparency and communication, accountability and redress, and explainability of model outcomes arise.
- Uses, or gives rise to the use, of alternative data sources where issues around data privacy and data security arise, or where there is less general consensus on whether the data is appropriate for use in the particular problem. In supervised ML applications, data labelling errors or biases can also present an issue when using new data sources.

This risk-based framework would then place at lower risk (and so lower control and governance requirements) those use cases where the AI system automates tasks but does not change the logic of decision making in any way, and at higher risk those use cases where new data sources are used, or when the use of AI changes the logic of decision making.

[33] Consequently, proportionality issues are being addressed as part of an ongoing review into Solvency II (EIOPA, 2019).

In fact, such an approach, although a useful starting point, may not necessarily capture all the requisite details in order to make an ethics-based risk assessment. Firstly, such an approach does not quantify in terms of an expected financial loss, or maximum financial loss, the risk that such AI systems pose to an insurer. For example, the use of financial data by algorithmic trading has been marked as ‘medium risk’ in Table 6, owing to the use of data that is currently available and used by traders. However, the risk of potential harm to the insurer could be material if there are unusual market events that the AI system has not been trained on that results in automated trading behaviour that generates substantial losses for the firm. Secondly, it does not aim to distinguish between the potential harm to the firm, against the potential harm to policyholders (or any other stakeholder groups). For example, although an individual’s unfair decision on underwriting a policy may not translate into a news event that damages the firm’s reputation, this could result in a significant impact on the individual’s livelihood. Finally, even in the use of existing data, any AI system that is continuously learning may face greater challenges of robustness, model drift, reproducibility and risk of adversarial attack than those that are recalibrated and validated at less regular intervals. Such considerations are not captured in this approach.

TABLE 6: AN ILLUSTRATIVE RISK-BASED ASSESSMENT OF AI USE CASES IN LIFE INSURANCE

	LOWER RISK	MEDIUM RISK		HIGHER RISK
	DECISION-MAKING LOGIC AND DATA USED UNCHANGED	**DECISION-MAKING LOGIC UNCHANGED, DATA USED CHANGED**	**DECISION-MAKING LOGIC CHANGED, DATA USED UNCHANGED**	**DECISION-MAKING LOGIC AND DATA USED CHANGED**
PRODUCT DEVELOPMENT AND DESIGN	▪ Lower touch products—speeding up sales or claims process by using NLP ▪ Lower touch products—using chatbots	▪ Lower touch products—speeding up sales process using facial recognition		▪ Creating more personalised products (using lifestyle data)
SALES, MARKETING AND DISTRIBUTION			▪ Existing customers—identifying upsell and cross-selling opportunities using existing internal data	▪ Providing product guidance using a robo-adviser ▪ Existing customers—identifying upsell and cross-selling opportunities using existing patterns derived from new external data sources ▪ Potential customers—more effective marketing (through targeted marketing campaigns and advertising) using ML to derive patterns in nontraditional data
PRICING, UNDERWRITING	▪ Augmented underwriting—using NLP, computer vision and speech recognition for processing data required for underwriting		▪ Applying ML methods in the derivation of pricing assumptions ▪ Individualised risk assessment in underwriting and pricing using existing data sets ▪ Algorithmic underwriting that applies ML to traditional data sources	▪ Individualised risk assessment in underwriting and pricing using new data sets ▪ Algorithmic underwriting that applies ML to alternative or nontraditional data sources
POLICY ADMIN, CUSTOMER SERVICE	▪ Self-service customer service using NLP, NLG, speech recognition and speech generation in chatbots and to answer calls		▪ Applying ML methods in the derivation of policyholder retention metrics (e.g., lapse rates) using existing data (e.g., policyholder data). ▪ Reserve approximation, capital modelling, daily solvency monitoring	▪ Applying ML methods in the derivation of policyholder retention metrics (e.g., lapse rates) using alternative data sources

	LOWER RISK	MEDIUM RISK		HIGHER RISK
	DECISION-MAKING LOGIC AND DATA USED UNCHANGED	DECISION-MAKING LOGIC UNCHANGED, DATA USED CHANGED	DECISION-MAKING LOGIC CHANGED, DATA USED UNCHANGED	DECISION-MAKING LOGIC AND DATA USED CHANGED
INVESTMENTS		▪ Using NLP, computer vision and speech recognition to gain investment insights to inform investment decisions from alternative data sources (e.g., tweets, audio data)	▪ High-frequency trading systems and ML use in automated trading (i.e., replacing human decision making), based on existing data sources used by traders ▪ Using ML to gain investment insights to inform investment decisions, based on current data sources (e.g., economic) ▪ Using AI to develop trade signals for human analysts, traders and portfolio managers that can be used in their investment decision-making process, based on traditional data sources	▪ High-frequency trading systems and ML use in automated trading (i.e., replacing human decision making) based on alternative data sources (e.g., tweets, news stories) ▪ ML algorithms that detect patterns in alternative data sources (e.g., tweets, news stories) that traders can use in their decision making ▪ Predicting uncertain events using sentiment analysis on alternative data sources like tweets and news stories
CLAIMS	▪ Using NLP to read handwritten documents in claims process ▪ Using NLP, computer vision to process images and visual data (e.g., an image of a death certificate and validating against death certificates on government websites) ▪ Using NLP and NLG in chatbots and virtual assistants to resolve simple questions and queries on claims	▪ Fraud detection—identifying known fraud cases using alternative data sources (e.g., social media data)	▪ Fraud detection—identifying known fraud cases using ML algorithms on claims data ▪ Fraud detection—identifying known fraud cases using a rules-based system on claims data	▪ Fraud detection—identifying unknown fraud using alternative data (e.g., social media data)

Thus, in order to apply a risk-based proportionate approach to ethics and AI, firms need to first determine and document their own criteria, with risk-based metrics, the thresholds and trigger points for such metrics, and what controls (e.g., human oversight, tools, mechanisms) and governance are applicable for each risk band. In doing so, firms could consider consulting domain experts, seeking public engagement, and benchmarking against industry best practices as available.

3.5 KEY ETHICAL CONSIDERATION: TRANSPARENCY

The ultimate goal of transparency in the use of AI is to present its use as justified and credible, and to have that perceived in the eyes of others by providing them with insight on what has been performed, and why. This was nicely put in (IBM, 2019), which alluded to us not trusting people without an understanding of their reasoning, and the same rules apply to AI.

Although most ethics guidelines outline a commitment to openness and transparency in the use of AI, this will mean different things to different stakeholders that will need different levels of insight, communicated in ways that are interpretable to them, in order to ultimately gain comfort in the use of AI (see Table 7).

TABLE 7: SOME INSURANCE STAKEHOLDERS, THEIR NEED FOR TRANSPARENCY, AND INTERPRETABILITY CONSIDERATIONS

STAKEHOLDER	MOTIVATION FOR TRANSPARENCY	TYPE OF DISCLOSURES NEEDED	INTERPRETABILITY
Public (policyholders, potential policyholders)	▪ Allows the public to seek appropriate redress and contest decisions that are made about them by an AI system ▪ To build trust that the AI system is working for their benefit, not to their detriment ▪ To facilitate the public's understanding of where, how and why AI is being used, to build trust	▪ **When** AI is being used to make a decision or recommendation that could impact them ▪ **When** the public is interacting with an AI system (e.g., chatbots) ▪ **How** AI is being used to make a decision or recommendation that could impact them (i.e., a record of its decision-making process, key factors or drivers) ▪ **How** a decision or recommendation made by an AI could be changed (e.g., what factors would result in a different outcome) ▪ **What** data sets were used by the AI system to arrive at its decision ▪ **Why** an AI system is being used, including its intent, capability, purpose and limitations	▪ Needs to be jargon free, easy to understand, useful and personal to their circumstances ▪ Needs to reflect and accommodate for a range of levels of understanding on AI ▪ Needs to make clear where help with understanding the outcome can be found ▪ Does not need to see source code of an AI model (too technical)
External certification bodies[34]	▪ To facilitate a review of the inputs, processes and validation methods used by developers of AI systems	▪ Data sets, model, algorithms, documentation on checks, validation and ongoing governance may be required	▪ More technical documentation is required that will assume a strong understanding of AI
Prudential regulators (e.g., PRA)	▪ To build trust that the AI system is robust, works as intended, with sufficient governance in place for ongoing review ▪ Allows regulators to build risk-based, proportionate regulatory guidelines or frameworks on the use of AI in insurance	▪ **Where** AI systems are being used (or may be used) in applications that could impact the financial security of a firm or impact the decision making of management (e.g., reserve calculation, daily solvency monitoring, investment decisions) ▪ **What** risks are created by a firm's use of AI systems in applications that may impact the financial security of a firm ▪ **How** the governance and controls contribute to sufficient risk management ▪ **What** are the key factors that drive an ML model output (e.g., to assess how the capital requirements would change in light of different economic conditions)	▪ Reasonably high level ▪ Some technical information may be required to support industry-wide consultation and to allow the regulator to provide guidance on best practices and its expectations
Conduct regulators (e.g., FCA)	▪ Allows regulators to investigate and contest the use of an AI system if it feels that it is unfair, discriminatory or not meeting other conduct of business requirements ▪ Allows regulators to build risk-based, proportionate regulatory guidelines or frameworks on the use of AI in insurance	▪ **Where** AI systems are being used (or may be used) in applications that could create new issues around fair outcomes for consumers ▪ **What** risks are created by a firms' use of AI systems in applications that may impact the ability to achieve fair outcomes for consumers ▪ **How** the governance and controls contribute to sufficient risk management around conduct ▪ **What** are the key factors that drive ML model output in AI use cases where the fairness of outcomes may be an important consideration	▪ May be high level ▪ Some technical information may be required to support industry-wide consultation to allow the regulator to provide guidance on best practice and its expectations ▪ Some technical information may be required for investigating issues of individual fairness to determine the source of the issue
Legal bodies	▪ Allows greater understanding of who is accountable when things go wrong	▪ **Who** in the firm is responsible for various parts of the AI system, including the quality of the data, model quality assurance, sign-off etc.	▪ Roles and responsibilities need to be clearly defined

[34] Responsible for reviewing the training data, model, testing, validation, documentation and governance to determine whether it meets certain standards and so can certify such standards.

Despite this, there is general convergence in guidelines to a few core requirements that life insurers should consider:

- **Explainability:** Explaining the technical process of the AI system, and tracing how it makes decisions
- **Interpretability:** Communicating explanations to interested parties in a way that they will understand
- **Traceability:** Recording data and model use to facilitate transparency
- **Communication:** Communicating openly to interested parties the information that they would want to know on AI systems

It should be noted that transparency around AI may not be needed in all use cases, such as where it has minimal impact on consumers (e.g., AI systems that convert handwritten paper documents into digital records) or financial markets.

EXPLAINABILITY

Insurers need to be able to: (1) explain the technical process of an AI system and (2) trace, and explain, how the AI system makes its decisions, in light of data to interested parties.

We discuss this in further in Section 3.8.

INTERPRETABILITY

This relates to the need for any communications made to stakeholders being understandable by those stakeholders, taking into account their level of understanding and requirements for transparency. Such communications should help policyholders better understand how AI is used, and thus, translate into increased public trust in the technology.

In particular, decisions, recommendations or other outputs of an AI system need to be understandable by interested parties.

Considering the use of AI in pricing or underwriting decisions as an example, although disclosing source code of the AI system would be considered transparent and open, this is not understandable (or of interest) to the general public, and likely to be proprietary to the firm. What the potential policyholder needs is an understanding of how underwriting decisions are made, with a range of detail that an insurer may be able to provide:

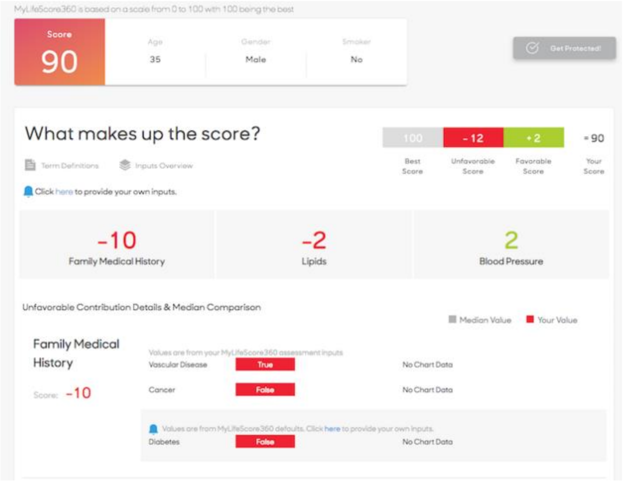
- Your premium was determined by an AI system that made its decision based on features X, Y and Z and used the following data items taken from sources 1, 2 and 3.
- Your premium was determined by an AI system and increased due to feature X increasing, but features Y and Z didn’t impact the price.
- Your premium was determined by an AI system and increased due to feature X increasing, but features Y and Z didn’t impact the price. Your premium would have been lower if your value of X was lower and/or your values of Y and Z were higher.

With this increased understanding, consumers will be empowered to dispute decisions made that impact them, or know how they can take (honest) actions available to them (for example, you cannot change your family history) to generate a more favourable decision in the future.

For example, via its InsurTech subsidiary LifeScore Labs, MassMutual has launched the MyLifeScore360 tool, which allows policyholders to see how various risk factors affect their life insurance premiums (LifeScoreLabs, 2020), allowing them to make more informed decisions (see Figure 12).

Being able to communicate in this way may be quite challenging, owing to the complexity of models, where the relationships between features and model output may not be linear.

FIGURE 12: MYLIFESCORE360 TOOL (LIFESCORELABS, 2020)



MyLifeScore360 is based on a scale from 0 to 100 with 100 being the best

Score	Age	Gender	Smoker
90	35	Male	No

What makes up the score?

Best Score	Unfavorable Score	Favorable Score	Your Score
100	- 12	+ 2	= 90

Click here to provide your own inputs.

Family Medical History	Lipids	Blood Pressure
-10	-2	2

Unfavorable Contribution Details & Median Comparison

Family Medical History

Score: -10

Values are from your MyLifeScore360 assessment inputs

Vascular Disease	True	No Chart Data
Cancer	False	No Chart Data

Values are from MyLifeScore360 defaults. Click here to provide your own inputs.

Diabetes	False	No Chart Data

Insurers may also be concerned about how customers would perceive the revealed information. If this helps customers understand what drives pricing or underwriting decisions, and feel more in control of these drivers, this positive engagement could provide a competitive advantage. However, there is also the risk that this transparency could instead undermine relationships with customers. For example, it could highlight the use of a commonly used rating factor that customers object to the firm using, unaware that it is standard across all insurers.

Finally, it may be necessary to ensure that any relationships between variables discovered by AI models are reviewed by those with sufficient subject matter expertise to be able to interpret them in a meaningful way and communicate findings to relevant stakeholders in the company. Whilst the model may tell you the relationship between variables, it may not tell you why there are relationships or why they act as they do; models may throw up counterintuitive results, and it may take a subject matter expert to interpret these results and assess whether the model is wrong or has discovered something new. If insurers fail to interpret their models in ways that people can understand, people will not have the confidence to use them, and some important business applications may be overlooked.

TRACEABILITY

Insurers need to ensure that any data sets, models or processes that make up an AI system are documented to a high standard in order to help with error-checking and investigation.

Namely, if AI fails to meet its aims, people need to know why, in order to:

- Understand where bias or discrimination arose in the AI system (e.g., the training data, the model features) in disputes around fairness
- Help those developing AI to correct any faults when a flaw in the system is identified
- Support the assigning of responsibility to those accountable in the event of disputes

However, openness and traceability need not be just ex-post considerations after an AI ‘failure’ but also ex ante; insurers can consider providing relevant information to external certification bodies to support an audit of the system (of which firms can notify policyholders) or disclose information on data and models used to prudential or conduct regulators to help support their understanding on what is being done and allow them to consider what regulation would be proportionate to ensure the use is in line with their aims.

COMMUNICATION

Firms need to be transparent with their policyholders, or potential policyholders, when they are talking to, or having decisions made (that could impact them) by, an AI system. This is particularly relevant when new algorithms or models are to be used, or new data sets that policyholders, or potential policyholders, may be unaware of.

This is important even in the use of very complex models where it is not necessarily possible to completely discern and explain why an AI system has made a particular decision, recommendation or arrived at an outcome. In such cases, notifying consumers of the intent to use an AI system, and how it will try to work, is still a meaningful step towards transparency—‘*Your premium will be determined by an AI system that will determine your risk profile based on your social media data, shopping history and credit history.’*

Furthermore, insurers could provide the option to opt out of interacting with, or having a decision made by, an AI system, in favour of interactions or decisions made by a human. Firms would need to consider how cost-efficient it would be to have both AI systems and human teams available to make decisions that could impact the lives of those interacting with the AI system. Indeed, providing a process for a human to review the decisions made by an AI system may be a more cost-efficient and realistic option.

OTHER CONSIDERATIONS

A challenge for firms will be to decide on, and put in place, measures for measuring and reviewing transparency. This may include a qualitative review to confirm that AI systems in operation (e.g., all chatbots) are notifying the public that they are interacting with an AI, and quantitative measures on traceability—that data and models are being documented such that they can be called on for error-checking or to resolve disputes.

Further consideration is required for insurers that are looking to either purchase external data (derived using AI or for use in an AI system) or use ML models or full AI systems from third parties. In such cases, it may make it more challenging for insurers to be open and transparent around the inner workings of its AI systems, as such systems may be protected by intellectual property rights. Insurers should consider this and how this may limit their ability to meet the transparency expectations of various stakeholders.

Firms also need to consider how the need for transparency to various stakeholders may constrain the use of AI systems for a particular use case; for example, the accuracy of the system may need to be sacrificed by using a less complex ML model, in order to ensure that the model decision making can be explained to consumers or regulators.

The style and level of detail of any public communications that aim to increase transparency must be weighed against the fact this could result in potential new risks (e.g., cyber-risk, to game the system) or confusion.

3.6 KEY ETHICAL CONSIDERATION: FAIRNESS, BIAS AND DISCRIMINATION

Insurance differentiates; it inherently is designed to group and classify people based on their risk to the firm, and to offer them coverage, or not, at a higher or lower price, dependent on their characteristics.

To enhance trust and confidence in the use of AI in life insurance, it is also important that there is—and is perceived to be—fairness in AI applications and that such applications do not unfairly discriminate against certain groups of customers.

Still, it can be difficult to define fairness, especially when the world around us is not fair. An interesting example comes from research showing that an online image search for CEOs returned results where 11% of the top images were women, while at that time, women accounted for 27% of CEOs in the U.S. Kate Crawford, cofounder of the AI Now Institute at New York University, raised the question of what the appropriate percentage would be—should the search return 27% images of women, in line with reality, or 50% images of women as would be the ideal percentage (Crawford, 2018)?

Most of the ethics guidelines we reviewed included fairness, bias and discrimination as key ethical concerns facing those developing and deploying AI systems; what is clear is that certain issues are unique to AI systems, and that fairness, bias and discrimination are intrinsically linked and thus must be considered in unison. Given AI's reliance on data, it is exposed to the bias that such data may have. Given AI's reliance on humans in feature engineering and model development, it is exposed to the bias that human modellers may have (conscious or unconscious). This bias can create unfair decisions, outcomes or predictions, and may result in differentiation (some of which will be fair, some of which will not be fair [i.e., they may be discriminatory]).

It is therefore useful to consider: the types of bias that an AI system could be susceptible to, the types of unfairness that the AI system could give rise to and the types of discrimination an AI system could create.

For life insurers, considerations of bias, fairness and discrimination will particularly be relevant when AI systems are used in making any decisions, recommendations or predictions that can impact people's lives (e.g., pricing, underwriting, claims management and product recommendations) where the risk of harm to consumers is greatest. Particularly in cases where ML may be used at scale (e.g., replacing the underwriting function completely) even small issues of bias can impact a large number of individuals.

BIAS

Since AI systems work on identifying patterns, the goal for those developing AI systems is to determine the bias that approximates the patterns in the data that one would want to identify, from those that are discriminatory or unintended. Commonly, it is considered that there are two main places where bias can arise in the model:[35]

- **Data:** When a ML model has been trained on data sets that contain existing societal prejudices, or where the data set is not sufficiently representative of the population that will be interacting with the AI system (i.e., it excludes certain demographic groups reflecting societal bias). It means that regardless of the model used, any statistical prediction model relying on pattern recognition will exhibit the same kind of biases present in the training data.
- **Modelling decisions:** Where the bias is not present in the input data at all and is added by the algorithm itself through the choice of its features and parameters.[36]

Data is often heterogeneous, comprising lots of subgroups that each have their own distinct characteristics. The goal of many ML algorithms is to discern the unique characteristics of each of these groups. A model that learns on biased data, however, may lead to decisions that are unfair, or inaccurate, or that discriminates against a particular group.

There are many forms of bias that can exist in data, some of which can result in unfair outcomes when used to train a ML model; in Table 8 we present some of these, including illustrative examples of how such bias could emerge in life insurance data sets and in some cases how this can transpire in creating unfair or discriminatory outcomes of the model. It should be noted that this is just a small selection of bias—(Meharabi, Morstatter, Saxena, Lerman, & Galstyan, 2019) list 23 types of bias that AI systems could be susceptible to.

[35] Although it is possible to also create bias at other stages (e.g., at the design stage where humans decide the purpose of the model). Imagine a model to optimise the repricing of a yearly renewable-term assurance; it could be computed to either maximise profits or maximise renewals. If a model determined that offering higher premiums than could be offered to new customers, or versus its competitors, to certain groups that are less likely to contest was a good way of maximising profits, it would end up learning this bias and engaging in predatory behaviour, even if that wasn't the intent. It is how the problem was framed that created the bias.

[36] Deciding what features to include, and which to ignore, can significantly influence the model's accuracy but also could introduce bias.

TABLE 8: TYPES OF BIAS—ILLUSTRATIVE LIFE INSURANCE EXAMPLES

BIAS	DESCRIPTION	ILLUSTRATIVE LIFE INSURANCE EXAMPLE
HISTORICAL	The data reflects reality, is perfectly sampled and representative, but historical bias in society is present in the data set.	Males purchasing life insurance have higher wages than females. This reflects historic wage inequality. AI system for recommending life insurance products may learn that women earn less than men, and offer a lower sum assured than for an equivalent male.
SAMPLING	The data used to train the ML model is finite, and thus does not fully reflect reality. Bias arises from the choice of training and test data, and the ability to fully reflect the whole population. The training data does not represent the environment that the model will operate in. Lack of individuals from under represented backgrounds in the training data.	Historically, fewer life insurance policies had been issued to minority groups at an insurer. Consequently they are underrepresented in a policyholder data set used to train an automated underwriting tool. For future applications using the tool, the system cannot determine an automated underwriting decision for policies from that minority group. It therefore refers the decision to a human underwriters to review, making the process longer for individuals from the minority group, and opening up their application to greater scrutiny than other policyholders.
MEASUREMENT	Arising from how the data items in a data set are chosen, used and measured.	If underwriters historically showed more leniency in accepting Caucasian people (compared to non-Caucasian groups) with high risk medical conditions for insurance (i.e., there is bias in their measurement of risk), an algorithmic underwriting system trained on this data may learn to also be lenient in this way.
AGGREGATION	When conclusions are drawn for a subgroup of a population that are not correct, based on observing the conclusion in other subgroups.	Pricing premiums using data provided from a health app which implies that individuals who eat a high number of calories have a higher risk of certain health conditions. Whilst this may be true for a certain subgroup of the population, this would not apply to a subgroup of athletes. Such a tool would unfairly discriminate against this subgroup, and if you were instead to breakdown the data into subgroups based on level of activity, different conclusions would be drawn.
GROUP ATTRIBUTION	A tendency to generalise for an entire group something that is thought to be, or is, true of some individuals that belong to that group. This includes in-group bias, where people have a preference for those with similar characteristics to themselves or to members of a group to which they belong, and out-group bias, a tendency to stereotype individual members of a group to which they do not belong, or to assume the characteristics of that group are more homogeneous.	Developers working on an insurer's activity tracker app choose to assign more 'reward points' (i.e., indicating a healthier activity) to activities in which they themselves partake to keep healthy. This is in-group bias as they assume that since they maintain their health by carrying out those activities, others who do those activities must be healthy too (which may not be the case).
SIMPSON'S PARADOX	A trend, association or characteristic observed in subgroups may be quite different from when these subgroups are aggregated.	A ML tool may learn from aggregated data that males, overall, are less likely than females to purchase a particular life insurance product. Thus, less marketing is spent in targeting males than females which may not be fair. However, the relationship between take-up rate and gender could be the result of other factors. For example, products sold by a financial adviser might have lower sales volumes, generally, than those products sold by the company's own sales team. Looking only at those who bought directly, sales volumes for men and women might be equivalent, or males could represent a higher proportion of sales. However, if men are predominantly buying via financial advisers and women predominantly buying directly, this results in a different relationship between gender and take-up rate at the aggregate level than observed at the sales channel level.
LONGITUDINAL DATA FALLACY	Treating cross-sectional data as though it is longitudinal, which may create bias through Simpson's paradox.	A ML model trained on policyholder data may identify that policyholders are choosing shorter policies over time on average. It therefore learns to advise policyholders to take up a shorter policy even if this is not appropriate for their situation. However, if you were to break the data down into subgroups, then you may find that for each sub-group the average policy term stays the same, or increases, and as such different conclusions should be drawn.
EXCLUSION/ OMITTED VARIABLE	Deleting valuable data that was thought to be unimportant, or the systematic exclusion of certain information.	Most of your policyholders are single life policies, so you delete joint life policies from a data set used to train a ML pricing model. However, this means the model will not identify features that are appropriate for joint life policies and would lead to a lower premium appropriate for certain policies.
OBSERVER	When developers project their expectations onto the data or model. This can be from the labelling of data, the choice of data or the choice of features.	A developer may expect there to be a relationship between mortality risk and occupational group. Thus a feature in your model is created based on occupational group, but other features are overlooked that are as, or more, appropriate in determining risk.
SELF- SELECTION	Where the subjects that will feature in a data set select themselves (a type of sampling bias).	Insurers using a health app or wearables in order to determine appropriate premiums may result in only individuals who are fit and healthy, affluent, and who understand the technology; this may result in a data set that omits those who are most vulnerable in society and result in poorer risk classification for those lives.
SOCIAL DESIRABILITY BIAS	Data subjects will overreport 'good' behaviour and underreport 'bad' behaviour.	Where a health tracking app relies on self-reporting rather than directly tracking activity, a desire to look good may cause reporting distortions (e.g., rounding up the amount of time spent exercising or underreporting alcohol and fast food consumption). This may be compounded by the potential financial benefit from appearing healthy.

As insurers increasingly start looking to use ML to harness alternative, nontraditional data sources (e.g., wearables, social media data or data provided in health apps) they should take note that despite the wealth of data out there that creates the impression we are moving towards full, complete data sets that are free of bias (particularly sampling bias)—this is not necessarily the case.

Further, data from such sources are retrieved from individuals who have the necessary resources, skills and interest to use certain digital devices, to access certain websites and use certain apps and platforms. This leads to self-selection bias, where certain lives cannot access a service, and the more exclusive a platform or a technology is, the higher chance it will result in self-selection bias. Furthermore, certain individuals choose to opt out of sharing data due to privacy concerns, meaning the data may systematically exclude certain groups for which these concerns are a characteristic (Richterich, 2018). Although such individuals may not meet a certain characteristic protected by law, insurers would want to consider whether their AI systems sufficiently allow access to such groups.

MEASURES OF FAIRNESS

In order to achieve fairness, one must first be able to define it. This is a challenge, given there is no universal definition of fairness (Saxena, 2019). It is possible to determine many metrics on fairness[37] in a ML system, and at present there does not seem to be a consensus on which is best. Indeed, there is often:

- **A cultural component to fairness:** What may be deemed fair to one group in society may not be deemed fair to another (see (Bolton, Keh, & Alba, 2018) and (Kipling & Iyer, 2008)).
- **An industry or application specific context to fairness:** For example, there are rules around fairness in relation to specific insurance practices (see (FCA, 2019) for fairness in relation to pricing in the U.K. financial service industry) or for specific industries (see (FCA, 2015) for guidelines on the fair treatment of customers in the U.K. financial service industry).

Likewise, what people deem as fair in one setting (e.g., criminal justice) may be vastly different to what they feel is fair in another (e.g., employment).

One definition of fairness that is relevant to the decisions made by AI (particularly in life insurance use cases) is to link it directly to discrimination; an AI system may be judged to be fair if it can evidence a lack of prejudice towards, or favouritism to, a particular individual or group of individuals based on their characteristics (both inherent and acquired) in all cases where such prejudices or favouritism cannot be legally justified. Then, an AI system is unfair if decisions are skewed in favour or against particular individuals or groups of people without reasonable justification.

This concept of discrimination is set out in human rights legislation such as the EU's Charter of Fundamental Rights, which broadly prohibits discrimination on the grounds of human characteristics, 'sex, race, colour, ethnic or social origin, genetic features, language, religion or belief, political or any other opinion, membership of a national minority, property, birth, disability, age or sexual orientation' (European Commission, 2020c). It is also set out in various national and international laws that make it illegal to discriminate on the grounds of protected characteristics.

WHAT TYPES OF DISCRIMINATION SHOULD DEVELOPERS BE AWARE OF?

- **Direct discrimination:** When protected attributes of an individual results in non-favourable outcomes towards them. For example, offering higher premiums to males than females.
- **Indirect discrimination:** When individuals look to be treated based on neutral and non-protected attributes, but protected groups of individuals still get treated unjustly through identification of their protected attributes by 'proxy.' For example, offering higher premiums to individuals who work in construction and who watch war movies, perhaps as a proxy for males.
- **Systematic discrimination:** Patterns of behaviour, policies or practices that are part of the structures of an organisation, and that create or perpetuate disadvantage for certain subgroups of a population. Historically, the practice of redlining was justified as it excluded 'risky' neighbourhoods from insurance. The result was that Black families and businesses struggled to buy or maintain their properties, creating a feedback cycle which generated broader and further inequality (Squires & Velez, 1988).

In life insurance, this could involve ensuring that decisions made about policyholders, or potential policyholders, by an AI system are based on relevant features only, and not features that the individual

[37] IBM's AI Fairness 360 (IBM, 2020), an open-source tool kit which can assess fairness in ML applications using over 70 metrics.

cannot influence in order to get a more optimal decision. That is, it is about finding the ‘explainable’ and justifiable differentiation that is permitted under antidiscrimination legislation and excluding from your data and AI modelling any differentiation that is not permitted. This is not a new problem for insurers introduced by AI, but one with a long history. For example, in pricing and underwriting it is necessary to differentiate between high and low risk groups in a risk pool, but insurers, regulators and legislators have a history of determining when it is not fair to do so; whilst gender is often an informative risk factor, the ruling on the European Gender Directive makes pricing based on this unlawful.

What is new though is that using AI, particularly in conjunction with big data, allows firms to consider a much broader number of risk factors, and thus widens the scope for ‘unfair’ characteristics to influence decisions. Some of these may be informative in their own right and some informative only in that they correlate significantly with protected characteristics the model is not allowed to lawfully consider.

Insurers must be able to evidence that their models do not discriminate by proxy—that is, the model is not using features to determine its output which are highly correlated to protected characteristics whilst excluding those protected characteristics from the model on the grounds that their use is not permitted (e.g., under antidiscrimination law). This would be classed as indirect discrimination.

However, a difficult challenge for insurers will be in auditing their models to evidence that they have been designed not to discriminate with respect to a protected class or a proxy. For example, to determine if an algorithm has bias with respect to *age*, you would need to, in theory, find an attribute or a set of attributes that act as a proxy for age or that correlate with age and then use techniques to demonstrate the bias (e.g., Statistic Parity Difference; see Table 9). This will, in a way, demonstrate that a proxy for age exists and it is then reasonable to assume that the model used this proxy to differentiate. This then becomes circular in that the model will need to be updated to ensure this differentiation by proxy is removed, with the audit then performed again on the new model.

Many factors may be correlated with protected characteristics, including some which are innately informative, aside from their link with the protected characteristic. For example, occupation and gender may be correlated, but occupation may also give useful information on an individual’s health.

This presents a challenge to insurers; insurers need to differentiate—they will not set the same premium for all lives, come to the same underwriting decision for all lives or approve all claims received without review. Thus, insurers instead have to determine that they are comfortable with how the model is differentiating (and to document this), rather than whether it is differentiating. The goal for insurers must therefore be to put in place mechanisms to measure, document, minimise and mitigate any potential bias that may put certain groups in a systematic disadvantage (or advantage).

DISCRIMINATION IN INSURANCE—WHEN IS IT ALLOWED?

This will vary from country to country. In the U.K., insurers can differentiate on certain *explainable* attributes where their reasoning is based on statistical evidence, but cannot when this is unjustified or unexplainable. What is, and is not, allowed, is set out in antidiscrimination legislation, including:

- **Equality Act 2010:** U.K. law that prohibits discrimination based on nine protected characteristics (age, disability, gender reassignment, marriage and civil partnership, pregnancy and maternity, race, religion or belief, sex, sexual orientation). The main change brought in by this legislation was to make it unlawful to discriminate based on age. One can decline or impose certain restrictions on disabled lives, provided the decision is based upon information or data that is relevant to the assessment of risk, and statistical evidence supports this.
- **European Gender Directive:** The European Courts of Justice ruled against Insurance opt-out provision in the European Gender Directive, meaning that from 2012 onwards EU insurers cannot use gender as a rating factor in their models.

It should be noted that direct discrimination via differential treatments of the grounds of age, is not discrimination under the Equality Act 2010, provided it is shown to be a ‘proportionate means of achieving a legitimate aim.’ Thus, price can vary by age, or lives can be banded into groups based on age, where this genuinely reflects risks or costs.

Indirect discrimination is not discriminatory provided it can be justified as a proportionate means of achieving a legitimate aim.

Insurers face an additional challenge in certain processes like pricing and underwriting. The extent to which they are allowed to differentiate based on protected characteristics will differ from one country to the next. Given this, insurers using ML in these contexts will also need to ensure that any rating factors or proxy rating factors are allowable under local regulation.

However, achieving fairness by focusing on discrimination is, in practice, incredibly difficult. One way of considering issues around fairness involves splitting it into two categories: what is fair for an individual (individual fairness) and what is fair for a group of individuals (group fairness).[38] For a given application of AI, it is important to consider both individual and group perspectives, noting that these are often considered to be competing. With these fairness targets, firms can look to define metrics that can be used to test and monitor fairness (see Table 9).

To achieve individual fairness in an AI system, it should be possible to show that two similar individuals experience similar outcomes from the application; that is, there is no *disparate treatment,* whereby there is unequal treatment of similarly situated individuals due to their protected characteristics. That way, independent of the outcome of those two individuals, the system has been fair in its decision making for those individuals. In an automated AI pricing or underwriting application, this would require consistent treatment of risk, whereby two similar risks would be provided with the same risk rating, and thus would have the same outcome. Group fairness on the other hand requires what is often referred to as statistical parity[39] of the outcomes between different groups. That is, that there is no *disparate impact* arising from the system, such that it adversely affects one group of people of a protected characteristic more than another.

TABLE 9: SOME FAIRNESS MEASURES FOR ML MODELS

FAIRNESS MEASURE	GROUP/ INDIVIDUAL MEASURES	DESCRIPTION
STATISTIC PARITY DIFFERENCE	GROUP	The rate of getting a good outcome for the protected group minus the 'good outcome' rate of the unprotected group. Seek to minimise.
DISPARATE IMPACT	GROUP	The rate of getting the good outcome for the protected group divided by the good outcome rate of the privileged group. Seek to maximise.
EQUAL OPPORTUNITY DIFFERENCE	GROUP	The difference between the rate of true positives[40] between a protected and unprotected group. Seek these to be equal.
AVERAGE ODDS (EQUALISED ODDS) DIFFERENCE	GROUP	The difference between the rate of true positives between two groups plus the difference between the rate of false positives[41] between two groups. Seek these to be equal.
COUNTERFACTUAL EXAMPLES	INDIVIDUAL	Review whether you get a different result for a test case when changing only a protected characteristic in isolation.

There are inherent difficulties with achieving both types of fairness at the same time. For example, imagine a life insurance pricing model that makes random predictions for the premiums to set for members of a group of similar individuals. The randomness would mean that under a group fairness metric the model is fair, but it would clearly lead to unfair premiums at an individual level.

On the other hand, assessing the *similarity* of individuals in order to assess individual fairness has its own complications. It is instinctive to say that in the pricing of life insurance an individual with the name John should be treated the same as an identical an individual named Mohammed (The Sun, 2018). However, in practice, the individual's name is not the only characteristic that will influence the pricing decision. Thus, in order to accurately check whether individual fairness has been achieved, it is necessary to determine an appropriate distance measure to compare the similarity of two individuals across all attributes that contributed to the final decision. This is not a straightforward task.

[38] There is also a consideration of 'sub-group' fairness, where you aim to obtain some of the best qualities of group and individual fairness (e.g., take a group fairness constraint and ask whether this holds over a large collection of sub-groups).

[39] The outcome is independent of the protected attribute.

[40] In validating a binary classification ML model, a true positive is an instance where the positive outcome is correctly assigned to an individual test data (i.e., expected a non-lapse, was a non-lapse).

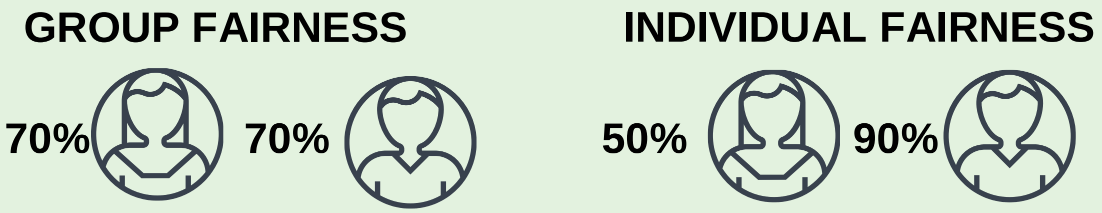
[41] In validating a binary classification ML model, a false positive is an instance where the positive outcome is incorrectly assigned to an individual test data (i.e., expected a non-lapse, was a lapse).

Insurers may be more comfortable with individual fairness, that otherwise identical candidates should be treated equally and aside from protected characteristics, underwriting and pricing should reflect an individual's risk. However, they must consider how this approach can perpetuate existing inequality.

INDIVIDUAL VS. GROUP FAIRNESS

Suppose individual A and individual B are both seeking the same amount of life insurance cover and have the same characteristics in their application except one: A is male, B is female.

If it is found that the AI application grants life cover to a larger percentage of men than women, the insurer may adapt its techniques to prevent this effect as it is unfair to women. This would be an example of the insurer aiming to achieve group fairness in its underwriting process, i.e., fairness between men and women. However, this may lead to stricter criteria for men than women. Despite their similarities, individual A could be declined life cover, while individual B was accepted. This would be considered unfair at an individual level.



WHAT CAN BE DONE?

Research is ongoing to determine technical methods that aim to address issues relating to bias in AI systems; various AI domains (e.g., deep learning, NLP) all have their own domain-specific issues to address and so the techniques and tools that are appropriate differ by domain—see (Meharabi, Morstatter, Saxena, Lerman, & Galstyan, 2019) for a reference to various research on methods to ensure fairness in ML applications. The methods can largely be broken down into:

- **Pre-processing:** Tools and techniques that aim to transform the data so that any bias or discrimination can be removed before the model is trained. A goal is to remove the undesired impact of protected characteristics on the outputs of the AI system, while retaining as much other information as possible (Chiappa & Gillam, 2018).
- **In-processing:** Tools and techniques to tackle bias during model training; for example by imposing constraints on the model's optimisation process[42] or using a technique called adversarial debiasing[43]—see (Zhang, Lemoine, & Mitchell, 2018) for a good resource on adversarial debiasing.
- **Post-processing:** Removing bias by modifying the model output so that it satisfies specific fairness definitions. The general methodology of post-processing algorithms is to take a subset of samples and change their predicted labels appropriately to meet a group fairness requirement (Lohia, et al., 2019).

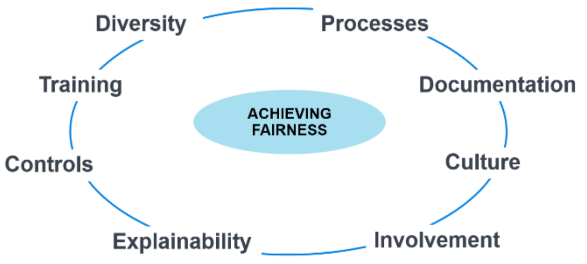
[42] The optimisation process involves changing overarching parameters of the model with the aim of improving the predictions it makes.

[43] You build two models. The first is predicting your target based on your current feature engineering and preprocessing on the training data. The second is the adversary and tries to predict, based on the predictions of your first model, the protected attribute. If there is no bias, the adversarial model should not be able to predict the protected attribute. The second model then can guide change to the original model that weakens the predictive power of the second model, until it cannot predict the protected attributes well.

Similar to the efforts in relation to Explainable AI (XAI) covered in Section 3.8, many technology firms have started to develop tools to help developers investigate a range of fairness measures in their ML models including Aequitas (Aequitas, 2020), TensorBoard's What-If-Tool (TensorFlow, 2020) and IBM's AI Fairness 360 (IBM, 2020). For example, the latter is an open-source tool kit which can assess more than 70 fairness metrics and can be used in common programming languages such as R and Python. These open-source solutions will support developers in their identification, documentation and communication (to relevant parties) of bias and fairness issues inherent in an AI system.

Only once fairness issues are understood can they be mitigated, and technical solutions in preventing bias only will be effective if teams are well placed to know what to look at when making decisions on these models, and those operating them continue to question.

FIGURE 13: NONTECHNICAL FAIRNESS CONSIDERATIONS



Consequently, firms should consider the following nontechnical factors that contribute to ensuring fairness in AI system applications:

- **Diversity:** Diverse teams (with regards to ages, gender, ethnicities, educational background, cultural perspectives, to name a few) help represent a wider variation of life experiences. Firms benefit from embedding diversity in the teams that are interacting with AI systems at various stages in the life cycle of an AI system as it provides greater scope for identification and consideration of actual and potential issues including fairness, bias and discrimination. However, achieving this diversity in the workforce is likely to take time and effort on the part of insurers— (Gardam, 2019) pointed out that the bias in AI systems may reflect those of the people who create such systems, who are not diverse. For example:
  - 17% of U.K. IT specialists are female, 8% are disabled (compared with 23% of the working population), and non-Caucasian IT employees are more than twice as likely to be in part-time work when compared with their Caucasian counterparts (Gardam, 2019).
  - In 2017, 20% of Google's technical employees were women, 1% were Black, and 3% were Hispanic. Facebook had similar diversity statistics in technical roles, and Twitter was 15%, 2% and 4%, respectively (Gardam, 2019).
  - The proportion of women graduating with a degree in computer science at American universities fell from 37% in 1984 to 18% in 2010 (Gardam, 2019). A similar rate (around 20% to 24%) applies to qualifiers from U.K. universities based on data from 2015 to 2017 (HESA, 2018).
  - On average 71% of the applicant pool for AI jobs in the U.S. are men (Shoham, et al., 2018).

  Diversity is beneficial in the teams that are responsible for making decisions about the AI system as well as those who influence decision making (i.e., not just developers), including teams making decisions on the data to use in an AI system (and that review it); teams that choose the optimal features for models; business users who are using and interpreting the outputs of AI applications; those responsible for providing sign-off for the use of a system; and those responsible for an AI system's ongoing review.

- **Processes and procedures:** Insurers need to review and ensure there are adequate processes and procedures for the consideration of bias, fairness and discrimination at various stages of the life cycle of the AI systems. They must ensure that sufficient time and effort is put into reviewing these issues and that they have individuals, with sufficient expertise and training, responsible for determining where bias and fairness issues may arise in the data, the model or its outcomes (so it can be tested), and wider issues around the fairness of the use of the technology.

The mechanisms put in place need to allow for regular review, in particular for ongoing learning models, to reduce the risk that a model determined to be free from bias and fair at the onset will have such issues introduced at a later stage as it learns from new data sets. Firms should consider the level, and regularity, of human intervention in interpreting the results from models, in order to spot any potential or actual issues pertaining to bias, fairness and discrimination.

For consumer-facing applications (e.g., pricing, underwriting, claims, marketing, customer servicing) insurers should ensure that mechanisms are in place that allow consumers to alert the company to any issues that they perceive in an AI system, with appropriate channels for such issues to be to be escalated, investigated and rectified.

- **Documentation is key:** To ensure accountability in cases where there has been discrimination, or unfair outcomes for consumers, firms must ensure that the decisions made about AI systems are clearly documented and retained.[44] In particular, insurers would need to record what it has, and has not, tested and considered in the AI system and any bias it is aware of inherent in the data, or features and results identified dynamically, throughout the AI system's life cycle.
- **Culture:** Insurers need to create a culture that promotes the consideration of bias, fairness and discrimination in its AI use cases. The culture needs to encourage and make it easy to highlight, escalate and investigate any potential, or identified, sources of bias, unfair outcomes or discrimination. Such a culture of speaking up will allow continuous learning that should act to help prevent similar issues in the future. This culture can be created through internal policies and procedures that those who develop, deploy or manage AI systems are aware of and agree to, by the monitoring and escalation mechanisms put in place, and through teams being encouraged to identify issues that they perceive could result in bias, issues of fairness or discrimination.
- **Involve multiple stakeholders and the public:** Determining what is, and is not, fair and what to look out for when developing and deploying AI systems requires getting as broad a range of views on what is fair as possible. Insurers could consider ensuring a range of inputs from internal stakeholder groups in in the planning and design stages for the AI system and engaging the public (see Section 3.10) to gain this viewpoint. This would facilitate developers identifying issues they had not originally considered, and to factor these viewpoints into the development of the AI system.

  Firms could also consider the use of external parties to perform an external audit of the system with a focus on bias, fairness and discrimination, to provide an independent viewpoint, benchmarked on best practice and the latest guidance and regulation on AI system development, that takes into consideration the learned experience from prior review work. For dynamic systems, such a review could take place both before and after deployment.
- **Explainability:** It is easier to discern the inner workings of more transparent and explainable models (and thus to spot and test for issues around bias, fairness and discrimination), and harder as they become more complex and black box. Generally, without being able to test for these issues, you cannot contest them. The trade-off between explainability and fairness considerations should be factored into the decision making prior to development as insurers consider what models will be best for their application.
- **Training:** Insurers have a responsibility for ensuring that those who are involved in the developing, reviewing or ongoing management of AI systems, and those deciding to introduce them for a business use case, have been trained on issues of bias, fairness and discrimination. Then, they will know what to look out for, but they should also look forward to ongoing research developments, and new tools and techniques, that could be incorporated in the AI system.

[44] This will include insurers determining, and documenting, their perceived definition of fairness for a particular AI use case. Those responsible for managing AI systems must ensure that controls and metrics are in place that reflect this definition and it is communicated to relevant parties that are responsible for business decisions and strategy—ultimately the board.

- **Controls embedded in the AI system:** In addition to human oversight, testing and review of data, models and model output, insurers could also consider whether controls could be embedded in their AI systems; for example, it has been proposed that models could learn when not to make a prediction and to pass a decision to a human if it knows a decision for a particular group does not have sufficient credibility (Madras, Pitassi, & Zemel, 2018).[45] Insurers could embed such controls into its sensitive consumer-facing applications (e.g., pricing, underwriting, claims, marketing, customer servicing).

OTHER CONSIDERATIONS

Fairness considerations can extend beyond whether the model avoids discrimination and ensures fairness for individuals and groups with protected characteristics; there are also questions around the model's very introduction and its methods for achieving its goals that those responsible for the introduction of AI systems in a business need to consider.

- **Hyper-personalised underwriting and pricing:** There are also fairness considerations in underwriting and pricing—a sufficient degree of risk classification ensures that policyholders are charged a fair amount for the financial services that they are purchasing. It aims to prevent policyholders that represent higher risks from profiteering from the use of heterogeneous groupings, and the use of AI in pricing may be seen as a method to ensure this. Conversely, greater risk classification can be seen at odds with the concept of pooling risks that is central to insurance; taken to its furthest conclusion, if AI systems can increasingly predict the mortality[46] of potential policyholders, there would increasingly become a section of society deemed too unhealthy to be accepted for life insurance at a reasonable price—an *uninsurable* section of society.

  This is at odds with the purpose of insurance as a social good, which may cause political opposition that may influence regulation on the use of AI systems in this area.

  This risk was raised as a concern by the U.K. financial services conduct regulator in a 2016 feedback statement on the use of big data in insurance, albeit it focussed on general insurance (FCA, 2016).

- **Unfair pricing practices:** The use of AI could increase the risk of unfair pricing practices where insurers could use it to learn the propensity for an individual to accept a premium at a higher price than is fair for their risk classification in order to optimise (for the insurer) the premiums that they will charge to potential policyholders (or existing policyholders in the case of renewable products).

  Categorising policyholders by their propensity to accept a premium is already de facto carried out in the U.K. insurance industry whereby insurers offer discounts for new customers and higher premiums to existing customers on renewals in general insurance. This practise has been deemed by the FCA as unfair and will soon be made illegal in the U.K. (Sillars, 2020). Using AI to analyse big data (e.g., retail loyalty card schemes or social media profiles) could provide insurers with the ability to derive lifestyle and behavioural factors that could potentially allow competitive deals to be offered to those likely to shop around and suggest policies with larger profit margins to customers who are less financially motivated (BEUC, 2020). In both cases, this would not be fair.

- **Vulnerable and less privileged groups:** The FCA (FCA, 2019) presented that insurers may need to place greater concern with regards to fairness problems in pricing for groups that are considered vulnerable (e.g., the old, the poor) than those that are wealthier and more financially stable.

  The introduction of AI into life insurance needs to consider its impact on vulnerable and less privileged groups; such groups may be less capable of understanding how algorithms and new data sources could influence decisions that are made about them, such that they cannot mitigate any adverse effects or seek appropriate redress when they have been victims of unfair practices.

[45] That said, this could result in unfair decisions by way of protected groups receiving more scrutiny than unprotected groups in their underwriting, pricing or claims decisions.

[46] For example, by identifying more precise and previously unseen indicators of risk for the first time (e.g., a previously undiscovered link between someone's occupational grouping and the chance of falling ill at work).

Products or AI systems that are built around gathering data from the use of apps or smartphones requires higher levels of digital and health literacy.[47] Those who can download a health app, who use social media or who utilise wearable technology are more likely to be younger, healthier and more affluent. This excludes those who may be in greatest need for insurance, and insurers need to deem if this is fair.

The use of social media data sets in AI systems for underwriting and pricing could also result in unfairness for certain groups. Would certain groups know to take down pictures from social media of them carrying out activities that the insurer may deem as 'high risk'? If you are healthy, do you know how to effectively signal that you are healthy on social media by posting information on your gym routines? If you do not know the appropriate manner in which to present yourself on social media for insurance purposes, then you would not be able to benefit from lower premiums or more favourable underwriting decisions. This isn't just about presentation, though, as the most vulnerable groups may not be able to afford the activities that would allow for favourable decisions by contributing to the data sets that drive the AI system's decision making (e.g., having a smartphone to use a health app or owning a wearable device).

This potential for unfairness is not an inevitable aspect of progress—AI could create greater fairness by increasing accessibility to life insurance products to certain groups. These groups may have found insurance unaffordable under traditional pricing methods but could be offered a lower, more accurate premium by insurers using granular pricing methods that apply AI to broader sources of data (e.g., using wearables data to offer lower premiums to active people who are age 50 or older).

AI may also serve to increase the availability of insurance for the most vulnerable groups in society; microinsurance has adopted AI to reduce the costs of reaching customers (e.g., via chatbots) and to utilise broader sources of data on customers with limited financial history (e.g., mobile phone payment records) (insight2impact, 2018).

Thus, insurers need to consider whether the use of AI is tackling the inequalities in access to insurance and the digital world for those that are most vulnerable or less privileged, or making these worse.

These considerations go beyond the scope of those who are developing the systems, and require wider ethical matters to be discussed with and approved by those responsible for the strategy of the business (i.e., the management and board), informed by relevant stakeholders with sufficient expertise to comment, including the firm's ethics boards (or other responsible governance body), regulatory bodies and technical experts.

3.7 KEY ETHICAL CONSIDERATION: ACCOUNTABILITY

If a human makes an error in decision making that adversely impacts a policyholder or that puts the financial security of a firm at risk, clearly defined roles and responsibility make it easier to ascertain culpability and ensure that the person can be held accountable for their actions. For individual decisions, recommendations or predictions that are made by an AI system, often without explicit human sign-off (i.e., automated) there is concern that it is not possible to determine who is accountable when the AI system results in adverse conclusions and harm. It is harder to determine the primary cause of a problem; was it poor design, a lack of understanding from those overseeing the AI system, or a lack of oversight and risk management from the board of a firm?

Firms need to consider not just issues of accountability regarding the final decisions, recommendations or output from an AI system but also any intermediate decisions made by an AI system that contribute to a final output. For example, consider an AI system that is making a recommendation for an individual. Suppose the system decides, based on the individual's profile, that the person is highly likely to be divorced within 10 years, and bases all recommendations or decisions on that conclusion. If the individual challenges those recommendations or decisions, is it reasonable for the justification to include that the AI believes that they will be a divorcee soon?

[47] Health literacy is the cognitive and social skills which determine the motivation and ability of individuals to gain access to, understand and use information in ways which promote and maintain good health (WHO, 2020).

Human judgement is required throughout the AI system life cycle; from determining whether an AI system should be developed and how, determining what data and models to use and how to decide whether the model is working well, to determining who in the firm has the required competence to review and sign off where and how AI will be used in the firm. Consequently, assigning accountability for decisions made by a seemingly objective AI system is, and must be, possible and will be facilitated by robust documentation and governance.

Also, AI can change the nature of the risk: for example, a normal car is not sold on the basis that it will not permit you to run over someone, but, not unreasonably, autonomous cars are being held to that standard. Automated underwriting may permit the rapid issue of policies through the exclusion of certain coverages to customers with particular health issues, but customers may not fully understand the implications or seek the explanation an insurance agent would provide.

Accountability is best considered by focusing on who may need to be held accountable for an AI system:

- **People:** Robust documentation and record-keeping of those who have made decisions that impact an AI system, its review or use in the firm
- **Firm:** Responsible for the introduction of an AI system and thus the ongoing governance and oversight of the system

ROBUST CONTROLS AND DOCUMENTATION

Insurers will need to keep internal documentation of the design process of an AI system, including who was responsible for making and signing off decisions regarding the choice of testing and training data sets, the type of model and parameters used, the level of testing performed and how it was determined that the model was fit-for-purpose. In particular, any trade-offs between various ethical principles should be documented, including who was responsible for the signing off the decision to proceed in light of any trade-offs. They must also keep a record of the exact data sets and model configuration used in testing in the event that these need to be called upon at a later stage (e.g., to assign responsibility to designers).

In the event that, prior to deployment, a firm chooses to get an external audit of the AI system in order to independently verify that it is fit-for-purpose (technical audit) or that there are sufficient governance and controls in place around the system, it must be clearly defined what was checked (i.e., including any limitations to the audit) and the extent to which, and circumstances in which, reliance can be place on the review.

Once AI systems are in deployment, the data input into the system and configuration of the AI system at any point in time needs to be recorded and documented to facilitate investigation in the event of errors or harm to consumers (or other adverse outcomes). Firms must keep clear documentation of the ongoing checks and controls that they are performing and how these contribute to their assurance in the competence of the system. In the event of changes to the AI system, dynamic record-keeping of the key decisions is required, including who made the change (including their authority to make changes), why a change was made, and what additional checks and controls have been performed in order to ensure the model continues to be fit-for-purpose. This may include creating record-keeping systems that make it possible to always trace who is legally responsible for a particular AI system, and for which components they are responsible.

GOVERNANCE FRAMEWORKS

‘We want to see boards asking themselves, “What is the worst thing that can go wrong” and providing mitigations against those risks.’ —Christopher Woolard, FCA

At its core, accountability relates to the ability to assign responsibility to individuals and firms involved in the design, deployment or ongoing monitoring of AI systems so that they can be held accountable for the proper functioning of the AI system, and also accountable when things go wrong. There should be no loss of accountability simply because a decision, recommendation or prediction was made by, or with the help of, an AI system, rather than solely by a human.

Accountability is well summarised in (ICO, 2020) that suggests firms must be:

- Responsible for the compliance of their AI systems
- Able to assess and mitigate any risks introduced by the AI system
- Document and show how the system is compliant, justifying any decisions made

For outcomes determined by an AI system there may be less of an opportunity for humans to call out or ‘whistle-blow’ improper practice than for decisions made by a human;[48] consequently, greater responsibility will lie with those responsible for the technical management of the models (i.e., those who are signing them off as fit-for-purpose), and the board of the firm that have made the decision to implement the system.

Thus, boards of firms need to ensure that they are establishing both clear lines of accountability for those people responsible for AI (and ensuring these are clearly documented and visible to concerned parties) but also put in place structures that provide sufficient control and oversight for AI systems due to their own responsibilities and accountability.

Some considerations for insurers looking to create robust structures for accountability include:

- A clearly defined, and regularly reviewed and managed, list of those who are responsible for an AI system including where their individual responsibilities begin and end. Those people need to be aware of their responsibilities and evidence that they have agreed to such responsibilities. This needs to be done both prior to deployment (i.e., ensuring developers are aware of their responsibilities) and post-deployment.
- Assigning roles in the firm for the day-to-day oversight and control of AI systems, who will be responsible for wider oversight (e.g., model quality assurance, risk management), and where those responsibilities start and end.
- Establishing policies and procedures for the development of AI and ensuring that designers and developers of AI (both internal and external) are aware of their responsibilities that they have agreed to in line with these documents, and in line with their job description.
- Structures within the firm that could incentivise human oversight and accountability in light of the introduction of AI systems within the firm. The Solvency II insurance framework addresses internal governance mechanisms that support strong risk management; insurers will need to consider how its internal oversight and control functions may need to be changed, in order to assign responsibility for the risks introduced by AI systems to an oversight body in the firm.
- Consideration of how current financial regulation can enhance accountability. For example in the U.K. there is an explicit regulatory requirement for insurers to assign certain roles and responsibilities under the Senior Managers and Certification Regime (SM&CR). This regime is intended to enhance accountability within firms by making those holding roles of responsibility more accountable for their conduct and competence. The SM&CR involves documenting and notifying financial regulators of who is responsible for what in the firm, and incorporates key C-suite roles and other roles more specific to risk management, and actuarial work (e.g., the chief risk officer and chief actuary). Firms would be required to assign responsibility for the risks introduced by AI, and this would allow full, transparent accountability at a firm-wide level.

Some considerations that the board of a firm are responsible for:

- What are the risks to the firm arising from the use of AI?
- Who on the board is ultimately responsible for oversight? Which committees should be considering which AI related issues?
- Should an AI ethics committee be created? If so, what issues does it address? What is its verdict on acceptable trade-offs (e.g., between accuracy and explainability)?
- Do our current risk assessment systems include AI specific risks? Is the firm’s AI risk management proportionate to the firm’s use of AI?
- How effective are the current compliance and audit functions when it comes to AI systems?
- Do staff, contractors and third parties have adequate training on ethics and AI?
- Does the board itself require additional training to properly assess the firm’s AI usage?
- Do reports to shareholders and regulators adequately cover AI use?

[48] There will still be some opportunity for whistleblowing; if the firm is monitoring the performance of the system, then someone may identify an issue. For example, an overseer may identify unfair outcomes for customers identifying themselves as from a specific background or region. If no change is made to resolve the issue, and there is reason to believe this may cause harm, an overseer could notify relevant parties both of the system’s failings and those responsible for the system who have failed to rectify any identified failings.

- Consideration of whether existing governance frameworks are sufficient to ensure efficient oversight and responsibility for the use of AI systems (e.g., by assigning responsibility to an existing committee, panel or group) or whether a separate group is required. Firms should also consider whether a single person, or a group of people (e.g., the panel), has responsibility and should be held accountable for oversight. Such individuals or groups will have the responsibility, and thus accountability, for ensuring relevant controls, safeguards and oversight mechanisms are in place to monitor the use of AI and ensure the systems continue to be fit-for-purpose.
- Insurers are part of a highly regulated industry and so already have robust governance structures in place for managing risk. However, insurers should look to include the use of AI into any existing risk management frameworks, in particular those introduced under Solvency II, to ensure there is sufficient systems and controls in place for monitoring, mitigating and controlling the risks posed by the use of AI.

As discussed, when looking to implement controls, insurers will naturally look to their existing frameworks to manage data and model risk, as the controls developed there should equally apply to AI models. However, there are particular risks arising from AI which must be considered in addition. Consequently, firms are responsible for creating specific controls for these that may be built into existing frameworks or developed from scratch (see Table 10).

Firms should remain aware of its obligations where part of an AI project has been outsourced, where third-party data is used in an AI system, or when off-the-shelf AI systems have been purchased where it is more challenging to perform due diligence. In such cases, it will be important to seek documented legal opinion on areas where it is not clear where the responsibilities of the insurer, and those of the third party, start and end. This legal advice should be sought prior to the deployment of an AI system.

TABLE 10: AI INTRODUCES NEW RISKS AND NEEDS NEW CONTROLS…

AI FEATURE	POTENTIAL CONTROL MEASURES
Automated processes require monitoring, in particular due to the risk of *model drift*, where the model continues to learn after development and model predictions seem to degrade because of changes in the underlying distribution of the feature set (data drift) or changes in the distribution of the dependent variable (concept drift). Ongoing testing of the system (i.e., not just at outset) is required.	▪ Agree on a designated person who has responsibility for each ML model that is in production. ▪ Performance monitoring for the ML model should be on a more frequent basis than for standard models. The firm needs to determine regular dates for the model to be retrained and when the system may be retired from use. ▪ Alert systems to flag unusual/unexpected actions. ▪ Human-In-The-Loop (HITL), [49] Human-On-The-Loop (HOTL),[50] or Human-In-Control (HIC)[51] mechanisms. ▪ Guardrails to automatically switch off an ML model if it produces undesirable outputs, with potential fall-back non-AI systems that can instead be used. ▪ Use a second AI system that can be switched to in the event that there is model drift that needs to be investigated.
Limited explainability of the models may prevent detailed auditing and checking.	▪ In-depth model validation exercises prior to deployment and post-deployment. ▪ Documenting explainability and model limitations including justification of why a more complex AI model is appropriate over a simpler (traditional) model. ▪ The firm's model risk criteria should be updated to ensure it is relevant for AI models used. ▪ Consider XAI approaches discussed in Table 12, investing time and money in developing internal explainability tools, focussed on the highest risk use cases of AI at the firm.
AI technology is emerging—the risks and benefits of AI models, and best practice for validation and governance of such models may not be known.	▪ Ensure that any process documentation created in the development stage includes the rationale for the approach, any risks and mitigations—justify why a new technique is preferred over a more well-established approach. ▪ Feedback, insight, guidance should be gained from those with technical knowledge in the firm and also through external advice from legal experts, ethicists, the public, domain experts, and from academia.
A lack of knowledge in AI in the firm that makes it difficult for them to audit and review AI systems, or AI projects generally, effectively.	▪ Include those with technical ML experience in the model quality assurance, model risk management and data governance functions where possible. ▪ Organise training for the board and management on AI, including its benefits, risks, limitations and ethical concerns they should be aware of. ▪ Create committees to advise the risk management function of the firm on ML specific questions.
Ethical considerations including those unique to AI systems (e.g., risk of bias and fairness issues).	▪ Perform reviews of ML models that focus on identifying potential bias in the data, or algorithms used. ▪ Establish internal frameworks, principles and practices on the use of AI in the firm that are widely communicated and ensure developers know what is expected of them. ▪ Ensure that the firm has an up to date knowledge of discrimination regulation and AI systems are reviewed in light of any changes in regulation. ▪ Establish a committee or function in the firm that are solely responsible for ethical issues in relation to the technology use (or AI specifically). ▪ Create new roles in the firm such as the role of chief ethics officer responsible for alerting the board to any concerns, developing internal frameworks on the ethical use of AI, and ensuring relevant staff have an appropriate level of training on ethics and AI.
Issues arising from the data used in AI models (risk of bias, errors, privacy infringements, unauthorised access).	▪ Perform thorough quality assessment on data used in AI models (for training and testing) to assess its quality, identify any errors and assess potential bias in the data. ▪ Ensure that the firm is aware of the origins of data including that appropriate consent for the intended use case was granted for any personal data. ▪ Ensure robust data security is in place for data used in AI systems, in line with local regulations on the use of personal data (e.g., the General Data Protection Regulation [GDPR] in the U.K.). ▪ Ensure robust cybersecurity controls are in place that are sufficient for the volume and nature of the data that will be used in the AI system.
Allocation of responsibility when using off-the-shelf models or data, open-source software or systems developed by third-party providers.	▪ Apply the same standards to externally sourced models or data as those developed, or sourced, internally. ▪ Ensure due diligence is performed on externally provided software to review the testing and validation performed on the model, and its suitability for the intended use case in the firm (staff will need to have sufficient knowledge and training to do this). ▪ Seek documented legal opinion on areas where it is not clear where the responsibilities of the insurer, versus the third party, start and end.

[49] Intervention in every decision cycle of the system.

[50] Intervention in the design cycle of the system and monitoring the system's operation.

[51] Oversee all activity of the AI system, with discretion on when, how and whether to use a system in any particular situation (including the ability to override a decision made by a system).

3.8 KEY ETHICAL CONSIDERATION: EXPLAINABILITY

Explainability is needed in situations where it is important to know why a predication, decision or recommendation was made by an AI system and not just what was predicted, decided or recommended.

The need for explainability is highly contextual; AI models that are used in lower risk use cases, where errors do not have serious consequences, may not require explanation. However, if there is a significant financial or social impact, or where there is potential for harm (physical or otherwise) to an individual, explainability becomes increasingly relevant.

In life insurance, a range of applications of AI would warrant explainability, guided by the various stakeholder groups that would require a level of understanding and what they may want to understand. Thus, there is a direct link between explainability and transparency requirements, such that explainability facilitates transparency, and is needed to an extent sufficient to support the required level of transparency.

Thus, focusing on the transparency requirements for various stakeholders, Table 11 highlights the different requirements for model explainability relevant to those higher risk life insurance use cases, where the blue shading indicates that an explainability requirement may be relevant to a particular stakeholder group.

TABLE 11: EXPLAINABILITY CONSIDERATIONS FOR VARIOUS LIFE INSURANCE STAKEHOLDER GROUPS

		EXPLAINABILITY CONSIDERATION				
Stakeholder group	**Some high-risk use cases relevant to stakeholder group**	**Key features that contributed to an individual outcome, decision or prediction**	**Key features that generally drive model output**	**High-level information on how the model is set up and works to come to its outcomes, decisions or predictions**	**How the model output would differ if an alternative non-ML approach was used**	**How the model would deal with data it has not seen before**
Public (policyholders, potential policyholders)	▪ Pricing decisions ▪ Underwriting decisions ▪ Claims decisions ▪ Guidance	🟦	🟦			
Internal developers	All high-risk use cases relevant	🟦	🟦	🟦	🟦	🟦
Internal model quality assurance	All high-risk use cases relevant		🟦	🟦	🟦	🟦
Internal risk management	All high-risk use cases relevant		🟦	🟦	🟦	🟦
Internal management	All high-risk use cases relevant			🟦	🟦	🟦
External audit/ certification bodies	Dependent on scope of work		🟦	🟦		🟦
Prudential regulators (e.g., PRA)	▪ Reserve calculations ▪ Daily solvency monitoring ▪ Capital requirement (e.g., SCR) calculations ▪ Investments ▪ Material assumptions used in reserving			🟦	🟦	🟦
Conduct regulators (e.g., FCA)	▪ Pricing decisions ▪ Underwriting decisions ▪ Claims decisions ▪ Guidance	🟦	🟦	🟦	🟦	🟦
Legal bodies	Reactive—following error that requires legal investigation	🟦		🟦		

Each of the various explainability considerations will require different tools and techniques that aim to achieve the required level of explanation for the stakeholder in question.

However, due to the complexity of many ML models,[52] particularly those nonlinear and non-deterministic in nature, it is challenging to understand and explain how the model is working or making its decisions, prediction, or recommendations. This is referred to as AI's black box problem, meaning that the inner workings of the model are not easily understood. Even in situations where the inner workings can be understood, translating this into a format that is easy for the concerned stakeholder to understand is a new challenge. For example, in the random forest example presented in Appendix 1, one could get a broad understanding of how an individual decision tree works, but how to summarise the workings when there are hundreds of trees becomes increasingly difficult.

The black box challenge has been met with the birth of the research field of XAI that broadly sets out to develop methods and techniques to use with AI that facilitate understanding and explanation.

Such techniques can be grouped by their goal, or the types of ML models that they can be used on (see Figure 14). Taking an underwriting decision as an example, global interpretation methods would be determined to understand what the key features are that drive an underwriting decision, whereas local interpretation methods would be needed to understand how an underwriting decision was made for an individual.

FIGURE 14: TYPES OF EXPLAINABILITY METHODS

- **Using simpler models:** Achieve explainability by using less complex ML models that are easier to interpret (e.g., use short decision trees or sparse linear models).
- **Post-hoc interpretation methods:** Applying methods that analyse the model after it has been trained on data—including partial dependence plots (PDP) and Shapley values.
- **Model-specific interpretation tools:** These can only be used for certain type of ML models.
- **Model-agnostic tools:** These can be used on any ML model and are applied after the model has been trained on data. These methods usually work by analysing feature input and output pairs.
- **Local interpretation methods:** These aim to explain an individual prediction, decision, or outcome made by an ML model.
- **Global interpretation methods:** These aim to explain the entire model behaviour.

There are methods that fall somewhere in between local and global.

We present a range of model-agnostic tools that insurers could consider for various explainability problems (the appropriate tool will depend on whether an answer is sought to local or global questions) in Table 12.[53]

Insurers considering the use of AI need to also consider the costs in investing in tools and processes and hiring staff with the required expertise to explain the results of the models, based on the features and models chosen.

Increasingly, open-source software libraries are creating tools to support explainability, such as TensorFlow's What-If Dashboard (available in Python, R and other packages) and other ML explainability plug-ins for coding packages. See (TensorFlow, 2020) and (Lundberg, 2020), respectively. Such tools may help support developers of ML models for certain global explanations, by facilitating their exploration of how data was interpreted by the model, and how changes to input data can result in different outputs.

However, it will remain a challenge to determine how to operationalise the use of explainability tools in the event insurers need to provide decision explanations to policyholders, or potential policyholders, in a quick time frame (e.g., online at the point of the decision). Instead, such explanation may need to be on request only to allow sufficient time to review an individual case, which may not meet the needs of the users interacting with the AI model.

[52] Noting that some are more complex than others (i.e., linear regression is a less complex model than a neural network).

[53] For a more technical explanation of some of the techniques, see (Molnar, 2019), which includes reference to packages in R that support the examples provided.

It should be noted that XAI is an emerging field of research and as yet does not have all the answers to all of the questions on explainability. Consequently, insurers need to consider, on a case-by-case basis, whether simpler, more-transparent models (that may be less accurate) are a better solution in certain use cases than highly complex ML models.

TABLE 12: SOME EXPLAINABILITY TECHNIQUES AND THEIR ADVANTAGES AND DISADVANTAGES

TECHNIQUE	GLOBAL/ LOCAL	DESCRIPTION	ADVANTAGES	DISADVANTAGES
Partial Dependence Plot (PDP)	▪ Global	▪ Shows marginal effect of one or two features[54] on a predicted outcome, decision or recommendation ▪ Can show if a relationship is linear, monotonic[55] or more complex ▪ Gives a statement about the global relationship of a feature with the predicted outcome	▪ Easy to understand and interpret	▪ Limited to one or two features of interest and so cannot explain more complex interrelationships ▪ Assumes features for which the PDP is considering are not correlated with other features, which may not be true
Individual Condition Expectation (ICE)	▪ Local	▪ Plot for each individual set of input data that shows how a predicted outcome, decision or recommendation would change as a particular feature changes value ▪ Similar to a PDP plot for an individual case ▪ Extensions to look at the change in an ICE curve with respect to a feature value (i.e., derivative ICE plot) that tell you whether changes occur and in what direction they occur	▪ Easy to understand and interpret ▪ Can use to understand interactions between input variables and how they may affect the predictions made by the model	▪ Only can look at one feature at a time ▪ If including multiple ICE plots in one chart it may be challenging to understand ▪ Derivative ICE plots are time consuming ▪ Assumes feature(s) for which the ICE is considering are not correlated with other features, which may not be true
Accumulated Local Effects (ALE)	▪ Global	▪ Similar goal to PDP plots ▪ Divide features into many intervals and calculate the differences in the predictions for features in that interval, against the average, for various regions	▪ Better than PDP plots in cases where predictive features are highly correlated	▪ Consideration required on intervals chosen—too few and complexity of the model is smoothed ▪ More complex to implement than PDPs
Global surrogate/ Reverse Engineering	▪ Global	▪ Interpretable model (e.g., linear regression, small decision trees) that is trained to approximate the predictions of a black box model ▪ Conclusions on the black box model can be interpreted from conclusions of the global surrogate	▪ Flexibility on the choice of the global surrogate model (provided it is interpretable) ▪ Can determined how well the surrogate model approximates the black box predictions (e.g., using R^2 measure)	▪ Surrogate model may not provide good approximations in all cases
Local surrogate (e.g., Local Interpretable Model-Agnostic Explanation (LIME))	▪ Local	▪ Interpretable model used to explain individual predictions of black box ML models ▪ Explanation is provided by the local surrogate model, which has been trained on the outputs of the less interpretable black box model	▪ Can determined how well the surrogate model determines the black box predictions	▪ Explanations may be unstable—two very close sets of input data could vary significantly
Shapley values	▪ Global	▪ Identifying and measuring the size and significance of the contribution each feature makes to a model to determine which are key to decision making ▪ Calculate the Shapley value[56] of individual features ▪ Use the Shapley values of features in a standard regression analysis to rewrite the ML model in terms of its Shapley values	▪ Allows for statistical testing of the model	▪ Intense computing requirement ▪ Explanations tend to look at all features, individuals may only be interested in specific features ▪ Doesn't work well when features are correlated
Permuted Feature Importance	▪ Global	▪ Provides the relative importance of features by estimating how the predictive power of the model changes when a feature is excluded (i.e., breaking the relationship between the feature and the true outcome) ▪ A feature is important if removing it increases the model error, or model output variance	▪ Easy to interpret and explain ▪ Does not require the model to be retrained	▪ Does not capture the interaction or correlation between features ▪ Hard to know whether to apply to training data or to validation data in the model

[54] A feature is a variable that is in your original data set or one created from the data as it may be of relevance to discerning relationships in the data. As a simple example, in an ML model to predict lapses from a policyholder data set, you may have the fields 'issue_date' (when the policy started) and 'cancellation_date' (when the policy ended), but you could create a new field 'duration_to_lapse' that is the calculated duration in-force immediately before the policy lapses. This composite feature will have a greater predictive ability for lapses than that of its parts.

[55] Entirely nonincreasing or nondecreasing.

[56] A concept from game theory—this provides the marginal contribution of the feature to the ML output.

3.9 KEY ETHICAL CONSIDERATION: DATA PRIVACY AND DATA SECURITY

The reliance that AI systems have on data for its training, testing, validation and operation means that data privacy, governance and security are at the forefront of ethical concerns. Although such issues are not unique to the use of data in AI, new challenges present themselves, with some of the common themes relevant to life insurance use cases presented below:

- **Use of big data:** Long-term records can be kept on people by anyone who produces storable data. The data on policyholders that insurers may be able to access ranges from internally sourced data from wearable devices, third-party data on shopping history and credit history, and public data such as social media data from posts, tweets and photos. AI systems can be tasked with extracting information on, and deriving patterns from, nonstandard data sets, creating additional data sets that require adequate privacy, governance and security considerations.

- **Creating new personal data:** When ML is used it can create new data sets that contain sensitive and personal data, and thus that require adequate controls in place to ensure the data is secure and anonymised. For example, AI systems may be able to discern an individual's personality traits based on their digital footprints more accurately than their close friends and family. See (Youyou, Kosinski, & Stillwell, 2015) and (Ahmad & Siddique, 2017).

- **Data regulation blind spots:** ML typically requires the use of large amounts of data. Data regulations such as the GDPR have been created with AI use cases in mind[57] and incorporates ethical considerations,[58] but the strength of data regulation to protect data subjects differs greatly between countries. Even the GDPR, which is perceived as a potential global standard for data protection (Barthelemy, 2019) has some blind spots with respect to AI systems; GDPR requirements do not apply to any anonymised data that is often used to train models, and the GDPR does not oblige insurers to ask for explicit consent to collect or use data on people, so long as there is another legal basis for processing that data (e.g., fraud prevention).

- **Awareness:** AI systems used lots of types of data, some of which data subjects will be aware as it has been provided directly by the user, or is being observed by a user (e.g., wearables) who is aware it may be used. In cases where AI allows for inferred data—where individual characterises are inferred from, what would seem, unrelated data—it becomes less transparent to users that such data is being collected, stored and processed. This lack of transparency makes it harder for individuals to object, and may create the impression that less data is being stored on them than is the case and make it harder to correct data that they feel is inaccurate. For example, how can someone object to data stored on them that identifies them as more or less loyal to particular products or services based on insights derived from loyalty schemes, credit history or social media behaviour? Firms could also take advantage of customers' lack of awareness—few customers will have the time or expertise to understand the details of how their data is used. It is unlikely to be clear how AI ethics vary across insurers—customers may assume all firms are likely doing the same thing. This is particularly likely to be the case for those vulnerable customers who are less technologically or financially aware.

- **Power asymmetries:** Purchasing insurance and making claims are part of a relationship between the consumer and the insurer. Some applications of AI are used to strengthen the capabilities and knowledge of the insurer, allowing them to better process applications and better assess an individual's risk level. Consumers face a limited number of options for insurance. Those with unusual circumstances or tight budget constraints may face even fewer. On this basis, consumers may be unable to opt out of AI use to which they object. There is a limit to transparency when the consumer lacks power.

[57] For example, Article 22 of the GDPR refers to automated decision making, profiling of individuals, and requirements for data protection impact assessments to identify and minimise data protection risks of AI systems.

[58] For example, consent, purpose limitation, data minimisation, accuracy of data, storage limitation and accountability via assigned data protection officers.

- **Retention period:** Certain insurers are looking to record more data on policyholders for the current, or potential, use in AI systems in the future (e.g., data from wearables). In cases where insurers have no current or immediate need for the data (and so are holding it for potential use in the future), this may go beyond the agreed purpose limitation between the data subject and the company. For any personal data that is collected, the longer these data sets are kept, the larger the cybersecurity threat. It also may raise the fear that we are being monitored in perpetuity.
- **New security threats:** Like any new technology system, firms need to develop a fall-back plan for the event that an AI system is faulty, or has an outage, and to ensure there are robust security measures in place to prevent or mitigate the risk of cyberattack. AI systems are also vulnerable to new threats such as:
  - **Model inversion:** it is possible to, in certain instances, reconstruct personal data, or data about individuals who were in a training data set from a model itself (Veale, Binns, & Edwards, 2018), or other model parameters that open the model to exploitation. Whilst there may be data privacy, governance and security laws around the use of personal data, there are no equivalent laws relating to a trained model. For example, whilst you may be able to object to your data being stored by a company, can you object to your data being stored in the *memory* of a ML model when you formed part of the training data?
  - **Model evasion:** For ML applications like fraud detection and in anti-money laundering (and other use cases) the goal is to detect the 'bad thing.' In such cases, the bad actors[59] have objectives, and in the case where they are trying not to be detected, they can consider ways in which they can change their behaviour in order to avoid detection. Evasion attacks are attacks on ML models whereby malicious objects deliberately are transformed to evade detection (prediction by ML that they are malicious), see (Moisejevs, 2020).
  - **Data poisoning:** In these cases, an adversary introduces malicious modifications to the data that is used for training ML algorithms. This could be by inserting instances in the training data that 'poison it' (see "New threats" above for an example), modifying instances in the training data or selectively removing some instances.

NEW THREATS

- A common example of **data poisoning** was Microsoft's Tay, an AI chatbot that was released on Twitter in 2016. The bot caused subsequent controversy when it started to post racist and other offensive tweets through its Twitter account, forcing Microsoft to shut down the service only 16 hours after its launch (Perez, 2016).
- According to Microsoft, this was caused by trolls who performed a 'coordinated effort' to 'abuse Tay's commenting skills,' with the bot making replies based what it had learnt from its interactions with people on Twitter who had been 'poisoning' the chatbot with racist and other offensive messages.

Insurers, as all organisations that are using AI systems, need to consider their approach to these new challenges. They must ensure that they are complying with local data regulation (e.g., the GDPR) with respect to storing and processing data used by, or collected by, AI systems. For example, for those subject to the GDPR this may involve ensuring that they do not pass such data on to third parties without the explicit consent of the data subject and only store it for as long as is required for their current processing activities. Indeed, a suggestion made by the Centre for Data Ethics and Innovation in the U.K. is that the insurance industry should consider creating regulation or industry standards (e.g., through insurance bodies such as the ABI) that discourage the storage of data that is not central to a firm's mission (CDEI, 2019c), with firms regularly reviewing data sets that they are holding to determine whether they are core to the business or can be removed. Guidance provided by other industry bodies that focus primarily on data regulation, such as the Information Commissioner's Office (ICO) in the U.K. (see Section 4), should also be consulted, particularly if there are concerns on whether a current, or proposed, AI use case could be in breach of local data regulations.

Insurers should ensure there is appropriate disclosure and transparency to the public around how it is using data in its AI systems, what data is being used (internal and third party) and what data it is sharing externally (if at all). Furthermore, individuals should be able to request, as it is possible under the GDPR, a full account of

[59] Those who do the 'bad thing.'

what data is being held on them. Together, this should empower some individuals to determine any unintended consequences that the firm's data use in AI systems may have that will impact them (in terms of decisions or outcomes from AI systems) and so to object to the use or storage of certain data. However, such discussions would need to be jargon-free and ideally, to the extent possible, standardised across the industry (in terms of what is disclosed, the layout in which it is disclosed, and how and when such information can be accessed) or more widely. If there is convergence to an agreed set of accountability standards, there will be less confusion in the public and more empowerment around data privacy and governance.

The approach cannot be just retrospective, informing people what data is being used; it also needs to ensure there are mechanisms in place that allow for individuals to object to the use of a service, to have their data captured or used by, or their data used in (e.g., training) an AI system prior to the interaction.

Documentation by a firm should also extend to its use of third-party data, with public registers kept on any third-party data suppliers (including the terms of the agreements where this is not commercially sensitive). For such data, and also for third-party software that uses AI for its operations, where insurers have less oversight over its source, quality and security, firms must do their due diligence to request sufficient assurances that the data or software provided meets the same data quality, governance and security standards it would assign to its own internal data (e.g., that the information they are accessing is accurate and collected with the knowledge of the data subject).

Finally, firms need to consider the tools that are available to identify the risk of cyberattacks to their AI systems (e.g., red team testing[60] or penetration testing[61]), those that may prevent cyberattacks (e.g., by being able to identify cases of model evasion or data positioning), and put in place fall-back plans when systems are determined to be faulty or where there has been an identified security threat to the AI system.

3.10 PUBLIC UNDERSTANDING AND ENGAGEMENT

Are you happy with Facebook photos of late-night junk food being used to increase your premium?

What if they were shared by a friend and you had no control over it?

Are you happy to exchange your data on gym workouts to receive a lower premium?

Readers will respond differently to each of these questions; we each have a view on what we are comfortable with in relation to AI, even if we haven't questioned ourselves before. And each view is unique, shaped by our individual experiences, values and cultures.

As discussed in Section 3.2, if life insurers want to benefit from the use of AI systems, in particular those that are consumer facing (e.g., pricing, underwriting, claims management, customer servicing), it is critical that the company understands the public's opinions on what is morally right, and what is morally wrong, so that it can align the firm's use of AI to meet the expectations of the public.

If they do not take into account public expectations, insurers risk building products or services that will be rejected or face backlash for overstepping boundaries that they didn't realise existed. This can create situations where public perception on what is acceptable is drastically misaligned with those of insurers. Even when firms can evidence that technology is working fairly, robustly, legally and in line with its expectations, the public may still decide they do not feel it is appropriate and so resist against its use.

[60] The goal of red team testing is to simulate the actions of malicious hackers going beyond that of a penetration test to include not only the IT systems but also the people, processes and physical security of facilities.

[61] The goal of penetration testing, or 'pentesting,' is to simulate the actions of malicious hackers, primarily focusing on IT controls and governance, to identify any flaws through which a hacker may try and target you, how your security measures would react, and the possible magnitude of a breach.

Consequently, ethical guidelines (European Commission, 2019), (Institute of Electrical and Electronics Engineers, 2019), (RSS and IFoA, 2019) along other influential bodies (CDEI, 2019b), (Ada Lovelance Institute, 2018)), stress the need to embed public engagement and deliberation mechanisms and improve the public understanding on AI use, namely that AI must be developed '*not alone and not behind closed doors. But hand-in-hand with the public*' (Dellot, 2020). Beyond ensuring that AI is not rejected, public deliberation and engagement allow for challenging questions to be addressed that neither developers nor the public can answer on their own. It also allows the public to be informed of issues and the opportunity to influence views.

To allow the public to articulate what it believes is ethically right and wrong regarding the use of AI, certain mechanisms life insurers could consider are:

- **Citizens' juries:** This involves presenting a group of randomly selected people (who are expected to be representative of the demographic that will use, or be impacted by, the use of AI) with a specific question (e.g., should individuals be given the right not to have their social media data used in pricing?), providing them with consultation from domain experts in a number of sessions over an extended period[62] to use the information they learn to come to a collective decision or recommendation.
- **Citizens' assembly:** This is similar to a citizens' jury; however, they may be larger and have a less prescribed question or decision to make (e.g., should we introduce regulation on the use of AI in life insurance pricing?). The process will be over a longer time period and will involve a learning, deliberation and decision-making stage for those involved.
- **Public opinion surveys and polls:** This involves presenting a questionnaire to a large group expected to be representative of the demographic that will use, or be impacted by, the use of AI in order to gauge public opinion. Typically, such surveys will not include a learning phase, and so are designed to capture the point-in-time views of the public. The questions asked must reflect the level of understanding that can be expected of their target audience.
- **Focus groups:** This typically involves assembling a diverse and representative group of people to participate in a discussion, or to provide opinions and feedback on a product, service, marketing campaign or other matter of interest prior to launch. The group's discussion will be facilitated by a mediator and has the format of a collective interview. In the context of AI systems, this could involve presenting a potential new business use case of AI to a group, and facilitating a discussion to gauge an understanding on its benefits and costs from a user perspective.

The various methods outlined above will provide an insurer with different insights based on the length of time and level of learning and discussion involved. It is important for insurers to ascertain both a clear view of the current level of public understanding and opinion through surveys and polls (based on their current education on AI and ML), and whether such views change in light of a greater level of understanding or education through longer focus groups, citizens' juries and citizens' assemblies.

[62] This can be from a few days to a few months depending on the scope.

CURRENT PUBLIC UNDERSTANDING AND CONCERNS ON AI AND ML

The current level of public understanding on AI and ML and how they are used is not very sophisticated.

In 2017 in the U.K.

9% of people surveyed

RECOGNISED THE TERM

'MACHINE LEARNING.'

3% UNDERSTOOD IT.

In 2017, the Royal Society, a fellowship of distinguished scientists, commissioned a study into public awareness and attitudes towards ML (The Royal Society, 2017). They ran a survey and conducted discussion groups with a selection of U.K. adults. They found that very few participants understood the concept of ML—9% recognised the term and only 3% said they had a good understanding of it. While people recognised some applications of ML, such as personalised shopping recommendations, they didn't understand the underlying mechanics of how it worked.

The study explained the concepts of ML to the discussion groups. The report highlighted that participants were content with the basic idea that data creates predictions, updated by feedback on how accurate those predictions are, and were not particularly interested in learning more. They saw AI as a neutral technology and believed the risks around its use primarily arose from the motives of the people applying it (i.e., good people use AI for good, bad people use AI for bad).

A low level of understanding should not be perceived as an opportunity for unbounded innovation. Insurers should instead perceive this as a risk; any missteps around the use of AI may well be attributed to intentional prejudice or misuse from the insurer, rather than perceived by the public as accidental and unintended.

Despite the low level of general understanding, the results of this exercise revealed that the public were aware of some of the key ethical concerns around the use of AI, in particular: (1) the potential for harm; (2) the societal impact on jobs; (3) the impact on human autonomy; and (4) the potential to restrict humans. Furthermore, the concerns raised are relevant to life insurance applications; see Table 13 where we have mapped the concerns raised in this study to life insurance use cases.

What is interesting is that for each of the concerns, AI could also provide benefits that alleviate the concern; although jobs may be lost to automation, new jobs will be created in data science, data labelling, AI system governance and monitoring. Whilst there is a risk of depersonalisation, in some cases AI can actually help fight against these issues through an increased personalisation of services. In addition, comparing the concerns raised against the ethical concerns raised presented in ethical guidelines in Appendix 2, there are some major omissions such as fairness, accountability and transparency. This points towards the need for insurers (although this is not an issue that is distinct to insurance) to consider how to improve public education on the benefits and risks of AI, in order to facilitate a more productive dialogue.

TABLE 13: PUBLIC CONCERNS ON AI FROM (THE ROYAL SOCIETY, 2017) AND LIFE INSURANCE IMPLICATIONS

CONCERN	LIFE INSURANCE USE CASES WHERE CONCERN MAY APPLY
This technology could harm me and others (e.g., autonomous car crash, misdiagnosis of illness, risk of harm from mistakes while the machine is learning).	This concern typically referred to physical risks which are not a concern for life insurance firms. Other risks where there is potential to harm customers: ▪ NLP wrongly records policy details which impacts decision making on underwriting or pricing. ▪ Discrimination and unfair decisions made in pricing or underwriting. ▪ Poor investment strategy taken based on guidance of AI, or algorithmic trading. ▪ Understated capital requirements from ML applications in reserve approximation and daily solvency monitoring resulting in solvency concerns that put the security of policyholders' benefits at higher risk.
This technology could replace me and others: ▪ Potential for mass redundancies ▪ Could lead to further overreliance on technology (and thus losing the ability to think critically)	▪ Routine tasks are increasingly getting automated through the use of NLP, computer vision and speech recognition throughout the life insurance value chain. ▪ AI issued to replace human decision making (e.g., automated underwriting, robo-advisers) or reduce the need for extensive research by humans (e.g., to support investment decisions). ▪ Customer service roles are being replaced by chatbots and virtual assistants. ▪ Human expertise in areas of pricing, underwriting, guidance may be lost over time.
This technology could depersonalise me and my experiences: ▪ Removal of enjoyable tasks (e.g., driving) by initially improving experiences then gradually taking over all functions ▪ AI is unable to treat people individually and meaningful human interaction is lost	▪ AI use may incrementally take over more and more insurance jobs; initially focussed on automation of manual tasks and augmenting the work of professionals whose work requires judgement, but as AI develops, these tasks too may be taken over by AI. ▪ Chatbots, virtual assistants in customer servicing and robo-advisers to replace face-to-face insurance guidance and reduce conversation that people may enjoy.
This technology could restrict me and others—humans are too unique to be understood by a machine especially on subjective topics.	▪ Contextual inferences from photos and tweets on social media used in pricing or fraud prevention (e.g., a photo from 10 years ago of you with a cigarette, may not mean you are now a smoker, but the ML algorithm may not deduce this subtlety). ▪ Ability for a robo-adviser to provide insurance guidance that considers the unique financial needs of an individual.

That said, regardless of any limitations, studies like this one can be used by insurers to gather a broad understanding of where they may need to engage with the public, provide reassurance to the public and areas where they should tread carefully before launching new use cases.

DEEPER MECHANISMS—INCORPORATING PUBLIC EDUCATION

To address the general lack of public understanding on what AI is and how it can be used being a barrier to efficient deliberation, alternatives that include an element of public education can be considered such as citizens' juries and assemblies. By allowing participants to learn about the technology and use cases in detail, they are able to give a more informed view on their opinions on ethical considerations.

It is also useful in cases where opinion is likely going to be divided, such as the use of AI in life insurance; background information, framing and facilitated discussion are required to come to a general consensus (Solomon & Abelson, 2012).

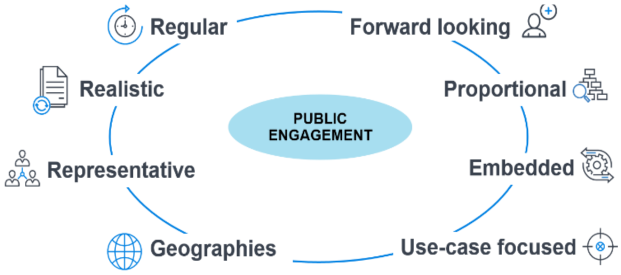
Examples of citizens' juries which have been carried out to gauge opinion on AI include the Forum for Ethical AI conducted by the Royal Society of Arts (RSA) and DeepMind (The Forum for Ethical AI, 2019) and Project ExplAIn conducted by the ICO and the Alan Turing Institute in the U.K. (ICO, 2019f)[63].

[63] See Section 4 for further information on this case.

PUBLIC DELIBERATION—KEY CONSIDERATIONS

In determining what public engagement and deliberation methods (or combination of methods) to use to gather information on AI use issues relevant to life insurance applications, insurers face a number of considerations that will inform this decision (see Figure 15) in particular:

FIGURE 15: SOME PUBLIC ENGAGEMENT CONSIDERATIONS



- **Public engagement must be regular.** Public opinion is not static and insurers need to keep pace with changing societal views and be prepared to adapt its use of AI to reflect changing views. This may require the consideration of fall-back options (i.e., not using an AI system) in the event that public perception materially changes. This may have greater implications for any processes that have been completely replaced by an AI system.

  In addition, regular engagement allows for the collection of public views to keep pace with the speed of innovation; an insurer cannot simply be comfortable that conclusions from an opinion poll or focus group conducted in 2017 are relevant for a new use case being developed in 2020.

- **Public engagement must be designed to solicit realistic views** of the public on the use of AI for specific use cases. For example, surveys suffer from the fact that people’s stated view on how they feel may not align with their actual views,[64] and the discussions in focus groups can lead to groupthink whereby individuals have their views influenced by others in order to facilitate a more harmonious discussion (GroupSolver, 2015).

- **Public engagement needs to gather a diverse range of views from society** by targeting a representative mix of people in terms of characteristics for the specific market that may be affected by the AI use case (e.g., age, occupation, ethnicity, culture and religion). This will allow opinions, views and considerations that were not apparent to the developers, deployers or those who are responsible for the ongoing review and management of AI systems to be shared, allowing for a greater alignment between the AI system that has been developed with public expectation. It will also highlight where views differ amongst groups that may need to be considered; it has been shown that younger people may be more supportive of the use of AI than older people (RSA, 2018).

- **Where an AI system is to be used in multiple geographies,** insurers cannot place reliance on public engagement from one country and assume it will apply to another; there will be cultural differences that need to be captured through separate public deliberation.

- **Public engagement needs to be focussed on specific use cases** since public opinion can be incredibly contextual. For example, a loosely defined question of *Are you happy that AI can be used to categorise your Facebook photos?* may solicit a different response to *Are you happy that AI can be used to categorise your Facebook photos* ***and be used to influence your life insurance premium****?* or *Are you happy that AI can be used to categorise your Facebook photos* ***and be used to decrease your insurance premium (but never increase it)****?*

[64] For example, it has been shown that in surveys, consumers say they’ll pay double for goods than they’re actually willing to spend (Crane, 2018).

- **The cost to the insurer of seeking public opinion should be proportional to the perceived risks** associated with the use case, and there should be consideration of cases where industry-wide public deliberation would be a more cost-efficient mechanism. In the U.K., industry bodies such as the ABI or financial regulators like the FCA may be well-placed to gauge general or broad public opinion on the use of AI, particular for use cases that are gaining more widespread adoption. However, insurers would have less control over what is asked or discussed, when the engagement takes place and to whom the public engagement is targeted; consequently, it may not get answers to specific concerns on newly developed or potential use cases, or not do so in a timely enough manner, or answer concerns that may apply only in specific geographies.
- **The scope of public engagement needs to be broad and forward-looking**, not only assessing the public's changing opinions on current AI use in life insurance or opinions on future use cases but also gauging emerging public views on noninsurance AI use cases that could have an impact on current or potential use cases in the life insurance sector.
- **Public engagement mechanisms need to be embedded into at all stages of the AI life cycle.** Owing to the continuous nature of AI systems, insurers need to consider how to create ongoing and sustained infrastructures for seeking public opinion throughout the design, development, deployment and ongoing governance stages of the AI life cycle to support decision making in line with public expectations.

The issue goes beyond pure deliberation and onto what an insurer would actually do in light of a greater understanding on public views. Firms must determine their own internal metrics or frameworks for making decisions in light of this information. Do you proceed in deploying an AI for underwriting using social media data because the majority of people say they are comfortable with it? What if you have this majority, but a particular group of people object? One option would be to tailor the use case of AI to provide the option for the public to decide whether it wishes to engage with an AI system (i.e., an opt-out option). In this case, care must be taken not to disadvantage or marginalise those groups that do not wish to use the AI system in any way. Firms may also seek to persuade dissenters over time of the benefits of the system.

3.11 COMPETENCE AND PROFESSIONALISM

Those who develop and create ML models must meet certain technical standards of competence. They must also meet nontechnical standards such as understanding the ethical concerns around the use of AI and how this should influence the way in which they build and document the models that they create.

Whilst evidence of the requisite technical skills to create ML models could be determined based on qualifications and experience, it is more challenging to determine if individuals meet standards on their consideration of ethical issues.

Professional bodies generally play a part in ensuring that roles of responsibility are filled by individuals with the requisite high level of technical expertise but also place requirements on ethical considerations and behaviour. Professional accreditation could be seen as a mechanism for ensuring that those involved in the development of AI systems have sufficient understanding of their responsibilities in relation to ethics. It also allows them to be held accountable if they do not meet the expectations of the profession in relation to ethical issues through the use of disciplinary actions, including (in extreme cases) expulsion from the profession.

In insurers, certain roles must be carried out by professionals (e.g., legal and actuarial professionals) who are accredited by respective professional bodies. There is currently no explicit requirement for certain roles in the development or ongoing validation and control of AI systems to be carried out by those in a professional position. Despite there being no current legal or regulatory requirement for professional accreditation of those using AI, given the technical and ethical training and background required to understand AI, insurers may look to fill roles of responsibility with individuals who are accredited by existing professional bodies. In the U.K., this would include the Institute of Analytics (IoA), the Royal Statistical Society (RSS) and the Institute and Faculty of Actuaries (IFoA). ML is used in actuarial modelling applications, so there will be cases where actuaries working in data science will be the developers of ML models and as such will already be bound by professional obligations that will equally apply when designing AI systems. For details on how IFoA professional guidance would apply to AI use, see Section 4.3.

Post-development, certain ethics guidelines place great onus on the developers of AI systems to do all they can to ensure that those who will operate the system are competent and have sufficient tools to mitigate the risks associated with the system (Institute of Electrical and Electronics Engineers, 2019). This includes requiring developers to communicate to those who will operate the AI system the types of knowledge and expertise that would be required to manage the system as a whole and the components of the system (e.g., data, parameters, controls). Those operating the system have the responsibility of ensuring that they meet these requisite competencies and have access to those with more specific technical knowledge if necessary.

Finally, those holding positions of wider oversight for AI systems (model assurance, risk management) should also be able to demonstrate evidence of competence, including relevant training and qualifications. In ensuring individuals are competent in their role, inspiration can be taken from existing regulation that applies to controlled functions or those requiring a certification, such as those required by financial regulators (FCA, 2020b), or as described under Solvency II's Fit and Proper requirements.

DEVELOPER VS OPERATOR COMPETENCE

Example: *A life insurer uses NLP techniques to process unstructured data regarding an applicant's health and medical history, and to draw out key information to assist in the underwriting decision.*

Developer competence:

- **Technical understanding:** On model and data selection (e.g., what the key words the application should search for) and the business problem (e.g., how the application is going to process the data).
- **Ethical considerations:** Safeguards for privacy and security concerns on using personal data, tools to identify and mitigate risk on bias in training data built into the system.

Operator competence:

- **Less technical understanding required:** Understand the underlying components of the NLP in case justification is needed where cover is limited or declined. Can elevate questions to a more technical team if needed.
- **Ethical considerations:** Understands ethical considerations in relation to the use of the NLP in underwriting, how to use tools and specifications provided by the developer, and how to identify and escalate concerns.

3.12 KEY TAKEAWAYS

- The ethics of AI questions both whether, and how, AI should be used.
- It also considers the need for a common, systematic application of ethical principles to guide such considerations, addressing this through the creation of guidelines, standards, frameworks or shared industry standards and norms.
- Recently, there has been a proliferation of guidelines from academia, the private sector, government bodies, industry bodies and think tanks that aim to provide guidance to firms on how to develop and use AI in an ethical, moral and responsible way.
- Such guidelines have shown a convergence to a number of core ethical principles that insurers should be aware of when developing AI systems for life insurance use cases:
  - **Transparency:** – This includesincluding disclosing when and where AI is being used to relevant parties, telling individuals when they are interacting with AI, and being open in providing details on how AI systems have been developed and work in making decisions, recommendations or other outputs.
  - **Fairness** – : Tthe decisions, recommendations and other outputs of AI applications must be perceived as fair, and not lead to unjustified discrimination to any groups or individuals.
  - **Accountability:** – AI systems cannot be held accountable for the decisions, recommendations or other outputs it they makes in cases that are deemedwhere these are unfair, discriminate or cause harmharmful. Thus, those responsible for decisions regarding AI systems, for a particular business use case, must be held accountable: from development, deployers, those managing the system, the firms' risk management function, the management of the firm, and the bBoard. Sufficient controls and governance frameworks must be put in place that enable accountability and responsibility for the technology.
  - **Explainability:** – This involves eensuring fthat firms aim to address the "black -box" problem of AI by aiming to understand how their models work to make decisions, recommendations or produce output, and how this can be communicated to relevant stakeholders in a way that they would understand.
  - Data privacy and security: – This entails eensuring that data privacy, governance and security challenges unique to the use of AI are considered and sufficiently addressed.
- Ethical requirements can be extensive, and insurers should consider a **risk-based approach** to allow proportionality in the governance of ethical issues. Focus should be placed on those AI applications that could result in the greatest harm to consumers, or the financial security of the firm.
- Insurers should not be making decisions regarding the ethics of AI in isolation but instead following the guidance of relevant local regulation, advisory groups (e.g., data protection, industry and data ethics bodies) and, importantly, the public who will ultimately have the strongest voice on what they deem is morally right and not right in terms of AI use. Seeking public views is imperative to build public trust, and thus, public legitimacy in the use of AI in the insurance sector, and more widely.

4. Regulation and guidance

Despite the wealth of ethical frameworks and guidance published by governmental bodies, civic groups and other organisations, see Figure 10 and Figure 11, relevant guidance will vary from country to country. In this section we examine some of the key organisations whose guidance is relevant from a U.K. life insurance perspective, considering their aims, focus and providing reference to any recent relevant guidance provided.

In particular, we focus on the following groups that may influence how insurers consider the ethical risks posed by AI presented in Section 3 of this report:

- **Financial supervisory bodies:** Such as the PRA and the FCA
- **Data regulation authorities:** Considering the GDPR and the ICO
- **Professional bodies:** Focusing on the standards that govern the work of U.K. actuaries working in AI and ML
- **Centre for Data Ethics and Innovation:** An independent advisory body that assesses the risks and benefits from data-driven technology for the U.K. government
- **European Commission:** Has published extensive guidelines on ethics that are of relevance to U.K. and European insurers
- **Independent of government, regulators and industry:** Other groups that are responsible for providing insight to government and regulators—the Alan Turing Institute and the Ada Lovelace Institute

We then conclude on some international considerations.

4.1 FINANCIAL SUPERVISORY BODIES

BANK OF ENGLAND AND PRUDENTIAL REGULATION AUTHORITY

As the central bank in the U.K., BoE has a goal to promote the good of the people by maintaining monetary stability and financial stability.

It is the requirement for financial stability, in particular, that there is public trust in U.K. financial services providers, which drives the BoE's AI interests. Consequently, the PRA, the part of the BoE responsible for the prudential supervision of insurers in the U.K. (along with banks, building societies, credit unions and major investment firms) will play an increasing role in ensuring that trust can be maintained or further developed, as insurers increasingly adopt AI technologies.

In particular, the PRA has two statutory objectives which will both be relevant in terms of the ethical adoption of AI:

- **Promote the safety and soundness of the firms it regulates:** The PRA will need to ensure that the risks associated with the use of AI are fully understood by the companies it regulates, including the impact these have on their operations and financial security. It will need to ensure that such risks are monitored, and managed, with sufficient controls and fall-back plans in place, such that the risk to firms' financial security is minimised.
- **Contribute to the protection of policyholders:** Unlike the FCA, which seeks to ensure that consumers are treated fairly in their interactions with insurers, the PRA is responsible for ensuring that insurers are able to meet their obligations to policyholders, particularly in the case of those obligations only emerging many years after a policy is taken up with an insurer. This is similar to the point above, that insurers' use of AI must not endanger the firm's financial security.

Consequently, the focus of the BoE, and the PRA, in relation to AI will be on the implementation of robust governance frameworks to prevent adverse outcomes to policyholders and the financial sector more generally. As such, their publications on AI to date have been focussed on two broad areas that align with their interests: transparency and explainability, and governance.

Transparency and Explainability

Much of the BoE's discussion on AI implementation has been on black boxes, and it has published academic literature on transparency tools and frameworks in use cases relevant to the financial services sector (e.g., Shapley value[65] and regression frameworks, and the Quantitative Input Influence method).

For example, the BoE has discussed use cases of AI (in particular ML) and the associated explainability challenges for the case of:

- A neural network to predict changes in unemployment one year ahead using other macroeconomic and financial variables (Joseph, 2019)
- Random forests, support vector machines and neural networks to determine the relationship between high-school financial education and financial proficiency and spending and saving behaviour (Joseph, 2020)
- A gradient boosting tree to predict mortgage defaults based on data of outstanding U.K. regulated mortgages from the FCA (Bracke, Datta, Jung, & Sen, 2019)
- A random forest ML algorithm to detect early warning signs of bank distress (Suss & Treitel, 2019).

Governance

As regulators feel that governance failings are the primary cause of prudential failure (Proudman, 2019), it is perhaps no surprise that robust governance around the introduction of AI to the financial service sector is a primary concern for the BoE and PRA. In June 2019, James Proudman, who was at that time the executive director for U.K. Deposit Takers at the BoE, gave a speech at the FCA Conference on Governance in Banking (Proudman, 2019) whereby he set three challenges for boards and management in relation to the use of AI in the financial services sector in the U.K., and three principles for governance in response to them:

1. **Governance of data and models** in particular:

    a. **Exclusion and privilege:** Can boards ensure that data is not being used to unfairly exclude individuals or groups of individuals, or otherwise promote unjust privilege to certain individuals or groups?

    b. **Bias:** Given AI systems use historical data, how can boards prove that such data does not contain long-standing bias towards certain groups that will result in the AI system perpetuating such bias?

    c. **Algorithmic bias:** How can boards ensure that the underlying algorithms are free from bias such that they do not exaggerate potential and/or unintended human prejudices?

    The speech identifies that firms should consider oversight mechanisms both in the design and deployment stage, followed by continued supervision and testing (e.g., HITL) and considerations of appropriate fall-back mechanisms and redress in the event that things go wrong.

    In the speech, it was suggested that boards should continue, and perhaps increase, their focus on the proper governance of data, in particular what data should be used, how it should be used and tested, and evaluating how reasonable the outcomes are from the model in light of the data provided.

2. **The role of people**

    The speech highlights that much regulatory reform over recent years has been focussed on overcoming 'human' problems in workplaces such as cultural failings and groupthink, poorly linked remuneration incentives, and short termism. Such responsibilities do not disappear with AI systems, despite the fact some decision making previously carried out by humans may be carried out by machines.

    In fact, issues around holding people accountable will become murkier with the introduction of such systems as it becomes harder to ascertain the primary cause of problems, and thus attribute accountability to individuals. For example, how do you determine whether the fault is due to poor design of the AI system (assigning accountability to the developer) or poor implementation and understanding of the management or board of the firm on how the AI system works? In addition, machines lack morals and cannot whistle-blow on improper practice, so accountability will have to be with those responsible

[65] A model-agnostic method for machine learning models to help assist explainability of the learning process of a model (see Table 12).

for the technical management of the AI systems, and ultimately the boards of firms that have made the decision to implement such systems.

Consequently, the speech suggests that firms will need to consider how to allocate accountability, including under the PRA's SM&CR.

The advice to boards provided in this speech was to continue to focus on the oversight of human incentives and accountability in light of the introduction of AI systems in their firms.

3. **Execution risks in the transition period**

   As financial services firms look to transition into using AI systems more widely, the risks and costs from this transition (i.e., from changes in processes, systems, data management, skill sets) need to be considered by boards. In particular, there needs to be sufficient oversight, with a focus on attracting talent, and training and developing skill sets and controls required to mitigate transition risks at both the senior level and throughout the organisation.

FCA

The FCA is responsible for regulating the conduct of financial services in the U.K. with the role of ensuring consumers are protected, that markets are honest, fair and financially stable, and promoting competition between financial service providers.

Given these responsibilities, ensuring the ethical use of AI in the financial service sector will be an increasing requirement of the FCA as implementation becomes more widespread in the U.K. market. For example, the role of the FCA may likely include (but is not limited to) ensuring that AI systems:

- Are designed, developed and used ethically and in the interest of consumers
- Do not result in unfair consumer outcomes (in fact can be used for positive consumer outcomes)
- Do not introduce any new ethical risks to the stability of financial markets

In 2020, the FCA is still at the stage of developing its own understanding on the extent to which AI is being developed, deployed and used in the financial services sector in the U.K. 'Machine learning in U.K. financial services' (BoE and FCA, 2019), coauthored with the BoE, in October 2019 was the first step in this journey (see 'Survey on machine learning' below for more details). The paper outlined the results of a survey on how embedded AI systems are (and how advanced such implementations are) in the U.K. financial sector, split by industry.

Financial Services AI Public Private Forum (AIPPF)

In 2020, the FCA and the BoE have established a forum, the AIPPF (FCA, 2020a), which aims to ensure there is dialogue between supervisors and the industry participants to better understand:

- The current (and potential) uses of AI in the financial services sector
- The costs and benefits of deploying AI in the financial services sector
- The risks associated with the application of AI

The AIPPF will be a closed (invite only) membership group bound by its terms of reference (BoE and FCA, 2020) comprising of the FCA and BoE, and selected members with relevant background and experience in the development of AI, public authorities and academics. Although the terms of reference do not explicitly mention ethics it does include areas that will inevitably involve discussion on ethics, such as governance, trade-offs between explainability and performance of the AI system, model risk management and validation, and issues around data.

To date the FCA has not published any principles, guidance, regulation and/or industry-specific best practice that U.K. life insurers (and financial service providers in the U.K. more generally) could use to ensure the safe and ethical adoption of AI systems. It will be the responsibility of the AIPPF to explore whether the development of these could help to ensure AI is implemented safely in the U.K. financial services sector.

SURVEY ON MACHINE LEARNING (BoE and FCA, 2019)

To gauge the extent to which AI (and in particular ML) is being used in the financial service sector, in 2019 the BoE conducted a survey of 106 financial service firms including banks, insurers and other financial institutions in the U.K., to get a better understanding of:

- The nature and level of deployment of machine learning AI systems in the U.K. (i.e., how many firms are using, to what extent are they used)
- The areas of the business where there has been the most implementation
- The maturity of applications (i.e., are they at the initial experiment phase (i.e., predeployment) or fully deployed in the firm)
- The benefits and risks associated with this technology

The results of this survey were released in October 2019. Regarding the ethical use of AI, the BoE stated in this report that it supports the ethical development and deployment of machine learning, citing that the role of the BoE and FCA will be to monitor the application of machine learning to ensure this (alongside understanding the role such technology will have on the U.K. economy).

Some key points from this study of relevance are:

- **Collaboration with the ICO on ethics:** The BoE stated that it will continue to collaborate with the ICO and other domestic authorities to address any potential ethical issues that arise from the use of machine learning.
- **Governance structures are not really evolving:** Of the 106 companies that responded, only four noted that they are establishing an ethics function that would address the particular issues raised by machine learning models and the use of new data sources.
- **Financial regulation is not seen as a barrier to AI:** Innovation is not currently being hampered by regulation. The majority of respondents (75%) do not consider that PRA or FCA regulation provides an unjustified barrier to deploying machine learning in their firm.
- **Regulation as an enabler:** The BoE stated that some firms reported that further clarity in the form of guidance or regulatory expectations for specific use cases (i.e., best practice examples) could help firms better understand how existing regulations apply to newly developed AI systems.
- **Confusion on ethics:** The survey also included a question specifically designed to identify each firm's perception of potential ethical issues arising from the deployment of machine learning applications. Interestingly, it received a somewhat confused response, with some firms answering it by addressing individual ethical issues (e.g., bias) and others taking a more holistic view, focusing on the firm's approach to addressing ethical issues resulting from the introduction of AI as a whole (e.g., responding that machine learning ethics are aligned with the firm's conduct and technology use rules). Furthermore, around 10% responded that they didn't know, and around 8% responded that they felt there were 'no ethical issues' as a result of machine learning deployment.

Taken together, the above may indicate that there is a lack of education on ethical issues facing firms deploying AI systems in the U.K. financial services sector, and furthermore a lack of real guidance on expectations from regulators, both contributing to confusion.

In particular, the results of the survey may signify that U.K. financial regulation has not kept pace with technological development. The BoE stated that 'well-judged regulation is intended—by design—to be a barrier to certain practices in order to maintain financial stability or protect consumers' and therefore in firms not perceiving such a barrier it could indicate that 'future regulation may need to be updated or adjusted to account for developments in machine learning.'

4.2 DATA REGULATION AUTHORITIES

THE GENERAL DATA PROTECTION REGULATION

The GDPR is the EU data privacy and security regulation which came into force in May 2018 and imposes a number of responsibilities on organisations which control or process personal data[66] (European Commission, 2018f). These strict provisions on acceptable use are intended to generate trust and are part of the EU's competitive approach to technology with ethics (European Commission, 2018b).

As of the 31 January 2020, the U.K. is no longer in the EU but remains bound by EU rules until the end of the transition period, ending 31 December 2020. After this date, the GDPR, as EU legislation, will no longer apply to the U.K. However, the British government has stated that it plans to incorporate the GDPR into U.K. data protection law as 'U.K. GDPR,' meaning there would be little practical change to firms in the immediate future (ICO, 2020c). Additionally, EU GDPR would still apply to any British firms processing or controlling personal data related to European citizens. The GDPR states that all data processing[67] must have a legal basis to justify it (European Parliamentary Research Service, 2020).

The legal basis may be:

- **Consent:** This must be specific and granular, with separate requests for different uses of the data, rather than blanket consent. Consent must be given freely and explicitly.
- **Necessity:** There are a number of ways organisations may justify processing as necessary to achieve a particular aim, for example:
  - To perform or enter into a contract
  - To comply with a legal obligation
  - To protect vital interests
  - To perform a task in the public interest or while executing public authority

  When this basis is used, it doesn't permit additional processing beyond what is needed to fulfil the aim (e.g., performing business analytics or profiling a customer for marketing purposes).
- **Legitimate interest:** There is a balance between the data controller's[68] legitimate interest[69] and the interests and rights of the data subjects.[70] If an individual's data is for use only in training data, strong security methods may be sufficient to ensure the person's data is safe and their interests are unaffected, such that this standard would be met. When a model is used to produce conclusions about an individual, the person is more affected by the processing and should be asked for consent and be able to opt out.
- **Repurposing:** Repurposing of data is only permitted when the new processing is compatible with the purpose given when the data was collected and where the new processing has a legal basis. This may limit the use of big data sets collected across many sources (e.g., combining an individual's financial records, medical history, social media profile and claim history). However, use for statistical processing, which doesn't affect an individual directly, may be permissible (European Parliamentary Research Service, 2020).

Processing of individuals' data, must be done while recognising the rights for data subjects as set out in the GDPR. We consider some examples of how these may impact the use of data in AI systems in Table 14.

TABLE 14: GDPR RIGHTS AND AI IMPLICATIONS

[66] Any information relating to an identified or identifiable individual, who can be identified from that data in combination with other information. This includes pseudonymised data but not truly anonymised data.

[67] The GDPR definition of processing covers any set of operations performed on personal data, such as collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.

[68] A data controller is the person or organisation who determines the purposes and means of the processing of personal data.

[69] This could be commercial interests, individual interests or broader societal benefits.

[70] A natural person who is identified or identifiable from the personal data processed.

RIGHT	IMPACT ON AI USE
The right to be informed	This includes the right to receive meaningful information on the logic, significance and expected consequences of automated decision-making systems affecting individuals. It does not go as far as to mandate a 'right to explanation' with a full description of how a decision was made in an individual case, though this would be desirable for consumers (The Alan Turing Institute, 2018).
The right of access The right to rectification The right to erasure	Where personal information is used to train an AI model (even if it has been processed and cannot obviously be linked to a particular individual), this is subject to the GDPR, and organisations are expected to share, correct or remove data if requested (ICO, 2019d). Where training data is needed to be stored for retraining, refining or evaluating the system, organisations are expected to take a case-by-case approach as to whether they can fulfil requests to rectify, or erase, the data from a model, taking into account considerations such as the impact that the rectification or erasure may have on the ability to demonstrate the correctness of the model's results. Models trained on personal data which has been subject to a request for rectification or erasure can remain unchanged, as long as they do not contain the data and cannot be used to infer it. Where the model contains training data (such as with SVM models which retain a small number of training data examples to be used by the model in deployment), the model would require deleting or retraining on an updated data set.
The right to restrict processing	When an individual has requested that their data be restricted from processing, this data cannot be processed except to store it[71]. This restriction is often temporary, while accuracy or legitimate use is contested (ICO, 2019e), so adds to the issue of the right to erasure (see above) around whether data can be removed from (and possibly re-added to) models.
The right to data portability	Data subjects can request a copy of all personal data they have provided to a controller, to be given in a structured machine readable format. This is only the case where data processing is based on consent. Data 'provided' by individuals includes data provided by the individuals (e.g., data input on an online application form) and data collected when tracking their activity (e.g., when they completed the form and on what device) (ICO, 2019e). Inferred or derived data about individuals (e.g., their risk grouping based on data provided) is not covered by this right, though if classed as personal data, individuals may request it via their right of access regardless.
The right to object	If the data subject has reason to request that the processing of their data be stopped, and the data controller cannot demonstrate compelling overriding grounds to continue, the processing of the individual's data must end. This right explicitly mentions profiling, where personal data is evaluated to analyse or make predictions about a person, a frequent use of AI. Individuals have an unconditional right to object to direct marketing profiling and a right to object to processing done for research, unless this is done for reasons of public interest (European Parliamentary Research Service, 2020). This right only applies when processing is done on the basis of public interest or legitimate interest, so would not apply, for example, to risk profiling done to price an insurance contract.
Rights in relation to automated decision making and profiling	There must be suitable safeguards around automated decision-making systems, where there is no human involvement, and the decision has a legal or similarly significant effect[72] on individuals—as a minimum individuals have the right not to be subject to a decision made in this way. The human involvement must be 'meaningful,' not just 'rubber-stamping' the decision, to fall outside the scope of this right (ICO, 2019b). However, use is permitted if the automated processing is necessary for a contract between the individual and data controller, authorised by law or based on explicit consent. This would permit, for example, an automated credit assessment performed before entering into a financial contract, and automated profiling done to detect fraud (ICO, 2019e). That said, the use of an automated system must be 'necessary' (e.g., due to the number of cases to be assessed, or where AI can outperform human judgement) (European Parliamentary Research Service, 2020). As this type of processing is considered high risk, firms must complete a Data Protection Impact Assessment (DPIA) to document their awareness and mitigation of possible risks.

[71] With exceptions for processing done with the individual's consent, as part of legal claims, to protect another person's rights or for reasons of important public interest.

[72] That is, affecting a person's legal status or their legal rights, such as assessing entitlement to government benefits, a credit application or making a recruitment decision.

The GDPR sets out the following seven key principles for processing data:

1. Lawfulness, fairness and transparency
2. Purpose limitation
3. Data minimisation
4. Accuracy
5. Storage limitation
6. Integrity and confidentiality (security)
7. Accountability

These bring up similar considerations around AI use to those raised by the rights and need for a legal basis under the GDPR.

The GDPR's principles of fairness and transparency require data subjects to be informed about how their data is used and that systems do not discriminate on protected characteristics.

The principles of data minimisation, purpose limitation and storage limitation could be taken to suggest that firms cannot hold any data other than what is essential, a potential barrier to creating the big data sets used in AI systems. However, this principle is intended to apply with a sense of proportionality—so where the storage of data is beneficial to a firm, and the risks from holding it are minimal (for example, when pseudonymised data is used in a training data set), firms can justify holding it. Additionally, the reuse of data for a different purpose is permitted under certain conditions, as outlined earlier, on the legal basis of repurposing data, broadening the scope for what data can be used in AI systems. The GDPR principles do not rule out AI use but do require firms to carefully consider how they use individual's data in their AI systems.

ICO

In the U.K., the ICO is the regulatory body responsible for data protection. Its remit is broad—investigating and enforcing data regulation across all industries—but its recent work includes a series on AI, focussed on the explainability of AI decisions and data protection in AI models.

Some noteworthy publications on ethics and AI include:

- **Project ExplAIn Interim Report** (ICO, 2019)**:** The ICO used citizens' juries to consider the trade-off between accuracy and explainability of AI systems in given healthcare, recruitment and criminal justice scenarios. They concluded that the optimal trade-off was heavily dependent on the particular use case, considering the urgency and impact of the decision, whether it was determined based on fixed characteristics or factors an individual could change, whether these were objective factors or subjective attributes, and the scope for bias and interpretation in the decision-making process. The ICO highlighted that this aligns with the GDPR requirement that firms provide meaningful information and adopt suitable safeguards,[73] where 'meaningful' and 'suitable' will be defined by the context.
- **Explaining Decisions Made with AI** (ICO, 2020a)**:** This provides best practice guidance on how organisations should explain decisions made by AI systems to those assessed by the system. The guidance covers:
  - What regulation, such as the GDPR, requires in the context of explaining AI
  - The broader benefits and risks of explanations
  - The type of explanations which may be provided
  - How the explanation may be affected by context
  - A detailed breakdown of how to provide an explanation, aimed at technical teams who are developing AI

[73] See the 'Right to be informed' and 'Rights in relation to automated decision making systems and profiling' in Table 14.

  - Governance guidelines, covering the responsibilities which lie with developers, managers, users and compliance teams when implementing an AI system
  - Why policies and procedures for explaining AI are necessary and how these should be implemented
  - The documentation required to show the AI design and deployment process was done responsibly, and to deliver appropriate explanations
- **AI Auditing Framework** (ICO, 2019a)**:** A final report has yet to be issued, but a structure and views on specific risk areas have been released. These consider the following risk areas:
  - That when human reviewers are included as a safeguard in an AI system, they either put too much trust in the system or are not able to interpret it, so they do not provide meaningful oversight
  - Accuracy of AI systems outputs and performance measures
  - Known security risks exacerbated by AI
  - Explainability of AI decisions to data subjects
  - Human biases and discrimination in AI systems

  ICO also reiterates the importance of carrying out a DPIA for AI systems which process personal data, something which in most cases will be a legal requirement under the GDPR (ICO, 2019C).

4.3 PROFESSIONAL BODIES—U.K. ACTUARIAL PROFESSIONAL REQUIREMENTS

Individuals working with AI systems may be subject to the regulations of professional bodies; for modelling carried out by U.K. actuaries, this will be the IFoA. Other professional standards will apply to data scientists who are members of other professions (e.g., the RSS, IoA).

U.K. ACTUARIAL PROFESSIONAL REQUIREMENTS

‘The ethical use of data is a cornerstone of the actuarial profession.’ —Colin Thores, Education Actuary at the IFoA (Bryce, 2020)

Actuaries in the U.K. are subject to the profession’s guidance and standards. These include mandatory ethical standards for all members of the IFoA. The requirements on actuaries are set out in three main areas:

- The Actuaries’ Code (IFoA, 2019b) published by the IFoA
- Actuarial Profession Standards and Guidance Notes published by the IFoA
- Technical Actuarial Standards (TAS) published by the Financial Reporting Council, the body that provides independent oversight of the regulation of actuaries

Furthermore, the IFoA has issued guidance for data science use and launched a data science credential designed specifically for actuaries. Looking at each of these in turn:

The Actuaries’ Code (IFoA, 2019b)

The code is a short principles-based document requiring of IFoA members:

- Members must act honestly and with integrity.
- Members must carry out work competently and with care.
- Members must ensure that their professional judgement is not compromised, and cannot reasonably be seen to be compromised, by bias, conflict of interest or the undue influence of others.
- Members must comply with all relevant legal, regulatory and professional requirements.
- Members should speak up if they believe, or have reasonable cause to believe, that a course of action is unethical or is unlawful.
- Members must communicate appropriately.

Whilst all of the principles are relevant to actuaries working in the field of AI, the third principle is particularly relevant. Actuaries need to ensure that their judgement is not reliant solely on output of an AI system, but must be able to justify their conclusions independently of such AI systems and be aware of, and allow for, any implicit or explicit bias in an AI system.

Actuarial Profession Standards and Guidance Notes

The 16, or so, actuarial standards and guidance notes do not directly address AI (IFoA, n.d.). Perhaps the most relevant is Actuarial Profession Standard X2: Review of Actuarial Work (IFoA, 2015). The standard requires actuaries to consider whether the work they are undertaking needs to be peer reviewed. Two of the circumstances to consider are: the degree of difficulty of the piece of work and its complexity, and whether the circumstances of the piece of work make it more likely that errors could be made.

Given the complexity and general lack of transparency of AI systems, both of these criteria are likely to apply to AI-related work, and it will require peer review by an appropriately experienced actuary who is independent of the work undertaken. Part of the peer review would be to understand why the results from an AI model were appropriate and what steps had be taken to ensure that the results were not biased.

Technical Actuarial Standards

TAS 100: Principles for Technical Actuarial Work (Financial Reporting Council, 2016) is a relatively short, high-level document that does not specifically mention AI but is particularly relevant to AI in its requirements relating to judgement, data and modelling:

- Where judgement is exercised it must be in a reasoned and justifiable manner; material judgements must be communicated to users so that they are able to make informed decisions. Where this judgement has been informed by AI it should be clear how and to what extent it relies on the results from AI.
- For data used in technical actuarial work, the checks and controls that have been applied to that data and any actions taken to improve insufficient or unreliable data has to be documented. This may be challenging in the context of AI involving unstructured data and training sets. It needs to be clearly communicated what data was used, why it was relevant and what its limitations and their implications are.
- Models used in technical actuarial work must be fit for the purpose for which they are used, and be subject to sufficient controls and testing so that users can rely on the resulting actuarial information. Controls and tests that have been applied to a model need to be documented. Communication of model results needs to include explanations of any significant limitations of the models used and the implications of those limitations. All these requirements may not be straightforward when using AI models and ML, which by its nature does not easily provide documentation of how the models work and their limitations.

The other TASs are focussed on specific areas of actuarial practice and do not add materially to the requirements on actuaries of TAS 100 when dealing with AI.

In summary, U.K. actuaries have quite onerous professional requirements placed on them. These requirements were not drawn up with AI in mind, but the general principles are relevant as illustrated in the example areas discussed above.

RSS and IFoA Guide for Ethical Data Science

The IFoA has produced specific guidance on data science in its publication A Guide for Ethical Data Science (RSS and IFoA, 2019), produced jointly with the RSS. This intends to complement the existing professional guidance by providing a practical framework for data-science-specific ethical issues, including an evaluation checklist.

The guidance suggests that:

- Practitioners must keep a broader awareness of the impact of modelling decisions.
- Ethical awareness must be built into a company's culture at all levels, with processes and checks to guide data usage and catch bad uses.
- Practitioners must apply professional judgement and standards to data science work.
- Public trust in AI systems must be built.
- Accountability and oversight for AI systems must be maintained.

Discussion at the launch event for that paper in October 2019 further expanded on this, specifically the risk of bias and the risk that models can function at the expense of vulnerable consumers (e.g., optimising profitability by identify groups unlikely to protest against a price increase and setting premiums accordingly) (RSS, 2020).

IFoA Certificate in Data Science

In response to the growth of data science work within the actuarial field, the IFoA's Certificate in Data Science was introduced in June 2019 (IFoA, 2019a). The course consists of six modules, one of which is 'Good practice of Data Science, and responsible AI', covering the responsibilities of actuaries working in the field of data science. The course covers ethical and professional safeguards and the ethical issues which might arise from the participant's organisation's use of data science.

This course is not aimed at creating data scientists but at raising awareness and enabling actuaries to work more closely with data scientists. This data-science-specific ethical training, in addition to the existing professional standards and Actuaries' Code should position actuaries well to address the ethical issues discussed in this paper.

4.4 U.K. CENTRE FOR DATA ETHICS AND INNOVATION

The U.K. government established the Centre for Data Ethics and Innovation (CDEI) in 2018 as an independent advisory body, to assess the risks and benefits from data-driven technology (UK Government, 2018). Its reports include recommendations on data ethics for government, regulators and industry. Some particularly relevant publications regarding the ethical use of AI are outlined below.

Landscape Summary: Bias in Algorithmic Decision Making (ROVATSOS, MITTELSTADT, & KOENE, 2019)

The publication reports the following key findings:

- Fairness is subjective and machines cannot be expected to judge trade-offs, for example, between measures of group and individual fairness. Any strategies to minimise bias in AI systems must consider what is 'fair' in the particular context.
- Bias in algorithmic decision making is partially covered by the existing legislation of the Equality Act 2010 and the GDPR and Data Protection Act 2018 in the U.K.
- Identification of bias is hindered by both the lack of transparency and explainability of ML algorithms and how firms limit access to their proprietary data and systems. This is especially the case in financial services where systems are often made deliberately opaque due to their commercial sensitivity.
- Bias in algorithmic decision making can manifest in a variety of ways depending on the context, affecting a few consumers dramatically or many consumers to a lesser extent.

Snapshot Paper—AI and Personal Insurance (CDEI, 2019C)

In September 2019, the CDEI took an in-depth look at insurance (focusing on general insurance). This report stressed the potential benefits to policyholders from lower prices (by reducing fraud and back office costs), opening up insurance to groups previously deemed too risky,[74] offering personalised risk reduction advice[75] and making interactions with firms more seamless.

This is counterbalanced by potential risks from making certain individuals 'uninsurable' by hyper-personalised risk assessments, the intrusion of monitoring schemes and nudging,[76] and the collecting of large troves of data, of which customers may not even be aware.

The paper sets out a number of recommended actions to use AI responsibly:

- Engagement with the public to assess what is considered 'reasonable,' for example by using citizens' juries
- Internal measures such as a customer data charter setting out what happens with customer data and/or a data privacy advisory panel to consider the firm's use of data
- Use of data discrimination audits
- Review third-party data and software suppliers, for example to ensure any data received is accurate, unbiased and collected with the knowledge of the data subject
- Make data privacy notices more accessible to the public
- Comply with the provisions of the GDPR and the Data Protection Act 2018
- Give customers the power to port their data and risk profiles between insurance providers
- Establish clear lines of accountability

Call for Evidence—Summary of Responses: Review into Bias in Algorithmic Decision Making (CDEI, 2019A)

The CDEI performed a broader assessment of bias in algorithmic decision making, in which it summarised responses from a variety of organisations[77] on four question areas:

1. **The use of algorithmic tools**

A repeated theme was the need for a set of ethical principles to underpin approaches to decision-making algorithms. A number of organisations mentioned using either an internal framework, the European Commission Ethics Guidelines for Trustworthy AI (European Commission, 2019) or the Singaporean FEAT Principles (Monetary Authority of Singapore, 2018). Regardless of the precise approach, there was significant overlap between the principles organisations proposed, with fairness, transparency and explainability coming through as especially strong themes, as also concluded by our review in Appendix 2.

Many responses discussed the challenges of designing unbiased algorithms based on data which is itself inherently biased (unrepresentative sampling, historical patterns of biased behaviour, proxy characteristics correlated with protected characteristics[78]). There was disagreement over whether organisations should collect more data on protected characteristics to evaluate outcomes for potential bias, given the data minimisation principle of the GDPR and general concerns around data security.

2. **Bias identification and mitigation**

Responses noted that bias can enter algorithms at multiple points and while particular stages may be especially high risk, bias identification and mitigation should be pursued at every stage of algorithm development and use.

[74] For example, by using AI to assess alternative data sources, which can show that individuals pose less risk than existing profiling would suggest.

[75] For example, by using telematics data to provide feedback on driving behaviour.

[76] A nudge is a way of influencing behaviour in a predictable way by altering the environment around a choice, without forbidding any options or significantly changing their economic incentives.

[77] Academic institutions, think tanks, nonprofits, companies, industry bodies and public sector organisations.

[78] See Section 3.6 for more detail on proxy discrimination.

Approaches mentioned to mitigate bias included:

- Maintaining individuals within the decision-making process to provide oversight, while ensuring they have sufficient training to minimise their own biases
- Interrogating underlying data sets for potential biases
- Performing ongoing performance monitoring of models
- Ensuring that staff designing and using tools are properly trained
- Ensuring senior staff can make informed judgements and act as intelligent customers when procuring externally
- Use of individual tools to identify bias, complemented by broader processes

However, it was noted that given bias can never be removed entirely, there is no consensus about what standards models should be held to, and what best practice looks like. Others argued that a standard approach was unlikely to be achieved given the wide variety of contexts that AI is used in.

3. **Public engagement**

Responses discussed the difficulty of conducting meaningful public engagement in an area that so few people fully understand. Some respondents additionally drew attention to the trade-off between large-scale public engagement which may struggle to encompass the complexity of the issues involved, and more targeted approaches such as focus groups, which while more nuanced are less representative.

Respondents also raised issues around the trade-offs between explainability and accuracy,[79] and there was no clear view on where that balance should lie.

4. **Regulation and governance**

The financial services sector responses were particularly cautious about the possibility of new regulatory or governance measures, noting that the sector has extensive experience with data and is heavily regulated.

AI Barometer (CDEI, 2020)

The CDEI's first edition of the AI Barometer assessed the risks, barriers and opportunities from the U.K.'s adoption of AI technology, including a review of its use specifically in financial services. Highlighted risks include bias, greater exposure to cyberattacks (both from the volume of data held and the potential for new methods of attack[80]) and market concentration due to data monopolies.[81]

The report stresses that there is great potential for social benefit from AI but that there are a number of barriers preventing ethical AI from being achieved, including:

- Data availability and quality
- Limited evidence on what impact new innovations will have on society
- Public disagreement on acceptable use and trade-offs
- Scarcity of data skills in the workforce
- An unrepresentative workforce who may not consider the needs of other groups in society
- Funding gaps for projects tackling complex types of misuse[82]
- The risk that new developments may threaten existing profits
- Varying regulatory approaches across sectors and regulators, with unclear lines of oversight
- Lack of transparency about AI usage limiting accountability
- Lack of trust in AI meaning beneficial developments are not adopted

[79] Certain AI techniques increase model accuracy but at the cost of producing less interpretable results.

[80] See 'New Security Threats' in Section 3.9.

[81] Where a small number of firms have large high-quality data sets which give them a significant competitive advantage, such that new entrants are discouraged and the market becomes uncompetitive.

[82] For example, detecting *deepfakes* spread on social media.

This lack of trust is highlighted as a fundamental issue which all the other issues contribute towards. The CDEI concludes that these issues require coordinated action to address, so government support and a national strategy are necessary. When looking at the finance sector CDEI notes the power imbalance between consumers and firms, and suggest top down regulatory action will be needed to ensure ethical AI usage.

4.5 EUROPEAN COMMISSION

The U.K. left the EU on 31 January 2020 and negotiations are still ongoing as to the future EU-U.K. partnership—known colloquially as Brexit. However, developments on ethical AI from the EC that are relevant to all member states in the European Economic Area (EEA) will remain relevant to U.K. life insurers for a number of reasons.

Firstly, given the U.K. was in the EU for 47 years, its views on human rights and how these are reflected in law will currently be reasonably consistent. Guidelines published by the EC may therefore also reflect the cultural ideals of those of the U.K.

Secondly, there continues to be significant levels of cross-border business between the U.K. and the EU. In response to Brexit, and to ensure continued access the EEA single market, a number of U.K.-based life insurers transferred all of their business to an authorised EEA firm and established a U.K. branch for regulated activity in the U.K. Others restructured by moving all of their EEA business to an authorised EEA firm (e.g., a subsidiary) and continue to manage the U.K. business from the U.K. entity. Consequently, for life insurers which write business in both the U.K. and in Europe, such guidelines remain relevant.

We discuss the EU’s broad approach to AI and ethics then consider the specific ethical guidelines released by the EC’s AI High-Level Expert Group (AI HLEG)—‘Ethics Guidelines for Trustworthy AI’ (European Commission, 2019).

EU’S APPROACH TO AI AND ETHICS

The European Commission’s journey towards ethical AI began in earnest in May 2017 when it pitched for ‘A Connected Digital Single Market for All’ (European Commission, 2017). In that paper the EC highlighted the importance of building on Europe’s scientific and industrial strengths with a goal of becoming a leading player in the development of AI technology.

Later that year, the European Council tasked the Commission to present a ‘European approach to artificial intelligence’ by early 2018 that would aim to find a balance between the urgency of innovation in Europe (that was falling behind the U.S. and Asia in terms of AI investment and innovation) and ensuring a high level of data protection, digital rights and ethical standards (European Council, 2017).

In April 2018, 25 member states,[83] including the U.K., signed up to cooperate on AI (European Commission, 2018d), to work towards a comprehensive and integrated EU approach to AI that put ethics at the forefront of Europe’s AI offering to the world. The countries signed up to create a EU-wide legal and ethical framework rooted in and reflecting EU fundamental rights and values, including privacy and protection of personal data, as well as principles such as transparency and accountability. Of particular interest to ethics, those member states have agreed to the following:

- Exchange views on ethical and legal frameworks related to AI in order to ensure responsible AI deployment.
- Contribute to sustainability and trustworthiness of AI-based solutions, for instance by working towards improved information security, promoting safety and vigilance in the design and implantation, and increasing accountability of AI systems.
- Ensure that humans remain at the centre of the development, deployment and decision making of AI, prevent the harmful creation and use of AI applications, and advance public understanding of AI.
- Exchange views on the impact of AI on labour markets and discuss best practices on how to manage such impacts.

[83] Austria, Belgium, Bulgaria, Czech Republic, Denmark, Estonia, Finland, France, Germany, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Netherlands, Norway, Poland, Portugal, Slovakia, Slovenia, Spain, Sweden and the U.K. Since then, Romania, Greece, Cyprus and Croatia have also signed up.

This was followed by a communication (European Commission, 2018b) in the same month, that set out the EU's Three Pillar Approach, broadly: (1) further investment in AI; (2) preparing for socioeconomic changes that AI will bring about (e.g., retraining schemes for jobs at risk of being automated); and (3) ensuring an appropriate ethical and legal framework exists for AI.

Pillar 1 includes research in explainable AI and pillar 2 includes ensuring the labour market is sufficiently skilled to understand and use AI, which both contribute to the ethics debate. Of particular relevance is the final pillar, with the EC setting the goal of developing draft AI ethics guidelines by the end of 2018.

The framework was to be based on the EU's values, in particular those set out in Article 2[84] of the 'Treaty on the Functioning of the EU' (European Union, 2016b) and in line with the 'Charter of Fundamental Rights of the EU' (European Union, 2016a). That is, the EC's framework must be founded on the human, economic and social rights and values enjoyed by those living in the EU and aim to ensure that AI systems are developed in a manner that allows humans to understand the basis of their actions in order to strengthen public trust, increase transparency and minimise the risk of bias or error from AI systems.

In December 2018, the EC published a communication (European Commission, 2018c) where it invited the European Council to endorse a coordinated plan, rooted in its Three Pillar Approach, to be implemented by all member states. An integral part of this plan is the development of its own ethical guidelines, the 'Ethics Guidelines for Trustworthy AI' covered below.

In the Annex (European Commission, 2018a) to that communication, the EC set out its aim of 'implementing, on the basis of expert work, clear ethics guidelines for the development and the use of AI in full respect of fundamental rights, with a view to set global ethical standards and be a world leader in ethical, trusted AI.' The EC also stated that:

- It will explore how to introduce an ethics by design principle, whereby ethical and legal principles are to be implemented from the beginning of the design process for AI systems, in any calls for proposals under its research programme.
- It believes that ethics and other nontechnical skills are equally important in the process of developing AI professionals.
- The EU will promote the AI ethics guidelines internationally, inviting discussions and cooperation from any non-EU countries and stakeholders from third countries that are willing to share its values.

Ethics Guidelines for Trustworthy AI

The EC appointed the independent AI HLEG to produce ethical guidelines by the end of 2018, with consultation through the European AI Alliance[85] until March 2019. The guidelines were published by the AI HLEG in April 2019 following an open consultation period that involved over 500 contributors (European Commission, 2019).

For trustworthy AI, AI HLEG state that the system must adhere to three core components, namely that the AI system must comply with the following:

- It must be **lawful**, abiding by all applicable laws and regulations, in particular:
  - **EU primary law:** Developed in line the Treaties of the European Union and its Charter of Fundamental Rights.
  - **EU secondary law:** Such as the GDPR, the Product Liability Directive, the Regulation on the Free Flow of Non-Personal Data, antidiscrimination directives, consumer law and Safety and Health at Work Directives.

[84] This article states that 'The Union is founded on the values of respect for human dignity, freedom, democracy, equality, the rule of law and respect for human rights, including the rights of persons belonging to minorities. These values are common to the member states in a society in which pluralism, nondiscrimination, tolerance, justice, solidarity and equality between women and men prevail.'

[85] A multistakeholder forum that are engaged in a broad and open discussion on all aspects of AI development, and its impacts.

  - **United Nations human rights treaties and the Council of Europe conventions:** Such as the European Convention on Human Rights and numerous EU member state laws.
  - **Domain specific rules:** In the case considered in this paper, those laws and regulations applicable to financial services, or more specifically, the life insurance sector in the U.K.

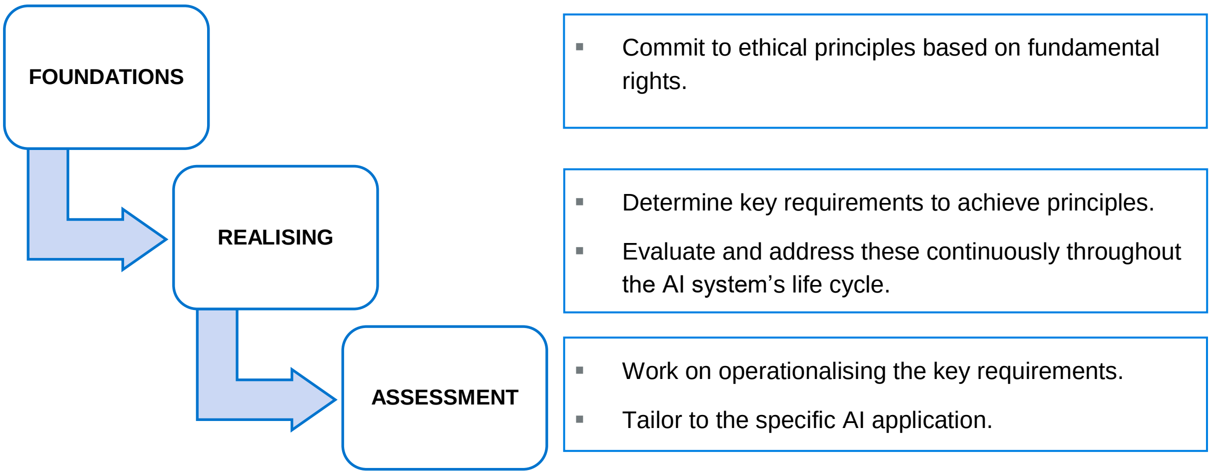
- It must be **ethical**, abiding by ethical principles and values. In part, AI HLEG presented this as a result of laws not always keeping pace with technological developments, resulting in situations where they are out of touch with ethical norms, or not well suited to addressing ethical issues.
- It must be **robust,** both in a robust technical sense but also socially robust, with AI HLEG identifying that even an AI system which has 'good intentions,' that is they are designed to be both lawful and ethical, can cause unintentional harm when considering the context and environment in which they operate. Thus, AI HLEG believes systems must be developed that perform in a 'safe, secure and reliable manner' with sufficient safeguards and monitoring then put in place in order to identify or prevent any unintended harm to individuals or society as a whole.

The framework itself focuses on how to ensure AI systems can be designed to be ethical and robust.

A key concept that the AI HLEG Guidelines promotes is the concept of the '**entire life cycle**' of the AI system.[86] AI HLEG acknowledges that certain AI systems will be continuously learning in response to new data and thus developing, and so it is not sufficient to simply ensure the AI is lawful, ethical and robust at its launch. Instead, an AI system must be continuously monitored for compliance throughout its entire life cycle: development, deployment and whilst in use.

AI HLEG's proposed framework is broken into three main sections, moving from broad, high-level and, in AI HLEG's own admittance, more 'abstract' principles for achieving trustworthy AI (the 'Foundations of Trustworthy AI'), through methods to realise those principles in practice (the 'Implementation and Realisation of Trustworthy AI') and finally how to operationalise those techniques and ensure that they become business-as-usual practices (the 'Assessment of Trustworthy AI'). See Figure 16:

FIGURE 16: AI HLEG FRAMEWORK FOR TRUSTWORTHY AI



FOUNDATIONS

- Commit to ethical principles based on fundamental rights.

REALISING

- Determine key requirements to achieve principles.
- Evaluate and address these continuously throughout the AI system's life cycle.

ASSESSMENT

- Work on operationalising the key requirements.
- Tailor to the specific AI application.

A detailed overview of each of these sections can be found in Appendix 3.

To move these concepts into a practical framework that companies can implement to embed ethics in their AI systems, the AI HLEG Guidelines present a non-exhaustive assessment list, broken down into its seven key areas. It also provides guidance on how firms could work the list into existing governance structures, or use it in newly created governance structures.

[86] See Section 2.6 for details about the AI system life cycle.

An initial list published in the AI HLEG Guidelines in April 2019 was subject to a piloting period involving more than 350 stakeholders, with a final document, and accompanying web-based tool, released in July 2020—dubbed the ‘Assessment List for Trustworthy AI’ (ALTAI) (AI HLEG, 2020), summarised in Appendix 3. This ALTAI is clearer and more focussed than the draft assessment list included in the AI HLEG Guidelines but captures the majority of its key considerations. The optional accompanying web-based assessment tool allows firms to input responses to the assessment list’s questions, and then the tool presents results as a spider-web chart to show how well firms meet each of the seven requirements. The tool also provides recommended actions, turning the list into an ongoing gap analysis.

The ALTAI may be useful for insurers seeking to determine what best practice looks like in this rapidly evolving field. However, as the ALTAI is self-assessment, there may be material differences in the interpretation of questions and how critical firms are towards their own practices.

Information provided by those filling out the tool should allow the AI HLEG to understand if there is harmonisation in the approach to AI ethics across Europe; this will be limited by the information requested, which does not cover the types of AI used (e.g., neural networks, NLP), the use case for the AI system, or the specific industries or sectors the AI system is to be designed for or deployed.

A number of the key themes in the ALTAI should not be new to U.K. life insurers that are part of a heavily regulated industry; putting in place strong governance (including resolution planning), determining risk appetite and levels for ethical risks, having robust models and systems that take into consideration data privacy and security, and the consistent monitoring of systems, are all areas where insurers have experience, making them well placed for compliance when compared with less heavily regulated industries.

Unlike some other industries, it may therefore be a case of ‘evolution rather than revolution’ for the insurance industry; instead of putting in place new mechanisms, processes and procedures, the focus will be on extending existing governance frameworks, policies and procedures, and processes to the ethical use of AI.

4.6 INDEPENDENT OF GOVERNMENT, REGULATORS AND INDUSTRY

There are a number of other bodies that are independent of government, regulators and industry, but that debate, study and research matters related to ethics in AI, each with varying focus. The below examples are some of the organisations whose work we have benefited from when writing this paper.

THE ALAN TURING INSTITUTE

A U.K. institute founded in 2015 that focuses on undertaking cutting-edge research in data science and AI, and applying this research to real-world problems. They have a broad selection of research areas spanning mathematical methods, computer systems and AI, but also have a research programme specifically on finance and economics, assessing trustworthy models and responsible AI use in financial services. The Alan Turing Institute has worked alongside the ICO to develop guidance in explaining decisions made by AI systems.

THE ADA LOVELACE INSTITUTE

This independent research institute established in 2018 assesses the social impact of AI and promotes public understanding. It gathers evidence of good and bad practice, runs public engagement work[87] and suggests regulatory changes to bring about socially beneficial use of AI.

The JUST AI (Joining Up Society and Technology in AI) network was established in 2020 as a partnership between the Arts and Humanities Research Council (AHRC) and the Ada Lovelace Institute. It aims to build a multidisciplinary research base in ethical AI, centred on practical issues of social justice (Arts and Humanities Research Council, 2020).

[87] For example, the institute recently ran online deliberation sessions with members of the public around how to build public trust in a coronavirus contact tracing app (Ada Lovelace Institute, 2020b).

4.7 INTERNATIONAL CONSIDERATIONS

We have focussed in this section on some of the bodies contributing to the ongoing debate on ethics and AI in the U.K. Naturally, there may be global variation in perspectives regarding ethics in AI. We consider some factors that influence the approach to AI for the U.S., China and the EU (i.e., the largest groups in terms of AI investment and research), in particular:

- National strategy
- Level and nature of investment in AI
- Financial and local regulation

NATIONAL STRATEGY

The U.S.'s national strategy towards AI was set out in 'Artificial Intelligence for the American People' (The White House, 2018) and stresses the U.S.'s core values of 'freedom, guarantees of human rights, the rule of law, stability in our institutions, rights to privacy, respect for intellectual property and opportunities to all to pursue their dreams.' They acknowledge the need for trustworthy AI but stress that this must be balanced against the risk of stifling innovation, arguing that regulation should be implemented once technology has matured (The White House Office of Science and Technology Policy, 2018). Government guidance has been issued stating that government agencies must assess the impact of any restrictions on innovation and only consider regulation after assessing that it is truly necessary (The White House's Office of Science and Technology Policy, 2019).

As discussed in Section 4.5, the EU has a coordinated plan on AI, to be enacted by member states through national initiatives. The EU has identified ethics as its competitive edge in developing AI (European Commission, 2018b) and intends to lead the global debate on ethics. The EU's strategy of 'AI made in Europe' states it has the key principle of 'ethics by design,' where ethical and legal principles are implemented in AI systems from the beginning of the design process (European Commission, 2018a).

China's 2017 AI strategy was to be one of the world's most advanced countries in AI by 2020, world leading in some AI technologies by 2025, and the primary innovation centre by 2030 (Chinese State Council, 2017). Establishing AI laws, regulations, ethical norms and guidelines is part of step 2, to be developed by 2025.

The three show differing priorities in terms of the balance between innovation and ethical constraints in the development of AI technology, and also with regards to the time frame for ethics to become the greater priority.

AI INVESTMENT IN RESEARCH AND DEVELOPMENT

To deliver their AI strategy, the U.S. government is planning a significant increase in AI investment, with $4.9 billion allocated to unclassified artificial intelligence and machine learning–related research and development in the 2020 fiscal year (Bloomberg, 2019). However, of this, $4 billion is defence related, while $850 million is for civilian government agencies.

In contrast, spending on AI by the Chinese government is estimated to be skewed to nonmilitary investments. The Center for Security and Emerging Technology estimates China's civilian AI research and development investment at $1.7 billion to $5.7 billion in 2018, compared with $0.3 billion to $2.7 billion spent on military research and development[88] (Center for Security and Emerging Technology, 2019).

In the EU, there is very little focus on the use of AI for the military (Franke, 2019). The European Defence Fund's Preparatory Action on Defence Research (PADR) allocates up to 8% of its budget to disruptive technologies which could include AI (European Commission, 2020a). This budget was around $106 million for 2017 to 2019.[89] This is low when compared to the total EU investment in AI for 2020 of $1.8 billion[90] over the similar period (European Commission, 2018b).

The research funding level and the differing emphasis on military versus domestic funding will presumably influence the prism through which ethics is viewed.

[88] It is more challenging to get a clear view on budgeted spend in China for AI research. The study by the Center for Security and Emerging Technology provides estimates of the upper and lower bounds of China's AI investments for 2018 by cross-referencing its ministry of finance's national expenditure report with other government agencies' funding calls and abstracts from researchers describing agency-funded projects.

[89] €90 million, converted to U.S. dollars using the exchange rate as of 7 October 2020.

[90] €1.5 billion, converted to U.S. dollars using the exchange rate as at 7 October 2020.

REGULATION

The insurance sector is heavily regulated, so the prioritisation of innovation, over preemptive regulation for the use of AI, in the U.S. more generally may not translate to the insurance industry. In the U.S., the National Association of Insurance Commissioners[91] has created an Artificial Intelligence Working Group, which this year adopted a set of principles to guide firms on expectations around appropriate use of AI (National Association of Insurance Commissioners, 2020). These are the familiar ethical principles seen in global standards that AI should be: fair and ethical, accountable, compliant, transparent, and secure, safe and robust. The Consumer Financial Protection Bureau has issued three policies outlining when the regulator will not take supervisory or enforcement action, such as when a firm is testing a service where there is regulatory uncertainty (The Consumer Financial Protection Bureau, 2019).

China's leading finance think tank, the China Finance 40 Forum, has called for a regulatory framework for AI usage and more technology use by regulators, noting that technology has advanced faster than regulation (Reuters, 2019).

In Europe, the European Insurance and Occupational Pensions Authority (EIOPA) has an expert group on digital ethics in insurance working to adapt the European Commission's guidance to fit the insurance industry (Bernardino, 2020).

LOCAL REGULATION

In the U.S. there is significant state-level regulation, varying from guidance that unconventional external consumer data may be used in life insurance underwriting if done transparently and shown to be nondiscriminatory (New York State Department of Financial Services, 2019) to Washington state's proposed privacy act aligned with the EU's GDPR (Future of Privacy Forum, 2020).

This creates additional complications around how and when insurers can apply particular aspects of AI for those who operate in multiple U.S. states.

SUMMARY

Despite these differences and worries about foreign technology developed with different ethical standards, as discussed in this report, there is an emerging set of broad themes among AI ethics guidelines and global convergence to a set of principles and norms.

For example, the Beijing Principles[92] (The Beijing Academy of Artificial Intelligence, 2019) promote AI which respects privacy, dignity, freedom, autonomy and human rights. The similarity to principles developed by Western organisations was noted as a positive sign by the World Economic Forum (MIT Technology Review, 2019).

FIGURE 17: INTERNATIONAL AGREEMENTS

There have been a number of international agreements on ethics in AI, aligning countries with shared ethical standards.

- **G7:** In 2018, the G7 endorsed a 'Common vision for the future of artificial intelligence (AI),' committing to promote human-centric AI (G7, 2018).
- **OECD:** In May 2019**,** OECD member countries approved the OECD Council Recommendation on Artificial Intelligence, expressing the goal of taking a 'human-centred' approach to AI (OECD, 2019).
- **G20:** In July 2019, the G20 agreed principles for responsible stewardship of 'trustworthy' AI, stressing human centred values, transparency, explainability and accountability (G20, 2019).
- **The United Nations Educational, Scientific and Cultural Organisation (UNESCO):** A draft text of a Recommendation on the Ethics of Artificial Intelligence (UNESCO, 2020) has been produced. There has been a global consultation on this draft, with a final recommendation due to be submitted for adoption in November 2021.

International convergence is helped along by international efforts where countries have pledged their support for ethical approaches to AI, see Figure 17.

[91] An organisation created and governed by the chief insurance regulators from across the U.S. states and territories, through which state insurance regulators establish standards and best practice, conduct peer review and co-ordinate regulatory oversight.

[92] A set of guidelines developed by the Beijing Academy of Artificial Intelligence (BAAI) as a partnership between China's Ministry of Science and Technology (MOST) and leading Chinese universities and technology firms.

4.8 KEY TAKEAWAYS

- U.K. insurers are, or will be, subject to AI regulation in terms of prudential regulation concerns (PRA), conduct regulatory concerns (FCA), and data privacy and security regulatory concerns (GDPR and the ICO).
- There is no current U.K. financial regulation covering AI, although existing regulation is relevant irrespective of whether AI is used. GDPR and ICO guidance have been created with AI in mind and contain provisions directly related to AI usage.
- For AI, the PRA prioritises transparency and governance and has issued firms with three challenges and principles on the governance of data and AI models, the role of people in AI systems, and execution risks when transitioning to an AI system.
- To date, the FCA has not published any principles, guidance, regulation and/or industry-specific best practice on AI for U.K. life insurers, though it is exploring whether any should be provided.
- The GDPR imposes restrictions on when individuals' personal data may be used and how, an important consideration given the volume of data used in training AI. Certain conditions must be met before data collected for one purpose can be reused for other purposes.
- The GDPR recognises individual's rights over their data. Firms may be asked to explain how they process data, remove or update individuals' data in their AI models on request, and provide people with data concerning them. Where processing is not necessary, individuals have a right not to be subject to an automated decision-making system.
- Actuaries working to create models that use AI must also comply with certain professional standards such as the Actuaries' Code, Actuarial Professional Standards and Technical Actuarial Standards. These documents should be consulted to consider their implications regarding the development of AI systems. Although the professional guidance does not specifically cover AI, the general principles, especially around modelling and judgement, remain relevant.
- Insurers in the U.K. may also find guidance on AI, and insight into proposed AI regulation, from the work of the CDEI, Alan Turing Institute, Ada Lovelace Institute, and others.

5. Concluding remarks

Life insurers that are using, or looking to use, AI in their business may be looking internally and externally at how to approach the questions raised by ethics. Internally, they may be reviewing existing policies and procedures, governance and controls, and training and talent acquisition to identify whether changes are needed in order to embed a risk-based[93] ethical approach to AI into the business. Externally, they may be seeking the views of experts in academia, regulatory bodies, civic bodies and the public to ensure that they are developing AI systems that are lawful, ethical and in line with public expectations. They will also be looking to shape public expectations through education and engagement through public debate.

One of the greatest challenges identified as part of this research is the breadth of the topic; in this report we have merely provided a flavour on the ongoing 'ethics and AI' debate, with a focus on the life insurance sector. What is apparent from our work is that although an abundance of holistic guidelines aim to provide a framework to govern all use cases of AI in all industries, such guidelines may not necessarily be practical in providing industry or use case specific insight that would be useful for those working in the life insurance industry to determine what best practice looks like for the particular problem at hand. Indeed, even in the cases where guidelines provided some technical methods to achieve ethical AI, rather than simply a philosophical mantra, those solutions again are not use-case specific. Instead, focusing on a single issue would be an interesting, and fruitful, avenue for further research. For example, (Lindholm, Richman, Tsanakas, & Wüthrich, 2020) have focussed on aiming to answer a single problem—how can we ensure that pricing in insurance is free from discrimination when using AI?

There is a clear indication that regulators in, and outside, the U.K. are taking a keener interest in how regulated entities are using, or considering using, AI in their business. Given the direction of travel of influential bodies for life insurers that we reviewed as part of this report appear to be moving (e.g., the EC, national finance and conduct regulators, data regulators, and professional bodies), we expect to see further regulatory guidance and potential rules and requirements on the use of AI in the life insurance sector in the coming years. Until then, further research that preempts such guidance and that presents a view on what best practice would look like, for specific use cases of AI in insurance, may be of value to the industry.

Our research into this field has shown that the ongoing ethical debate can be rooted in human rights legislation,[94] influenced by past legislation and expectations from regulators,[95] and heavily political with countries at a national level weighing up the extent to which it balances innovation with ethics. This is most clearly seen with the EU setting out ethics as its key differentiator in the global competition for dominance in AI technology. Owing to this, and also informed by other macro-factors such as culture, religion and tradition, there could be quite a distinctly different view on what is morally right and wrong on the use of AI between countries. An interesting avenue would be a survey of the public on ethics and AI use in the life insurance sector. The focus of the survey would be to determine whether there are geographical differences in the public's attitude to ethics, such as questions on privacy, what it means to be fair and how transparent firms should be, focusing on current and potential life insurance use cases. Such considerations would be useful for life insurers that are operating in multiple markets, to provide valuable insight of where concerns may, and may not, lie, that could further guide strategic decision making regarding AI implementation.

Most of the guidelines, advice and concerns on ethics and AI that we observed as part of this research focussed predominantly on the preservation of current ethical norms and requirements. This is likely reflecting that such technologies are seen as relatively new, so currently the focus is on ensuring firms put in place appropriate risk management and governance, building AI systems with ethics in mind—'ethics by design.' As AI systems become more commonplace, we may instead see the ethics conversation move more in favour of the ethical benefits AI can create. In life insurance, such benefits may include creating greater fairness in underwriting by deriving insight from alternative data sources that allow insurers to accept risks that would previously have been deemed 'too risky,' or to provide greater transparency on an individual's risk factors and what they can do to get an improved premium or underwriting decision. On the latter example, although insurers can already notify policyholders on the factors they use to determine their premiums and underwriting decision in cases where AI has not been used, the increased transparency and explainability requirements for AI systems may make this a requirement for decisions made by an AI system.

[93] Focussing on those applications that have the greatest potential for harm to consumers, or to the financial security of the firm.

[94] For example, the European Commission's AI HLEG.

[95] Such as the data regulation, conduct of business regulation, financial regulation and antidiscrimination legislation.

Finally, one of the challenges faced in drafting this report was getting a clear view on the current state of play of AI in the life insurance sector. The 2019 BoE and FCA Study mentioned in this report provides an incredibly useful snapshot of the use of machine learning in the U.K. financial services sector. However, due to the broad scope of that review, which includes life insurers, general insurers, banks and other financial service providers, the survey does not allow for more nuanced industry-specific, or use-case-specific, questions on the life insurance sector specifically. A similar survey that focuses on surveying life insurers in the U.K. and further afield, would provide valuable insights on how AI is being used, where it is being used, and the controls and governance structures considered appropriate by industry participants for various use cases. This would provide firms with the ability to benchmark their current practices against their peers, allowing for greater convergence on best practices in the industry.

Appendix 1—Machine learning examples

In this Appendix we provide an illustration of the application of some machine learning techniques to the categorisation of 4,601 electronic messages between true emails and spam.

For each message, 57 characteristics are available:

- 48 measures of the percentages of the words in the message that match a given word (such as ‘remove’ or ‘free’)
- 6 measures of the percentage of characters in the message that match a given character (such as ‘$’ or ‘!’)
- The average length of continuous sequences of capital letters in the message
- The length of the longest continuous sequence of capital letters
- The sum of the length of the continuous sequences of capital letters

We outline the application of a (binary) decision tree, a random forest, adaptive boosting and a neural network to this example.[96]

In practice the best performing technique depends on the nature of the data to which it is applied.

A (BINARY) DECISION TREE

Of 4,601 messages, 1,536 (33%) were randomly selected to be the test set so that 3,065 formed the data to be used to train the model.

The top (**root**) node of the tree starts with all of the training data. The randomness (entropy[97]) of the data at the node is measured, and then that data is split between two child nodes so that the weighted average[98] of the entropy at the two child nodes is as low as possible[99].

A key requirement for the entropy measure is that its value at the parent node will be greater than or equal to the weighted average of its value at the two child nodes[100].

To decide how to split the training messages at the root node, we consider in turn each of 57 characteristics available for each message that may be a good indicator of whether it is email or spam.

For example, if the first characteristic is the percentage of characters in the message that are ‘$’ ($f_{‘\$’}$), and the values in the training data are 0.1%, 0.2%, 0.3% and 0.4%,[101] then the entropy would be calculated for the two child nodes created by each of the splits $f_{‘\$’}$<0.2%, $f_{‘\$’}$<0.3% and $f_{‘\$’}$<0.4%. A split $\mathbf{s_1}$ that produces the lowest entropy for the children would be determined. In the example considered here $\mathbf{s_1}$=0.0555.

Of the 57 splits $\mathbf{s_1}$ to $\mathbf{s_{57}}$ determined in this manner, suppose $\mathbf{s_1}$ is one that produces the smallest entropy. Then each of the 3,065 training messages at the root node are split between two child nodes depending on whether the percentage of characters in the message that are ‘$’ is higher or lower than s_1=0.0555.

Then the process is repeated for each of those child nodes, with all 57 characteristics being considered once again.

[96] These application are synthesised from (Hastie, Tibshirani, & Friedman, 2009).

[97] There are a number of ways of measuring entropy.

[98] The weight applied to a child node is the proportion of the messages in the parent node that are allocated to that child.

[99] This means the information gain from the split is maximised. Information gain is the entropy at the parent minus the weighted average of the entropy of the two child nodes.

[100] The spam example illustrated in this appendix uses the cross-entropy (deviance) methodology. For example, the 3,065 training messages comprise 1,852 emails and 1,213 spam messages. The entropy (randomness) at the root node of the training data is calculated as 0.928 = [-p x $\log_2$p—(1-p) x $\log_2$(1-p)] where p =1,852/3,065 = 0.604, the proportion of emails in the training messages at that node.

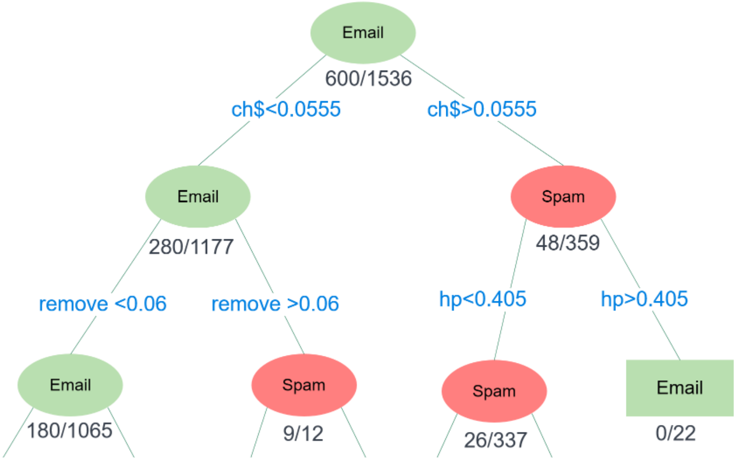
[101] In practice there will be far more variance than this. Assuming that all the training messages have ‘$’ appearing with a frequency of 0.1%, 0.2%, 0.3% or 0.4% is just a simplification for explanatory purposes.

The size of the tree is determined at outset. The preferred strategy is that each node is split until it contains a small number of training data points (messages in this example); minimum size could be 5 or 1. The large tree that results is then pruned to reduce the overfitting to the training data[102] that would be likely to result from using the whole tree.[103]

Terminal (leaf) nodes have no children and are labelled email or spam according to whether the majority of the training emails that reach that node are emails (majority voting).

The following diagram shows the first few splits of the fully built tree (after pruning), and the results of applying the 1,536 test messages to it.

FIGURE 18: A BINARY DECISION TREE FOR SPAM FILTERING—BASED ON (HASTIE, TIBSHIRANI, & FRIEDMAN, 2009)



The root node is labelled ‘email’ as only 600 of the 1,536 test messages are spam.

The most effective initial split of the training data was whether the frequency of the character ‘$’ in a message is greater than 0.0555. So this criterion is applied first to each test message.

The leftmost node in the second layer indicates that of 1,117 test messages where the frequency of ‘$’ is less than 0.555, the majority are emails (so the node is labelled ‘email’) and 280 are actually spam.

The rightmost node in the second layer indicates that of 359 test messages where the occurrence percentage of ‘$’ is greater than 0.0555, the majority are spam, and 48 are actually emails.

For the 359 test messages that reach the rightmost node in the second layer of the tree, the next characteristic measured is the occurrence percentage with which ‘hp’ appears. This results in 22 test messages being allocated a final categorisation of ‘email’ as the rightmost node in the third layer is a terminal node. In this case this classification is correct: all 22 of the test messages reaching this node are emails.

The 337 test messages with $f_{‘\$’}$>0.0555 and $f_{‘hp’}$<0.405 are subject to further processing through the tree until they reach a terminal node, which provides their classification (email or spam).

The attraction of a decision tree is its relative simplicity.

A major drawback is its high variance: A small change in training data may result in a very different series of splits. The hierarchical nature of the process is the major reason for this: An error in a split is propagated to all of the splits below it.

[102] Overfitting is where a model captures the features of the training data too closely and performs poorly when applied to other data.

[103] The example described here was pruned using the *misclassification rate*, which is the percentage of false positives plus the percentage of false negatives.

RANDOM FORESTS

A random forest is an extension of the decision tree approach. It comprises a large number of de-correlated decision trees. Each tree has an equal say: The result is determined by majority voting when categorisation is involved, or else by averaging the outcome across all trees in the forest.

Each tree is built on a random sample N of the training data;[104] each sample being drawn from the whole of that data (i.e., sampling with replacement). At each node of a tree a random sample of m of the possible data characteristics[105] is used to split data at that node.

The random sampling used to generate individual trees reduces the variance of the prediction from the random forest, overcoming one of the disadvantages of decision trees.

Interpreting the final model is very difficult as it contains many independent decision trees.

ADAPTIVE BOOSTING

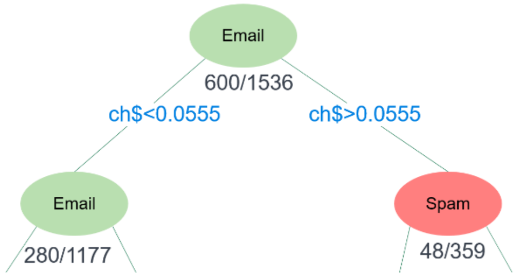
Adaptive boosting uses a sequence of tree stumps, each comprising only a parent and two child nodes. Each stump aims to improve the predictive power of the previous one in the sequence by placing greater weight on the training data misclassified by the previous stump.

The prediction of each stump is assigned a credibility factor, and the prediction of the model is determined as the sum of the credibility weighted prediction of all the stumps.[106]

The first tree stump is built in the manner described above for determining the first split of a binary decision tree from the root node. At this stage each training data point has equal weighting of 1/N, where N is the number of training data points.

For illustrative purposes, we assume the first stump is as follows.[107] That is, we assume the smallest entropy that can be produced by splitting the training data is 0.7392 (this is the entropy of the stump shown[108]), and this is achieved by separating messages depending on whether the frequency with which '$' appears in the message is higher or lower than 0.5555.

FIGURE 19: ADAPTIVE BOOSTING—AN ILLUSTRATIVE EXAMPLE



For illustrative purposes, the credibility of stumps is calculated as $\log_{10}$ [(1- weighted error rate)/weighted error rate].[109]

If the stump performs better than random guessing (i.e., the error rate is lower than 50%), then the credibility will be positive, it will zero if the error rate equals 50% and otherwise it will be negative.[110]

[104] In the spam example, N will be less than 3,065.

[105] In the spam example, m will be much less than 57.

[106] Credibility-weighted majority voting for a categorisation example such as the spam identification.

[107] This stump actually shows the effect of applying the first step of a fully grown decision tree for the spam example to testing data, but nevertheless it can be used to illustrate the process.

[108] There are 1,536 training messages, so the initial weighting for each data point is 1/1536. The probability p of a message being email in the left node is the sum of the weights of the email that reach that node divided by the total weights of the messages that reach that node: (897/1536) / (1,177/1,536) = 0.7621. The entropy of the left node is therefore 0.7915 = [$-p\log_2(p) -(1-p)\log_2(1-p)$]. The entropy of the left node is 0.5675. So the weighted average of the two child nodes is 0.7392 = [1,177/1,536x 0.7915 + 0.5675 x 359/1,536].

[109] There are a number of ways of measuring credibility.

[110] If the weighted error rate is zero (all messages correctly classified), then the credibility factor is taken as +1. If the weighted error rate is 100%, then the credibility factor is taken as -1.

For the above stump the credibility weighting is 0.5662.[111] So it is better than random guessing but a weak model.

The next stump is determined using the same testing data points after increasing the weightings of the 328 (=280+48) misclassified points by multiplying by 3.68 (=(1- weighted error rate)/error rate)).

Using these revised data weightings, the entropy of the above stump increases from 0.7392 to 0.9860.

The next stump will be a split of the training data that produces the lowest entropy using the revised weightings, and so on.

One method of determining how many stumps to include is **random search cross validation**. For this a minimum and maximum is set, together with the number of stumps to be randomly selected within this range. For example, suppose a minimum of 3 and a maximum of 30 stumps is set, and 6 random selections are 4, 8, 14, 18, 24 and 28. Models with each of these number of stumps are built, and the selected model is that which produces the lowest **k-fold validation error** for some selected k.[112]

Tree stumps are easier to interpret than full decision trees; however, interpreting the model is difficult because of the weighting of each stump.

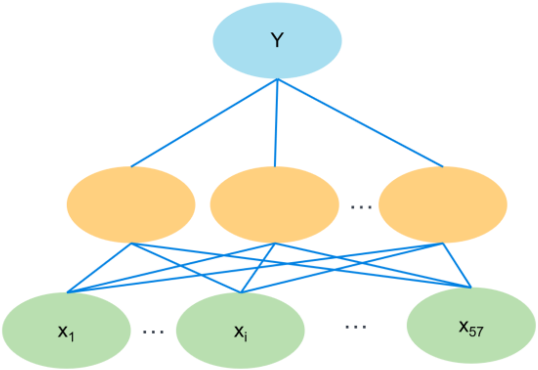
NEURAL NETWORKS

A neural network comprises:

- A set of input neurons (the data items, shown in green in Figure 20)
- A set of output neurons (only one for the message filtering example, which indicates whether a message is predicted to be email or spam, shown in blue in Figure 20)
- Hidden neurons which connect the input neurons to the output neurons (shown in orange in Figure 20)

The number of layers of hidden neurons, and the number of neurons in each layer, are specified by the designer of the network. The user of the model does not interact with the hidden neurons.

FIGURE 20: NEURAL NETWORK—AN ILLUSTRATIVE EXAMPLE



In the message filtering example, the data items (labelled x_i) are the 57 characteristics known for each message and the network comprises one layer of 10 hidden neurons.

The value of the output neuron labelled Y would lie between 0 (email) and 1 (spam), with the prediction being whichever of 0 and 1 the output is nearer.

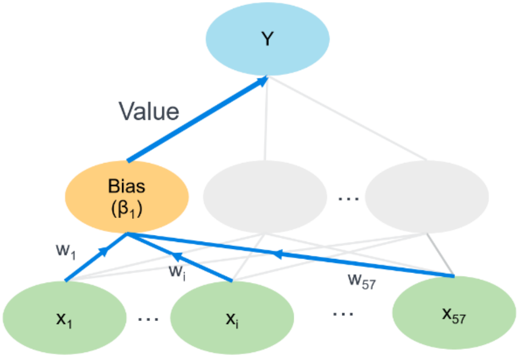
The following diagram illustrates the typical operation of each hidden neuron in a feedforward neural network (i.e., there is no looping back):[113]

[111] The stump classifies 1,536 messages (=1,177+359). Initially the weighting applied to each training data point is 1/1,536. So the weighted error rate is [280/1,536 + 48/1,536] = 0.2135, and the credibility applied to this stump is $\log_{10}$ [(1-0.2135)/0.2135] = 0.5662.

[112] K-fold cross validation involves splitting the training data into k subsets (folds), and iterating through the folds taking one of the k subsets as the test set and the rest of the subsets as the training set, and repeating until all of the folds have been used as a test set. The error rate is determined after each iteration, and the average of the error rates across the k iterations is the k-fold cross-validation error of the tree.

[113] Only three of the 57 inputs are shown here.

FIGURE 21: TYPICAL OPERATION OF EACH HIDDEN NEURON IN A FEEDFORWARD NEURAL NETWORK



In the message filtering example, each of the 57 inputs to the ith hidden neuron is multiplied by a weight $w_{i,1}$, $w_{i,2}$, … $w_{i,57}$. Then the weighted inputs and a bias β_i (specific to the hidden node) are combined using an **activation function** to determine a value for the node to be passed to the output neuron.[114]

If the weighted sum of the inputs plus the bias is large and positive, then the output value determined at the hidden neuron will be close to 1, and if the weighted sum of the inputs and the bias is large and negative, then that value will be close to zero.

The output neurons have exactly the same structure as the hidden neurons, except that the number of inputs depends on the number of hidden neurons that feed into each output neuron.

In the message filtering example there is one output node, and this has 10 inputs (one from each of the nodes in the single hidden layer).

Initially the weights and biases of the hidden and output nodes are determined randomly. Their values are then changed through the training process.

This involves running each of the training messages through the network.

A training message would be represented, for example, as (0.04, 0.01, 0.001, ….) where each of the 57 elements of the vector represents the occurrence percentage of each characteristic. For example, 4% of characters in the message are '$', 1% are 'hp' and so on.

Suppose the training message is an email, and using the (randomly generated) starting weights and biases the output value (the prediction) for it is 0.4. Prediction error is measured as the sum of the squares of the differences between the prediction and the actual classification of the message: (0.4-0)^2 = 0.16 in this case.

The fit *F* of the model is the total prediction error summed over all the training data.

The aim of training the network is to make small changes to the weights and biases to improve the fit of the model.

One way to incrementally reduce *F* (i.e., improve the fit) is to use **gradient descent**. This involves making a small increase to just one weight or bias (call it w_1), leaving all the others unchanged, and rerunning all 3,065 training messages through the network to calculate the resulting change in *F.*[115] If the result is a decrease in *F* of 1%, say, then shortly w_1 will be increased by 1% of η (a constant known as the **learning rate**). If the result is an increase of *F* by 0.5%, say, then shortly w_1 will be decreased by 0.5% of η.

Now w_1 is set back to its original value and a change is made to another weight or bias (call this w_2), leaving all others unchanged. Suppose that after increasing w_2 and rerunning all 3,065 training messages through the network *F* increases by 0.2%. Then shortly w_2 will be decreased by 0.2% of η.

Once the manner in which each and every weight and bias should be changed has been determined, all the changes are made in one go and *F* is recalculated; call its new value F_1 which will be smaller than *F*. That completes the network's first lesson.

Now the lesson is repeated using F_1 instead of *F* to determine whether to increase or decrease each of the new weights and biases, and to determine the size of the changes (the method of determining the size of the change

[114] A commonly used activation function is the sigmoid function.

[115] In effect the first derivative of *F* with respect to w_1 is being estimated.

does not differ, and nor does η). Then all the changes are made in one go to calculate F_2 (which will be smaller than F_1). The next lesson uses F_2 to calculate F_3 and so on.

Where the training data set is large a less time-consuming approaches can be used, such as **stochastic gradient descent.**

This involves taking a small, randomly chosen, subset of the messages (known as a **mini batch**) and carrying out gradient descent on this subset. Then gradient descent is applied to the next mini batch and so on. When the training messages have been exhausted by mini-batches (lessons), an '**epoch**' of training is said to have been completed (i.e., each training data sample has been used to update the weights and biases).

The challenge is deciding how few epochs to use, how small a mini-batch can be (larger batches are more accurate but slow down the rate of fitting), and how large a learning rate should be used.

Neural networks are one of the most popular machine learning techniques. However, they are black box models in the sense that they are difficult to interpret. The work being done at an individual hidden node, or in a hidden layer, is not readily explained.

Appendix 2—AI ethics guideline review

Table 15 provides an overview of the key principles, messages and requirements included in 18 published AI ethics guidelines or frameworks, where the blue shading indicates that an ethical concept is considered (either as a key principle or is passing) in that document (this is subjective and based on the views of the authors). The list is not exhaustive, and aims to provide a snapshot that covers multiple geographies and stakeholder groups (e.g., government, civil bodies and private sector).

TABLE 15: AN OVERVIEW OF KEY PRINCIPLES FROM ETHICAL GUIDELINES

			Transparency, openness, communication	Fairness, non-discrimination, diversity	Accountability (auditability, reporting negative impacts, redress)	Explainability, interpretability	Data privacy	Data security + cybersecurity	Societal and environmental well-being	Prevention of harm/misuse of AI	Robustness (reproducibility, accuracy, reliability)	Human rights, human autonomy etc.	Human oversight and control	Public awareness, education on AI and its risks	Impact on employment	Potential use in weapons	Cultural considerations[116]	Competence and skills of those using AI
Source	**Type of Body**	**Country**	**17**	**15**	**15**	**13**	**13**	**12**	**11**	**10**	**9**	**7**	**6**	**6**	**5**	**4**	**4**	**3**
(European Commission, 2019)	Intragovernmental	International	■	■	■	■	■	■	■	■	■	■	■	■	■	■		
(IBM, 2019)	Private sector—technology	U.S.	■	■	■	■	■	■									■	
(Institute of Electrical and Electronics Engineers, 2019)	Professional association	International	■		■	■	■		■	■	■	■		■				■
(Institute for Ethical AI & Machine Learning, n.d.)	Civil Society	U.K.	■	■		■	■	■			■		■		■			
(Microsoft, n.d. a)	Private sector—technology	U.S.	■	■	■		■	■	■		■							
(OECD, 2019)	International organisation	International	■	■	■	■	■	■	■		■	■	■		■			
(Monetary Authority of Singapore, 2018)	Government	Singapore	■	■	■	■	■											
(World Economic Forum, 2018)	Civil society	International	■	■	■	■										■		
(University of Montreal, 2018)	Academia	Canada	■	■	■	■	■	■	■	■	■	■	■	■		■	■	
(Beijing Academy of Artificial Intelligence, 2019)	Government	China	■	■	■	■		■	■	■	■	■		■	■			
(OpenAI, n.d.)	Civil society	U.S.							■	■		■						
(DeepMind, n.d.)	Private sector—technology	U.S.	■	■	■		■		■	■					■		■	

[116] Different groups and cultures may have different ethical views.

			Transparency, openness, communication	Fairness, non-discrimination, diversity	Accountability (auditability, reporting negative impacts, redress)	Explainability, interpretability	Data privacy	Data security + cybersecurity	Societal and environmental well-being	Prevention of harm/misuse of AI	Robustness (reproducibility, accuracy, reliability)	Human rights, human autonomy etc.	Human oversight and control	Public awareness, education on AI and its risks	Impact on employment	Potential use in weapons	Cultural considerations[116]	Competence and skills of those using AI
(Google, n.d.)	Private sector—technology	U.S.	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦						🟦	🟦	
(Partnership On AI (e.g. Apple, Microsoft, Google), n.d.)	Private sector—technology	International	🟦		🟦	🟦	🟦	🟦	🟦					🟦				
(BMJ, 2020)	Academia	U.K.	🟦	🟦		🟦		🟦		🟦	🟦							
(National Association of Insurance Commissioners, 2020)	Civil society	U.S.	🟦	🟦	🟦		🟦	🟦		🟦	🟦		🟦					🟦
(Insurance Europe, 2020)	Civil society	International	🟦	🟦	🟦	🟦												
(RSS and IFoA, 2019)	Professional association	U.K.	🟦	🟦	🟦		🟦	🟦	🟦	🟦		🟦	🟦	🟦				🟦

Appendix 3—European Commission's ethics guidelines for trustworthy AI and assessment list

This appendix provides further detail on the AI HLEG Guidelines introduced in Section 4.5 covering the following three sections of the proposed framework:

- **The Foundations of Trustworthy AI:** More 'abstract' principles for achieving trustworthy AI
- **The Implementation and Realisation of Trustworthy AI:** Methods to realise those principles in practice
- **The Assessment of Trustworthy AI:** How to operationalise those techniques and ensure that they become business-as-usual practices

FOUNDATIONS FOR TRUSTWORTHY AI

AI HLEG states that one of the initial steps in creating its guidelines was to reflect on what type of future, in terms of the values and society we want to have, is desired in order to determine how AI should be researched, developed, deployed and used in order to realise this vision; for the EC, the vision is for 'an AI-immersed future … where democracy, the rule of law and fundamental rights underpin AI systems and where such systems continuously improve and defend democratic culture' (European Commission, 2019).

The four families of fundamental rights that the AI HLEG believe are particularly applicable to AI systems are summarised below:

1. **Respect for human autonomy:** Each human has an intrinsic worth that AI systems should not diminish, compromise or repress. Humans must be able to keep full self-determination over their lives, with AI systems never being used in a way that would sift, sort, score, herd, condition, deceive, subordinate, coerce or manipulate humans. AI systems should be developed using human-centric design principles, with due regard to the fair allocation of labour and functions between the human and AI, to ensure human oversight of the AI system, and to leave opportunity for human choice (i.e., even if the AI provides a solution or decision, humans can choose to apply their own judgement in whether to accept that solution or decision).

2. **Prevention of harm:** AI systems must be developed, deployed and used in a way that ensures that they do not exacerbate harm (including intangible harm: social, cultural and political) to an individual or to a collective group, or otherwise adversely affect human beings (including the way of living of individuals and social groups). They must be safe, secure and technically robust to ensure they are not open to malicious use. AI HLEG draw out that particular care should be provided to:

   - Vulnerable persons
   - Situations where the system could cause or exacerbate asymmetries of power of information (e.g., between employee and employer, consumers and businesses, citizens and governments).

3. **Fairness:** AI HLEG believes there are two dimensions to fairness to consider—substantive (i.e., monitoring and measuring that AI systems are operating fairly) and procedural (i.e., creating procedures that ensure AI is applied fairly).

   The substantive dimension involves ensuring that the benefits and costs of AI systems are distributed in an equal and just way, with no individuals or groups being victims of unfair bias, discrimination or stigmatisation by an AI system.

   The procedural dimension involves ensuring there are mechanisms to contest decisions made by an AI system and by the humans operating them (which in itself requires accountability to be identifiable and decision making of AI systems to be interpretable), and an opportunity to seek appropriate redress in situations where AI systems have produced judgements that are not fair.

4. **Explicability:** AI systems must be transparent with the processes, capabilities and purpose of the system documented and openly communicated.

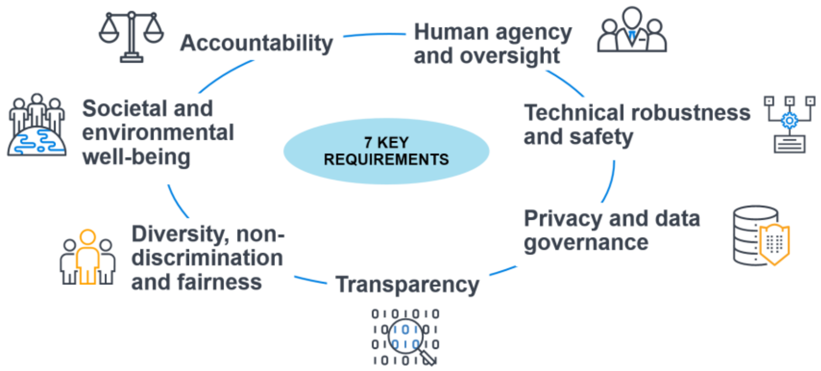
> Decisions made by AI systems must be explainable, to the extent possible, to those directly, and indirectly affected in order to facilitate any contests on fairness. Where such explanations are not possible (i.e., a black box model or algorithm) other explicability measures may be required, see Table 12 of this report for some explainability measures.
>
> The extent of explicability measures that should apply to an AI system is dependent on the context in which it is being used, and the severity of the consequence if its output is, for example, in error or inaccurate.

AI HLEG acknowledges that tensions may arise between these principles. For example, in the life insurance pricing context, the principle of prevention of harm and the principle of human autonomy may be in conflict: An AI tool may more accurately price risks, reducing premiums and ensuring more individuals can benefit from the social good of life insurance, but in doing so could entail processing and surveillance of individuals to a level that does not honour their intrinsic worth or respect each individual's privacy. In such cases, AI HLEG states that democratic principles require 'due process and open political participations and methods of accountable deliberation to deal with such tensions.' In certain situations no ethically acceptable trade-offs can be established.

IMPLEMENTATION AND REALISATION OF TRUSTWORTHY AI

In its guidelines, AI HLEG acknowledges that ethical principles alone are too abstract for those working in the field of AI to use in the creation of ethical AI systems. Consequently, the second part of the guidelines provide seven key requirements that build on the core principles.

FIGURE 22: KEY REQUIREMENTS FOR TRUSTWORTHY AI



Accountability
Human agency and oversight
Technical robustness and safety
Privacy and data governance
Transparency
Diversity, non-discrimination and fairness
Societal and environmental well-being
7 KEY REQUIREMENTS

The requirements are applicable to various stakeholders (developers,[117] deployers,[118] end users and society as a whole) that engage with the AI system at various stages of its life cycle, noting that each stakeholder had a unique role to play in the application of the requirements:

- **Developers:** Should implement the requirements 'by design,' ensuring the development of AI systems meets the requirements.
- **Deployers:** Have a duty to ensure that the systems they use, and the products and services they offer, meet the requirements.
- **End users and the broader society:** Must be aware of the requirements and must be able to request that the requirements are upheld by developers and deployers

The requirements are summarised in Figure 23.

[117] Those who are involved in researching, designing and/or developing AI systems.

[118] Public or private organisations that are using AI systems in the products and services they provide, or in their business processes.

FIGURE 23: 7 KEY REQUIREMENTS FOR REALISATION OF TRUSTWORTHY AI

1. Human agency and oversight

   - *Fundamental rights:* A fundamental rights impact assessment should be undertaken prior to the development of an AI system if there are situations where the AI could adversely impact fundamental rights. Such an assessment should determine if such risks can be reduced or mitigated, and more generally whether they can be justified. There should be a mechanism in place to receive external feedback on the system should there be a risk the AI system infringes on fundamental rights. Presumably such external feedback could be sought during development of the system, post-deployment, or both.

   - *Human agency:* Ensure people have sufficient expertise and tools available to understand and interact with the AI system, and, wherever possible, provide challenge to how the system is working. In addition, users of an AI system should have the right not to be subject to a decision based solely on automated processing if the decision would have a legal effect on the user, or would otherwise significantly impact them.
   - *Human oversight:* Humans should provide oversight to the AI via HITL,[119] HOTL[120] or HIC[121] approaches to ensure human autonomy is not undermined, and to monitor for adverse effects from the AI system.

2. Technical robustness and safety

   - AI systems should be developed such that they prevent risks and unacceptable harm, to minimise unintentional and unexpected harm.

   - AI systems should be secure and resilient to attack.

   - Contingency plans should be in place in the event that issues arise. For example, being able to switch from the AI tool to a rules-based procedure, or by the AI system flagging to a human operator that an issue may have arisen, and asking whether it is appropriate to continue in their operation.

   - AI systems must be accurate and reliable (i.e., working well for a range of inputs and in a range of situations), and the results must be reproducible.

3. Privacy and data governance

   - *Privacy and data protection:* This must be guaranteed throughout the entire life cycle of an AI system. Any data that is collected on a person (preferences, sexual orientation, age, gender, religious or political views) must not be used unlawfully (i.e., not in breach of data regulation such as GDPR) or used to unfairly discriminate against that person.

   - *Quality and integrity of data:* The quality of the data must be reviewed (tested and documented) to ensure it does not contain inaccuracies, errors, mistakes or contain any socially constructed biases (e.g., racism, sexism). Data should also be reviewed for its integrity to ensure it is not malicious and will change the behaviour of an AI system in an adverse way. Such testing and documentation should be undertaken when planning, training, testing and deploying an AI system.

   - *Access to data:* Ensure data protocols are in place that outline who has access to data and under what circumstances. Only those who are qualified to use the data, and who need to use the data, should be allowed to do so.

4. Transparency

[119] Intervention in every decision cycle of the system.

[120] Intervention in the design cycle of the system and monitoring the system's operation.

[121] Oversee all activity of the AI system, with discretion on when, how and whether to use a system in any particular situation (including the ability to override a decision made by a system).

- *Traceability:* Data sets and processes of an AI system that feed into decision making must be documented to a high standard to facilitate error-checking and prevent future mistakes.
- *Explainability:* There must be the ability to explain both (1) the technical process of an AI system and (2) its decision making. A human being should be capable of understanding a decision made by an AI system and be able to trace how that decision was made.
- *Communication*: Humans have a right to know that they are not interacting with another human but an AI system, or that a decision made that will impact them was determined by an AI system. A human should have the right to decide against an interaction with an AI system and to opt for human interaction.

5. Diversity, nondiscrimination and fairness

- *Avoidance of unfair bias:* Both the training and operational data set may suffer from unintended historical bias, and its use in an AI system could lead to unintended prejudice and discrimination to certain groups. Identifiable, discriminatory bias should be removed in the data collection phase if this is possible.

  In addition, there must not be intentional exploitation of the biases of consumers, and the AI system should not engage in unfair competition (e.g., homogenising prices via collusion or non-transparency in the market).

  Furthermore, to reduce the risk of bias in the way AI systems are developed (i.e., how they are programmed), oversight processes could be put in place at the development stage to transparently analyse and review the system's purpose, constraints, requirements and decisions. Moreover, encouraging the teams responsible for developing and delivering AI systems to recruit people who reflect the diverse backgrounds and cultures that will engage with the system (i.e., end users), and society as a whole, could reduce the risk of unfair bias in the development process.
- *Accessibility and Universal Design:*[122] An AI system must be designed to be user-centric so that all people can access and use the products or services regardless of their age, gender, abilities or other factors.
- *Stakeholder participation:* Mechanisms should be put in place to receive regular feedback from any stakeholders who may be directly, or indirectly, affected by the AI system during its life cycle.

6. Societal and environmental well-being

There must be consideration of broader society, the environment and of other sentient beings at all stages in the life cycle of an AI system.

- *Sustainable and environmentally friendly AI:* Consider whether the development, deployment and use processes in the AI system's life cycle are environmentally friendly (and opting for more environmentally friendly choices), taking into consideration the whole supply chain (e.g., energy consumption, resource usage)
- *Social impact:* Monitor and consider the effects on social relationships and our social agency[123] as AI systems become more widely used in all areas of our lives.
- *Society and democracy:* Consider the impact of the AI system from a societal perspective, considering its implications for institutions, democracy and society at large.

[122] Universal Design is an enabling and empowering approach to design that aims to ensure equitable access and active participation of all people for the product or service, as opposed to a 'one size fits all' approach.

[123] The capacity of individuals to act independently and make their own free choices.

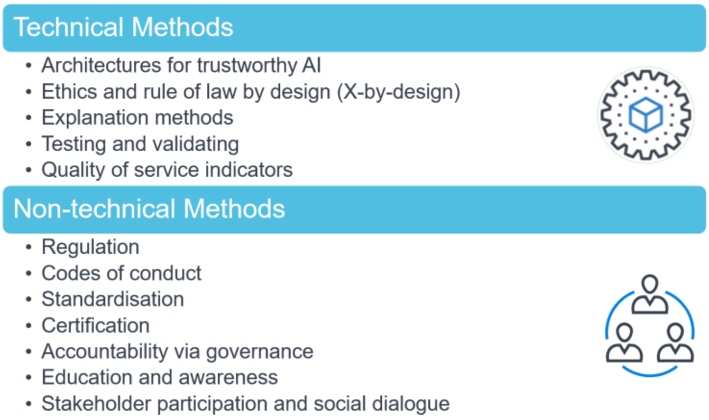
7. Accountability It is necessary for mechanisms to be put in place that ensure responsibility and accountability for AI systems and their outcomes, both before and after their development, deployment and use.

   - *Auditability:* AI systems could be evaluated by internal and external auditors responsible for assessing the algorithms, data and design processes. Availability of such reports would contribute to trustworthiness. In cases where fundamental rights (including safety) are potentially at risk, there should be an independent external audit of the system.
   - *Minimisation and reporting of negative impacts:* It is important to identify, assess, document and minimise the potential negative impacts of AI systems, with impact assessments performed at all stages of the AI system's life cycle, proportionate[124] to the risk posed.

     There must be the ability to report on the actions or decisions made by an AI system that contribute to a certain outcome, and the ability to respond to the consequences of such an outcome. There must be protection for those who report concerns about an AI system (e.g., whistleblowers, nongovernmental organizations).
   - *Trade-offs:* In situations where conflicts between certain requirements arise, trade-offs must be documented and evaluated against their risk to ethical principles—notably those relating to fundamental rights. In situations where no ethically acceptable trade-offs can be identified, then the development, deployment and use of the system should not proceed in that form.
   - *Redress:* There should be mechanisms put in place that provide accessible and adequate redress in situations where an adverse impact on a person, or group, occurs.

In order to operationalise the above requirements, AI HLEG presents technical, and nontechnical methods for firms to consider:

FIGURE 24: AI HLEG METHODS TO REALISE TRUSTWORTHY AI



Technical Methods

- Architectures for trustworthy AI
- Ethics and rule of law by design (X-by-design)
- Explanation methods
- Testing and validating
- Quality of service indicators

Non-technical Methods

- Regulation
- Codes of conduct
- Standardisation
- Certification
- Accountability via governance
- Education and awareness
- Stakeholder participation and social dialogue

[124] The use of assessments must be proportionate to the risk that the AI system pose.

The technical methods comprise:

- **Architectures:** The requirements must be turned into procedures and/ or constraints on procedures that form part of the AI system's architecture**.** AI HLEG suggest that this could be through the creation of a set of rules (behaviours or states) that the system should always follow (i.e., an acceptable list), and a set of prohibitions (i.e., an unacceptable list) for the system. Then, compliance monitoring can be performed.

  In addition, a 'sense-plan-act' approach to ensuring ethics is integrated in all stages of the life cycle could be adopted (1) sense—develop the system so that it understands what is needed to ensure adherence to the requirements; (2) plan—develop a model, ruling out those which do not adhere to the requirements; and (3) act—the actions available to the system should be restricted to those that realise the requirements.
- **Ethics and rule of law by design:** Ethics-by-design and rule of law by design whereby, the above ethical requirements that reflect desired ethical principles are embedded in the design of the AI system. Given firms will be reviewing how the system could avoid negative impacts and uphold ethical principles from the design stage, the system should be developed to be secure in its processes, data and outcomes, resilient to adversarial data and attacks, and contain contingency plans such as mechanisms to shut down (and restart) the system in the event of attacks.
- **Explanation methods:** As noted in Section 3.8, XAI is a field where there is a lot of recent and ongoing research and investment to target one of the biggest perceived shortcomings of AI—the ability to explain a system's mechanisms and how it comes to its solutions. There are techniques that facilitate the explanation of outcomes and decisions made by an AI system (see Table 12 for some examples), and firms should be using available tools to support interpretability, particularly in cases that can negatively impact policyholders. It should be noted that the more complex the AI system the more challenging explanation becomes, with deep neural networks still facing interpretability challenges.
- **Testing and validation of systems:** Throughout the life cycle of the AI system, all components of the system (e.g., data, models and environments) should be monitored to ensure that, despite the non-deterministic nature of the system, the model produces stable, robust and sensible results using realistic data. Testing should be completed by diverse groups of people who will be able to provide their perspective on whether the results look sensible and fair. Adversarial testing[125] could be carried out by diverse groups aiming to identify vulnerabilities in the system.
- **Quality of service indicators:** A set of indicators relating to security and safety could be used as reference so that it is possible to establish whether the AI system has been tested, trained and developed with such considerations in mind.

The nontechnical methods, to be applied on an ongoing basis, described by the AI HLEG framework included:

- **Regulation:** A review of current legislation to determine if it needs to be revised or adapted, or if new regulation needs to be introduced, to ensure AI is used in an ethical manner.
- **Codes of Conduct:** A range of ideas are presented by the AI HLEG, including that a firm could:
  - Sign up and adhere to an established framework or set of guidelines (for example the AI HLEG Guidelines).
  - Adapt its charter of responsibility to include an ethical perspective on AI development, deployment and use.
  - Adjust its key performance metrics and indicators to include some that monitor adherence to ethical AI.
  - Adapt its internal policy documents or codes of conduct to include ethical AI development, deployment and use.

[125] Adversarial machine learning attempts to 'fool' models by supplying deceptive input.

- **Standardisation:** In certain industries there are established standards that work as a quality management and accreditation system for firms (e.g., ISO Standards, TAS for actuarial work in the U.K.).

  AI HLEG suggests that standards could be established for the ethical development, deployment and use of AI, with firms signing up to these standards (and to be audited against those standards) such that users and consumers can use this to inform their decisions on the companies from which they buy products and services.

- **Certification:** Firms can attest to the public that their AI systems adhere to ethical principles through a certification. Such a certification would be specific to how the AI is being applied, and the AI techniques at use, and aligned with industrial and societal standards.

- **Governance frameworks for accountability:** A firm can establish governance frameworks that ensure accountability for the ethical decisions associated with the development, deployment and use of AI systems. Potential frameworks could include the appointment of an individual, or an internal or external panel or committee, responsible for ethics relating to AI systems and that provides oversight and advice to the board of the firm.

- **Education and awareness:** In order to facilitate the widespread understanding in society of the impact and implications of AI systems, communication, education and training are required. This will also ensure that the public is aware that they have a role to play, and should play, in determining how society develops this new technology.

- **Stakeholder participation and social dialogue:** Firms should be transparently and openly discussing the use and impact of AI systems with the wider public, actively seeking participation and dialogue.

- **Diversity and inclusivity in teams designing AI systems:** To ensure AI systems are objective and consider the different perspectives, needs and objectives of all groups in society, teams involved in the life cycle of the AI system (i.e., those that design, develop, test and maintain, deploy and procure AI systems) must reflect the diversity of users and of society in general. AI HLEG states that ‘ideally, teams are not only diverse in terms of gender, culture, age, but also in terms of professional backgrounds and skill sets.’

ASSESSMENT OF TRUSTWORTHY AI

Finally, to actually move these concepts into a practical framework that companies can implement to embed ethics in their AI systems, the AI HLEG Guidelines present a non-exhaustive assessment list, broken down into its seven key areas. It also provides guidance on how firms could work the list into existing governance structures, or use it in newly created governance structures.

An initial list published in the AI HLEG Guidelines in April 2019 was subject to a piloting period involving more than 350 stakeholders, with a final document, and accompanying web-based tool, released in July 2020—dubbed the Assessment List for Trustworthy AI (ALTAI) (AI HLEG, 2020). This ALTAI is clearer and more focussed than the draft AI HLEG Guidelines but captures the majority of its key considerations.

The ALTAI requires tailoring to the specific industry and context of the AI system and is summarised below in Figure 25.

FIGURE 25: ALTAI ASSESSMENT CHECKLIST

1. HUMAN AGENCY AND OVERSIGHT	
Human agency and autonomy	▪ Has the AI system been designed to interact, guide or take decisions by human users, in a way that affects humans or society? o Could the system cause confusion to some or all users on whether a decision, content, advice or outcome has been made by an AI system or a human? Have users, and other humans, been informed should this be the case? ▪ Could the AI system affect human autonomy by: o **Generating overreliance by users**. What safeguards are in place to ensure there is not overreliance on the system by users? o **Interfering with the decision-making process of users** in an unintended or adverse way. What safeguards are in place to avoid the system indirectly affecting human autonomy? ▪ Is the AI system stimulating human social interaction between its users? ▪ Is the AI system at risk of creating **adverse psychological behaviours**, such as human attachment to the AI system, the stimulation of addictive behaviour or the manipulation of the behaviour of its users? If so, what measures are in place to minimise or mitigate such risks?
Human oversight	▪ Is the system: o Self-learning or autonomous? o Overseen by a HITL, HOTL or HIC? ▪ Have those humans overseeing the system (i.e., HITL, HOTL, HIC) been provided training on how to sufficiently exercise oversight? ▪ Have you put in place any mechanisms to identify and respond to any adverse effects of the AI systems for users? ▪ Is there a procedure in place to safely stop an AI system from operating if required? ▪ Are there any oversight and control measures in place to deal specifically with the self-learning or autonomous nature of the AI system?
2. TECHNICAL ROBUSTNESS AND SAFETY	
Resilience to attack and security	▪ In the event of design or technical faults, defects, outages, attacks, misuse, or inappropriate or malicious use of the AI system, can the system have damaging effects to human safety (and more general of societal safety)? ▪ Has the AI system been certified for cybersecurity, or any other security standards?[126] ▪ How vulnerable is the AI system to cyberattacks? o Has the system been tested for potential forms of attack to ensure it is resilient? o Have you considered attack points: ▪ Data poisoning, such as the manipulation of training data? ▪ Model evasion, whereby attackers ensure that data is classified in the way they would like? ▪ Model inversion, whereby the attacker infers the model parameters, opening the model up for exploitation? ▪ Have you put in place measures to ensure that, throughout the entire life cycle of the AI system, the system will be resilient to potential attacks, robust, and retain its integrity? ▪ Have you carried out red team testing[127], or penetration testing[128] of the AI system? ▪ Have you kept users informed of the duration of any security coverage and updates? How often do you expect to provide security updates for the AI system?

[126] The ALTAI provides the EU Cybersecurity Act as an example.

[127] The goal of red teaming is to simulate the actions of malicious hackers going beyond that of a penetration test to include not only the IT systems but also the people, processes and physical security of facilities.

[128] The goal of penetration testing, or ‘pentesting,’ is to simulate the actions of malicious hackers, primarily focusing on IT controls and governance to identify any flaws where and how a hacker may try and target you, how your security measures would react, and the possible magnitude of a breach.

General safety	▪ Have you defined risks, risk metrics and risk levels of the AI system in each use case that the system could be deployed for? ○ Have you put in place processes that allow you to continuously monitor and measure these risks? ○ Have you communicated these potential risks to users? ▪ Have you identified if any potential threats exist in the AI system (e.g., design faults, technical faults, environmental threats), and any potential adverse results of such threats? ○ Have you considered the potential risk of malicious use, misuse or inappropriate use of the AI system? ○ For such risks, have you defined safety criticality levels (e.g., related to death or serious injury as a result of the system)? ▪ For critical AI systems, have you analysed whether the decisions made by the system are reliant on the system having stable and reliable behaviour? ▪ Do you have a plan when faults are identified in your AI systems (e.g., switching from a complex AI system to a rules-based procedure or redirecting to a human operator)? ▪ Are there procedures in place to determine when an AI system has been changed or updated, such that a new review of its technical robustness and safety is required?
Accuracy	▪ In the event of a low level of accuracy of the AI system, could there be any critical, adversarial or damaging consequences? ▪ What measures do you have in place to ensure all data used to develop the AI system is complete, recent, of high quality and accuracy, but also representative of the environment in which the system will be deployed? ▪ Have you put in place processes and procedures to monitor and document the accuracy of the AI system? ▪ Have you determined whether the AI system could affect its environment and invalidate the data or assumptions used to train it (including how this could result in adversarial effects such as attempts to 'game' the system)? ▪ Have you put in place processes and procedures to ensure that the AI system's level of accuracy has been communicated to users?
Reliability, fall-back plans and reproducibility	▪ In cases where a model has low reliability and/ or low reproducibility, is it possible for the AI system to cause critical, adversarial or damaging consequences (e.g., death or serious injury)? ○ Have you put in place systems to monitor if the system is operating as intended? ○ Is the AI system reproducible in specific contexts or conditions? ▪ Have you tested whether there are any specific contexts or conditions that need to be taken into account in order to ensure the model is reproducible? ▪ Are there documented (e.g., logged) verification and validation methods in place to review different parts of the AI system's reliability and reproducibility? Are such testing and verification methods part of an operationalised process? ▪ Have you created a fall-back plan in the event of errors in the AI system? ○ Have you defined thresholds and built a governance structure that triggers your fall-back plan or other alternatives to using the AI system where there has been a failure? ▪ For events where the AI system produces a result with a low confidence score in terms of accuracy or reliability, do you have procedures in place for handling these cases? ▪ Is your AI system using continual learning?[129] ○ If so, have you considered possible adverse results from the AI system adapting (in response to new data) to learning novel or unusual methods to score well on its objective function?
3. PRIVACY AND DATA GOVERNANCE	
Privacy	▪ Have you considered what the effect your AI system has on the right to privacy, the right to physical, mental and/or moral integrity and the right to data protection? ▪ If required based on the use case, do you have in place mechanisms for people to flag data privacy issues (e.g., in the process of data collection [i.e., for training and operation] or processing)?

[129] AI systems that learn continuously (e.g., using online data) and are constantly adapting autonomously are described as continual learning.

Data governance	▪ Is your AI system being trained, or was it developed (or trained), by using or processing 'personal data' as defined by the GDPR (including special categories of personal data)? ▪ Have you put in place measures (some of which are mandatory) under the GDPR, or a non-European equivalent, such as: carrying out a data protection impact assessment; designating a data protection officer (involved in early stages of development, procurement or use phases); oversight mechanisms for data processing; measures to achieve privacy-by-design as default (e.g., encryption, pseudonymisation, aggregation, anonymisation); and data minimisation (in particular for personal data)? ▪ Have you included the right to withdraw consent, to object, and the right to be forgotten into the development of your AI system? ▪ Have you considered the privacy and data protection implications of data collected, generated or processed over the course of the AI system's life cycle? ▪ For any non-personal data used (as training data or elsewhere) by the AI system, have you also considered the privacy and data protection implications of the AI system? ▪ Has your system been aligned with relevant data management and governance standards (e.g., ISO, IEEE)?
4. TRANSPARENCY	
Traceability	▪ Have you ensured measures are in place to address the traceability of the AI system during all stages in its life cycle? ▪ Have you ensured measures are in place to continuously assess the quality of the input data to the AI system (e.g., automated quality assessment of data input for missing values, gaps, errors, etc.)? ▪ Have you ensured measures are in place to continuously assess the quality of the output of the AI system (e.g., automated quality assessment of output such as checking prediction scores are within certain ranges and/or anomaly detection)? ▪ Is it possible to trace back the exact data that was used by the AI system to make a particular decision or recommendation? ▪ Is it possible to trace back the exact model or rules that led to a particular decision or recommendation? ▪ Have you ensured that measures are in place to adequately record any decisions or recommendations made by the AI system?
Explainability	▪ Have you explained the decision, or decisions, made by the system to its users (e.g., through workshops or other ways of drawing out any misunderstandings from users on how the system makes decisions)? ▪ Do you have processes in place to continuously survey whether users understand the decision, or decisions, made by the AI system?
Communication	▪ In cases of interactive AI systems (e.g., chatbots), have you told the users of the AI system that they are engaging with an AI system not another human? ▪ Have you ensured measures are in place to inform users about the purpose, criteria and limitations of any decisions made by the AI system? ▪ Have you made the benefits of the AI system known to users? ▪ Are users made aware of any technical limitations and potential risks of the AI system to its users (e.g., its level of accuracy and/or error rates)? ▪ Have you provided users with sufficient training material and disclaimers on how the AI system can be used?

5. DIVERSITY, NONDISCRIMINATION AND FAIRNESS	
Avoidance of unfair bias	▪ Both in the input data used, and the algorithm design itself, have you put in place a strategy to ensure you do not create, or reinforce, unfair bias in the system? ▪ Have you considered whether your data is diverse and representative of users, and tested the system for specific groups or potential problematic use cases that the data may not sufficiently describe? ▪ Have you researched and used any technical tools at your disposal to ensure you understand the data, the model and how the model is performing? ▪ Have you established processes for testing and monitoring for bias in your system at all stages in the life cycle? ▪ If relevant, are you considering the diversity and representativeness of end users in the data used by the AI system? ▪ Have you put in place any educational and awareness initiatives to help those responsible for designing and developing AI to become more aware of any potential bias they may impose in their role in designing or developing the system? ▪ Have you put in place a mechanism to allow people (both users and other groups potentially indirectly affected) to flag issues related to bias, discrimination or unfair outcomes, including clear steps on how and to whom such issues can be raised? ▪ Have you determined a definition for 'fairness' that is applied when you design AI systems? ○ Is this definition commonly used and implemented in the process of designing and developing the AI system? ○ Did you consult with any groups that could be vulnerable to discrimination, bias or unfair outcomes to determine their definition of fairness and see how it aligns with yours? ○ Do you have any quantitative metrics to measure and test whether the system is producing fair outcomes? ○ Have you established mechanisms to ensure fairness in your system?
Accessibility and universal design	▪ Have you made sure your system has been designed to meet the needs of a wide range of users with different preferences and abilities? ▪ Can those with special needs or disabilities, or those at risk of exclusion interact with your AI system's user interface? ○ Can information on the AI system, and the system's user interface, be accessed by those who use assistive technologies? ○ Were those users who need assistive technology consulted (or otherwise involved) in the planning and development of the AI system? ▪ Have Universal Design principles been considered at all stages of the planning and development process? ▪ Have you taken into account the impact of the AI system on the potential users of the system? ○ Have you assessed whether the development team engaged with the possible target users of the system? ○ Have you considered whether certain groups could be disproportionately affected by poor outcomes from your AI system? ○ Have you assessed whether there is a risk of unfairness of the system to certain communities who use the system?
Stakeholder participation	▪ Have you put in place a mechanism to include the participation of different stakeholders in the AI system's design and development?

6. SOCIETAL AND ENVIRONMENTAL WELL-BEING	
Environmental well-being	▪ Are there any negative environmental impacts from the system? ▪ Have you put in place mechanisms to measure the environmental impact of the system at all stages of its life cycle? ▪ Have you put in place measures to reduce the environmental impact of the system at all stages of its life cycle?
Impact on work and skills	▪ Are there any consequences for human work and work arrangements due to the AI system? ▪ Prior to introducing the AI system in your company, did you inform and consult those workers that would be impacted by its introduction (including their representatives such as trade unions)? ▪ Have you put in place measures to ensure that the social impacts of the AI system on human work are well understood? ○ Do workers understand the functionality, capabilities, and limitations of the AI system? ▪ Is there a social risk that the AI system de-skills the workforce? If so, have measures been put in place to mitigate these risks? ▪ Does the AI system promote or require new digital skills, and if so, have you provided training for re-skilling and up-skilling?
Impact on society at large or democracy	▪ Could the AI system have a negative impact on society at large, or democracy? ○ Have you assessed the societal impact of using the AI system beyond those users that are directly impacted (e.g., indirectly affected stakeholders or society at large)? ○ Have you put in place measures to minimize any potential societal harm arising from the use of the AI system? ○ Have you put in place measures to ensure that democracy is not adversely impacted by the use of the AI system?
7. ACCOUNTABILITY	
Auditability	▪ Have you put in place mechanisms that allow the AI system to be audited? ▪ Have you put in place measures to ensure the system is subjected to an independent audit?
Risk management	▪ Did you receive any external guidance, or put in place an external auditing process to oversee ethics and accountability? Does such involvement extend beyond the development phase of the AI system? ▪ Have you provided training on risk? Does training include the potential legal framework applicable to the AI system? ▪ Have you considered putting together a panel, committee or similar mechanism responsible for discussing and reviewing accountability and ethics practices? ▪ Have you established processes to monitor and assess the system's adherence to the ALTAI? ○ In such processes, do you identify and document any conflicts between the requirements of the items in the checklist other than accountability, or between different ethical principles within a requirement? Are you documenting explanations of any 'trade-off' decisions made? ○ For those employees responsible for monitoring and assessing the adherence to the ALTAI, are you providing training (and does this training include the legal framework applicable to the AI system)? ○ Have you put in place a mechanism for third parties or workers to report any issues they have identified with the system (e.g., vulnerabilities, risks, biases)? ○ Are changes required to your risk management processes to ensure you are monitoring and assessing the systems adherence to the ALTAI? ▪ Have you put in place a mechanism to ensure there is adequate redress for those that are harmed or adversely impacted by the AI system?

The EC has also created an optional accompanying web-based assessment tool that firms can use for their own self-assessment. For those using the web-based tool, the ALTAI will provide results from your self-assessment by way of a spider-web chart, providing a rating for each of the seven key requirements listed in the table above. The tool also provides recommendations in light of answers provided (albeit most are along the lines of '*put in place A, B and C*,' '*establish mechanisms to ensure X, Y and Z*') that turns the list into an ongoing gap analysis for those that are using it.

Given the ALTAI has been subject to a robust testing and consultation phase involving over 350 stakeholders, the ALTAI may be useful for those insurers wanting to determine what best practice looks like in this rapidly evolving field.

Acknowledgements

We would like to thank Samuel Burgess, Jessica Crowson, Dana-Marie Dick, Amy Nicholson, William Smith and Kelly Yang for their contributions to the research and drafting of this paper.

References

ABI. (2020, September 7). *Detected insurance fraud—new data shows that every five minutes a fraudulent claim is discovered*. ABI. Retrieved November 16, 2020, from https://www.abi.org.uk/news/news-articles/2020/09/detected-insurance-fraud/.

ABI. (n.d.). *Fraud.* ABI. Retrieved November 16, 2020, from https://www.abi.org.uk/products-and-issues/topics-and-issues/fraud/.

Ada Lovelance Institute. (2018, May 15). *About us.* Ada Lovelance Institute. Retrieved November 16, 2020, from https://www.adalovelaceinstitute.org/about-us/.

Ada Lovelace Institute. (2020a, August). C*onfidence in a crisis? Building public trust in a contact tracing app.* Ada Lovelance Institute. Retrieved November 16, 2020, from https://www.adalovelaceinstitute.org/wp-content/uploads/2020/08/Ada-Lovelace-Institute_COVID-19_Contact_Tracing_Confidence-in-a-crisis-report-3.pdf.

Ada Lovelace Institute. (2020b, August 17). *Rapid, online deliberation on COVID-19 technologies.* Ada Lovelance Institute. Retrieved November 16, 2020, from https://www.adalovelaceinstitute.org/our-work/rethinking-data/rapid-online-deliberation-on-covid-19-technologies/.

Aequitas. (2020, October 5). *The Bias and Fairness Audit Toolkit for Machine Learning.* Retrieved November 6, 2020, from https://dssg.github.io/aequitas/#.

Ahmad, N. & Siddique, J. (2017). Personality assessment using Twitter tweets*.* Elsevier B.V.

AI HLEG. (2020, July 14). *The Assessment List for Trustworthy Artificial Intelligence for Self Assessment.* European Commission. https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=68342.

Aleandri, M. (2019, May 10). *Data science in insurance—some introductory case studies.* [SAO Research Dissertation, Institute and Faculty of Actuaries]. https://www.actuaries.org.uk/system/files/field/document/AleandriM_DataScienceinInsurance_20190510_SA0_dissertation.pdf.

Algorithmwatch. (2020, September 27). AI Ethics Guidelines Global Inventory. Algorithmwatch. Retrieved November 16, 2020, from https://inventory.algorithmwatch.org/.

Amina, A., Al-Obeidat, F., Shah, B., Adnan, A., Loo, J., & Anwar, S. (2019). Customer churn prediction in telecommunication industry using data certainty. *Journal of Business Research*, pp. 290–301.

Artís, M., Ayuso, M., & Guillén, M. (2002). Detection of automobile insurance fraud with discrete choice models and misclassified claims. *The Journal of Risk and Insurance*.

Arts and Humanities Research Council. (2020, February 5). *New humanities-led network will put social justice at the heart of AI research*. Arts and Humanities Research Council. Retrieved November 16, 2020, from https://ahrc.ukri.org/newsevents/news/new-humanities-led-network-will-put-social-justice-at-the-heart-of-ai-research/.

Aviva. (2020, September 25). *'What's the value of my pension?' Just ask Alexa.* Aviva. Retrieved November 16, 2020, from https://www.aviva.co.uk/services/more-from-aviva/alexa/.

Babaoglu, C., Ahmad, U., Durrani, A., & Bener, A. (2017). Predictive modeling of lapse risk: An international financial services case study. *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC).* Banff: IEEE.

Babel, B., Buehler, K., Pivonka, A., Richardson, B., & Waldron, D. (2019, February 19). *Derisking machine learning and artificial intelligence*. McKinsey & Company. Retrieved November 16, 2020, from https://www.mckinsey.com/business-functions/risk/our-insights/derisking-machine-learning-and-artificial-intelligence#.

Balasubramanian, R., Libarikian, A., & McElhaney, D. (2018, 30 April). *Insurance 2030—the impact of AI on the future of insurance*. McKinsey & Company. Retrieved November 16, 2020, from https://www.mckinsey.com/industries/financial-services/our-insights/insurance-2030-the-impact-of-ai-on-the-future-of-insurance.

Barthelemy, L. (2019, May 24). *One year on, EU's GDPR sets global standard for data protection*. Phys.org. Retrieved November 16, 2020, from https://phys.org/news/2019-05-year-eu-gdpr-global-standard.html.

BBC. (2018, March 31). *Tesla in fatal California crash was on autopilot*. BBC News. Retrieved November 16, 2020, from https://www.bbc.co.uk/news/world-us-canada-43604440.

BBC. (2020, August 20). *A-levels and GCSEs: How did the exam algorithm work?* BBC News. Retrieved November 16, 2020, from https://www.bbc.co.uk/news/explainers-53807730.

Beijing Academy of Artificial Intelligence. (2019). *Beijing AI principles*. Beijing Academy of Artificial Intelligence. Retrieved November 16, 2020, from https://www.baai.ac.cn/news/beijing-ai-principles-en.html.

Bernardino, G. (2020, April 29). *Digital responsibility and the role of actuaries*. EIOPA. Retrieved November 16, 2020, from https://www.eiopa.europa.eu/content/digital-responsibility-and-role-actuaries_en#RoleofEIOPA:EIOPA%E2%80%99sExpertGrouponDigitalEthicsinInsurance.

Bestow. (n.d.). What is algorithmic underwriting? Bestow. Retrieved September 25, 2020, from https://support.bestow.com/what-is-algorithmic-underwriting.

BEUC. (2020, May 19). *The Use of Big Data and Artificial Intelligence in Insurance.* BEUC. https://www.beuc.eu/publications/beuc-x-2020-039_beuc_position_paper_big_data_and_ai_in_insurances.pdf.

Cornillie, C. (2019, March 28). *Finding artificial intelligence money in the fiscal 2020 budget*. Bloomberg Government. https://about.bgov.com/news/finding-artificial-intelligence-money-fiscal-2020-budget/.

BMJ. (2020). Machine learning and artificial intelligence research for patient benefit: 20 critical questions on transparency, replicability, ethics and effectiveness. *BMJ*.

BoE and FCA. (2019, October). *Machine Learning in U.K. Financial Services*. Bank of England. https://www.bankofengland.co.uk/-/media/boe/files/report/2019/machine-learning-in-uk-financial-services.pdf.

BoE and FCA. (2020, January). *Financial Services Artificial Intelligence Public-Private Forum: Terms of reference*. FCA. https://www.fca.org.uk/publication/corporate/financial-services-artificial-intelligence-public-private-forum-terms-of-reference.pdf.

Bolton, L. E., Keh, H. T., & Alba, J. W. (2018). How do price fairness perceptions differ across culture? *Journal of Marketing Research*.

Boodhun, N., & Jayabalan, M. (2018). Risk prediction in life insurance industry using supervised learning algorithms. *Complex & Intelligent Systems*, 145–154.

Borovkova, S., & Tsiamas, I. (2019). An ensemble of LSTM neural networks for high-frequency stock market classification. *Journal of Forecasting*, 600–619.

Bracke, P., Datta, A., Jung, C., & Sen, S. (2019, August). *Staff working paper No. 816: Machine learning explainability in finance: an application to default risk analysis.* Bank of England. https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/machine-learning-explainability-in-finance-an-application-to-default-risk-analysis.pdf?la=en&hash=692E8FD8550DFBF5394A35394C00B1152DAFCC9E.

Britannica. (2020, September 27). *Ethics*. Britannica. Retrieved November 16, 2020, from https://www.britannica.com/topic/ethics-philosophy.

Bryce, C. (2020, March). *IFoA certificate in data science—Colin Thores interview*. APR Actuarial Solutions. Retrieved November 16, 2020, from https://aprllp.com/certificate-in-data-science-colin-thores-interview.

Cao, L., Preece, R., & Baker, G. (2019). *AI pioneers in investment management.* CFA Institute. Retrieved November 16, 2020, from https://www.cfainstitute.org/-/media/documents/survey/AI-Pioneers-in-Investment-Management.ashx.

Cantle, N. (2017, June). *Emerging risk analytics: Application of advanced analytics to the understanding of emerging risk.* Milliman. https://milliman-cdn.azureedge.net/-/media/milliman/importedfiles/uploadedfiles/insight/2017/emerging-risk-analytics.ashx.Castellani, G., Fiore, U., Marino, Z., Passalacqua, L., Perla, F., Scognamiglio, S., & Zanetti, P. (2018). *An investigation of Machine Learning Approaches in the Solvency II Valuation Framework.* SSRN.

Castro, D., McLaughlin, M., & Chivot, E. (2019, August 19). *Who Is Winning the AI Race: China, the EU or the United States.* Center for Data and Innovation. Retrieved November 16, 2020, from https://www.datainnovation.org/2019/08/who-is-winning-the-ai-race-china-the-eu-or-the-united-states/#_edn58.

Cave, D. S. (2019, January 28). *The ethical and political questions raised by AI*. Ada Lovelace Institute. Retrieved November 16, 2020, from https://www.adalovelaceinstitute.org/the-ethical-and-political-questions-raised-by-ai/.

CDEI. (2019a, October 10). CDEI bias review—Call for evidence: Summary of responses. CDEI. Retrieved November 16, 2020, from https://www.gov.uk/government/publications/responses-to-cdei-call-for-evidence/cdei-bias-review-call-for-evidence-summary-of-responses.

CDEI. (2019b). Introduction to the Centre for Data Ethics and Innovation. CDEI. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/787205/CDEI_Introduction-booklet.pdf.

CDEI. (2019c, September 12). Snapshot paper—AI and personal Insurance. CDEI. Retrieved November 16, 2020, from https://www.gov.uk/government/publications/cdei-publishes-its-first-series-of-three-snapshot-papers-ethical-issues-in-ai/snapshot-paper-ai-and-personal-insurance.

CDEI. (2020, June). AI Barometer Report. CDEI. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/894170/CDEI_AI_Barometer.pdf.

Center for Security and Emerging Technology. (2019). Chinese Public AI R&D: Provisional findings. Georgetown University. https://cset.georgetown.edu/wp-content/uploads/Chinese-Public-AI-RD-Spending-Provisional-Findings-2.pdf.

Chiappa, S., & Gillam, T. P. (2018). Path-Specific Counterfactual Fairness. arXiv.org.

Chinese State Council. (2017, July 30). China's new generation of artificial intelligence development plan. Foundation for Law & International Affairs. Retrieved November 16, 2020, from https://flia.org/notice-state-council-issuing-new-generation-artificial-intelligence-development-plan/.

Crane, M. (2018, June 27). *In surveys, people say they'll pay twice what they're actually willing to spend*. Phys.org. Retrieved November 16, 2020, from https://phys.org/news/2018-06-surveys-people-theyll-theyre.html.

Crawford, K. (2018, July 23). *You and AI—the politics of AI* [Video]. YouTube. https://www.youtube.com/watch?v=HPopJb5aDyA.

Dataiku. (2019). *A glance at ADA: Aviva's algorithmic decision agent*. Dataiku. Retrieved November 16, 2020, from https://www.dataiku.com/stories/a-glance-at-ada-avivas-algorithmic-decision-agent/.

DeepMind. (n.d.). DeepMind ethics & society principles. DeepMind. Retrieved November 16, 2020, from https://deepmind.com/about/ethics-and-society.

Dellot, B. (2020, February 13). Reflecting on the CDEI's snapshot on AI & personal insurance. *Centre for Data Ethics and Innovation Blog*. https://cdei.blog.gov.uk/2020/02/13/reflecting-on-the-cdeis-snapshot-on-ai-personal-insurance/.

Deloitte. (2017). *From Mystery to Mastery: Unlocking the business value of artificial intelligence in the insurance industry.* Deloitte. https://www2.deloitte.com/content/dam/Deloitte/xe/Documents/financial-services/Artificial-Intelligence-in-Insurance.pdf.

Digital Fineprint. (2017, August 21). Working with Metlife. Digital Fineprint. https://digitalfineprint.com/2017/08/working-with-metlife/.

Dignan, L. (2017, June 26). *How Haven Life uses AI, machine learning to spin new life out of long-tail data*. ZDNet. https://www.zdnet.com/article/how-haven-life-uses-ai-machine-learning-to-spin-new-life-out-of-long-tail-data/.

Dong, W., Li, J., Yao, R., Li, C., Yuan, T., & Wang, L. (2016). *Characterizing driving styles with deep learning.* arXiv.org.

Doyle, D., & Groendyke, C. (2019). Using neural networks to price and hedge variable annuity guarantees. *Risks*.

EIOPA. (2019). Consultation paper on the opinion on the 2020 review of Solvency II. EIOPA. https://www.eiopa.europa.eu/content/consultation-paper-opinion-2020-review-solvency-ii_en.

Enterprise Big Data Framework. (2019, January 9). Data types: Structured vs. unstructured data. Enterprise Big Data Framework. https://www.bigdataframework.org/data-types-structured-vs-unstructured-data/.

Ethos Life. (2020, September 25). Ethos Life. Retrieved November 16, 2020, from https://www.ethoslife.com/life/ethos-life-insurance/.

European Commission. (2017, May 10). A connected digital single market for all. European Commission. Retrieved November 16, 2020, from https://eur-lex.europa.eu/resource.html?uri=cellar:a4215207-362b-11e7-a08e-01aa75ed71a1.0001.02/DOC_1&format=PDF.

European Commission. (2018a, December 7). Annex to the coordinated plan on artificial intelligence. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56017.

European Commission. (2018b, April 25). Communication artificial intelligence for Europe. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/digital-single-market/en/news/communication-artificial-intelligence-europe.

European Commission. (2018c, December 7). Coordinated plan on artificial intelligence. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=56018.

European Commission. (2018d, April 10). Declaration of cooperation on AI. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=50951.

European Commission. (2018e, June 6). European budget: Commission proposes €9.2 billion investment in first ever digital programme. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/commission/presscorner/detail/en/IP_18_4043.

European Commission. (2018f). The GDPR: New opportunities, new obligations. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/info/sites/info/files/data-protection-factsheet-sme-obligations_en.pdf.

European Commission. (2019, April 8). Ethics guidelines for trustworthy AI. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai.

European Commission. (2020a, June 15). European defence fund: €205 million to boost the EU's strategic autonomy and industrial competitiveness. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/commission/presscorner/detail/en/IP_20_1053.

European Commission. (2020b). European enterprise survey on the use of technologies based on artificial intelligence. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/digital-single-market/en/news/european-enterprise-survey-use-technologies-based-artificial-intelligence.

European Commission. (2020c). Non-discrimination. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/info/aid-development-cooperation-fundamental-rights/your-rights-eu/know-your-rights/equality/non-discrimination_en.

European Commission. (2020d, September 23). Shaping Europe’s digital future. European Commission. Retrieved November 16, 2020, from https://ec.europa.eu/digital-single-market/en/artificial-intelligence.

European Council. (2017, October 19). European Council meeting (19 October 2017)—Conclusions. European Council: https://data.consilium.europa.eu/doc/document/ST-14-2017-INIT/en/pdf.

European Parliamentary Research Service. (2020, June). The impact of the General Data Protection Regulation (GDPR) on Artificial Intelligence. European Parliament. https://www.europarl.europa.eu/RegData/etudes/STUD/2020/641530/EPRS_STU(2020)641530_EN.pdf .

European Union. (2016a). Charter of Fundamental Rights of the European Union. *Official Journal of the European Union*, 389–405.

European Union. (2016b). Consolidated Versions of the Treaty on European Union and the Treaty on the Functioning of the European Union. *Official Journal of the European Union*, 1–388.

European Union. (2016c). The General Data Protection Regulation. *Official Journal of the European Union*.

FAT ML. (n.d.). Fairness, accountability, and transparency in machine learning. FAT ML. Retrieved November 16, 2020, from https://www.fatml.org/.

FCA. (2015, May 12). *Fair treatment of customers*. FCA. Retrieved November 16, 2020, from https://www.fca.org.uk/firms/fair-treatment-customers.

FCA. (2016, September 21). *FS16/5: Call for inputs on big data in retail general insurance*. FCA. Retrieved November 16, 2020, from https://www.fca.org.uk/publications/feedback-statements/fs16-5-call-inputs-big-data-retail-general-insurance.

FCA. (2019). *FS19/04: Fair pricing in financial services: Summary of responses and next steps.* FCA.

FCA. (2020a, January 23). Financial Services AI Public Private Forum. FCA. Retrieved November 16, 2020, from https://www.fca.org.uk/news/news-stories/financial-services-ai-public-private-forum.

FCA. (2020b). *Fit and Proper test for Employees and Senior Personnel Sourcebook.* FCA. Retrieved November 16, 2020, from https://www.handbook.fca.org.uk/handbook/FIT.pdf.

Fidelity. (2020). Fidelity Go. Fidelity. Retrieved November 16, 2020, from https://www.fidelity.com/managed-accounts/fidelity-go/overview.

Financial Reporting Council. (2016, December). *Technical Actuarial Standard 100: Principles for Technical Actuarial Work.* Financial Reporting Council. https://www.frc.org.uk/getattachment/b8d05ac7-2953-4248-90ae-685f9bcd95bd/TAS-100-Principles-for-Technical-Actuarial-Work-Dec-2016.pdf.

Franke, U. (2019, December 18). Not smart enough: The poverty of European military thinking on artificial intelligence. European Council on Foreign Relations. Retrieved November 16, 2020, from https://www.ecfr.eu/publications/summary/not_smart_enough_poverty_european_military_thinking_artificial_intelligence.

FT Partners Research. (2020, July). *Q2 2020 Quarterly Insurtech Insights.* FT Partners Research. https://ftpartners.docsend.com/view/kbdj58rp6qnfg5jz.

Future of Privacy Forum. (2020, January 13). It’s raining privacy bills: An overview of the Washington State Privacy Act and other introduced bills. Future of Privacy Forum. Retrieved November 16, 2020, from https://fpf.org/2020/01/13/its-raining-privacy-bills-an-overview-of-the-washington-state-privacy-act-and-other-introduced-bills/.

G20. (2019, June 9). G20 Ministerial Statement on Trade and Digital Economy. Ministry of Foreign Affairs of Japan: https://www.mofa.go.jp/files/000486596.pdf.

G7. (2018, June). Common Vision for the Future of Artificial Intelligence. Government of Canada. https://www.international.gc.ca/world-monde/assets/pdfs/international_relations-relations_internationales/g7/2018-06-09-artificial-intelligence-artificielle-en.pdf.

Gao, G. & Wüthrich, M. V. (2019). Convolutional neural network classification of telematics car driving data. *Risks*.

Gardam, T. (2019, March 8). *Data science and the case for ethical responsibility*. Ada Lovelace Institute. Retrieved November 16, 2020, from https://www.adalovelaceinstitute.org/the-culture-of-computing-and-the-case-for-ethical-responsibility/.

Good Shopping Guide. (n.d.). Ethical insurance ethical comparison table. Good Shopping Guide. Retrieved November 16, 2020, from https://thegoodshoppingguide.com/subject/ethical-insurance/.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning. Vol. 1.* Cambridge: MIT Press.

Google. (n.d.). Artificial intelligence at Google: Our principles. Google AI. Retrieved November 16, 2020, from https://ai.google/principles/.

GPT-3. (2020, September 8). A robot wrote this entire article. Are you scared yet, human? *The Guardian*. https://www.theguardian.com/commentisfree/2020/sep/08/robot-wrote-this-article-gpt-3.

Gregory, B. (2018). *Predicting customer churn: Extreme gradient boosting with temporal data.* arXiv.org.

Grootendorst, M. (2019, September 26). *Validating your machine learning model—Going beyond k-Fold Cross-Validation*. Towards Data Science. Retrieved November 16, 2020, from https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7.

GroupSolver. (2015, August 27). *Avoiding groupthink in focus group discussion*. Medium. Retrieved November 16, 2020, from https://medium.com/groupsolver/avoiding-groupthink-in-focus-group-discussion-f8fd6c7a8091.

Hainaut, D. (2018). A Neural-Network Analyzer for Mortality Forecast. *ASTIN Bulletin: The Journal of the IAA*, 481–508.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, interference, and prediction, Second Edition.* Springer-Verlag. Retrieved from https://web.stanford.edu/~hastie/ElemStatLearn/.

Hejazi, S. A., & Jackson, K. R. (2016). *A neural network approach to efficient valuation of large portfolios of variable annuities.* Insurance Mathematics & Economics.

HESA. (2018, January 11). *Higher education student statistics: U.K., 2016/17—Qualifications achieved*. Higher Education Student Statistics. https://www.hesa.ac.uk/news/11-01-2018/sfr247-higher-education-student-statistics/qualifications.

IBM. (2019). *Everyday ethics for artificial intelligence.* IBM. https://www.ibm.com/watson/assets/duo/pdf/everydayethics.pdf.

IBM. (2020, October 02). *AI fairness 360*. IBM. Retrieved November 16, 2020, from https://aif360.mybluemix.net/.

ICO. (2019, June 3). *Project ExplAIn: Interim Report.* ICO. https://ico.org.uk/media/about-the-ico/documents/2615039/project-explain-20190603.pdf.

ICO. (2019a). *AI Auditing Framework*. ICO. Retrieved November 16, 2020, from https://ico.org.uk/about-the-ico/news-and-events/ai-auditing-framework/.

ICO. (2019b, April 12). Automated decision making: The role of meaningful human reviews. ICO. Retrieved November 16, 2020, from https://ico.org.uk/about-the-ico/news-and-events/ai-blog-automated-decision-making-the-role-of-meaningful-human-reviews/.

ICO. (2019c, October 23). Data protection impact assessments and AI. ICO. Retrieved November 16, 2020, from https://ico.org.uk/about-the-ico/news-and-events/ai-blog-data-protection-impact-assessments-and-ai/.

ICO. (2019d, October 15). Enabling access, erasure, and rectification rights in AI systems. ICO. Retrieved November 16, 2020, from https://ico.org.uk/about-the-ico/news-and-events/ai-blog-enabling-access-erasure-and-rectification-rights-in-ai-systems/.

ICO. (2019e, May 22). Guide to the General Data Protection Regulation. ICO. Retrieved November 16, 2020, from https://ico.org.uk/for-organisations/guide-to-data-protection/guide-to-the-general-data-protection-regulation-gdpr/.

ICO. (2019f, June 3). *Project Explain—Interim Report*. ICO. https://ico.org.uk/media/about-the-ico/documents/2615039/project-explain-20190603.pdf.

ICO. (2020a, May 20). Explaining decisions made with AI. ICO. https://ico.org.uk/media/for-organisations/guide-to-data-protection/key-data-protection-themes/explaining-decisions-made-with-artificial-intelligence-1-0.pdf.

ICO. (2020b, February 14). Guidance on AI auditing framework: Draft guidance for consultation. ICO. https://ico.org.uk/media/2617219/guidance-on-the-ai-auditing-framework-draft-for-consultation.pdf.

ICO. (2020c, July 10). Information rights at the end of the transition period: Frequently asked questions. ICO. https://ico.org.uk/media/for-organisations/documents/brexit/2617110/information-rights-and-brexit-faqs-v2_3.pdf.

IFoA. (2015, July). *APS X2: Review of actuarial work.* IFoA. https://www.actuaries.org.uk/system/files/documents/pdf/20150122-aps-x2-final-version.pdf.

IFoA. (2019a, June 26). *Certificate in data science*. IFoA. Retrieved November 16, 2020, from https://www.actuaries.org.uk/news-and-insights/news/data-science-credential.

IFoA. (2019b, April). *The actuaries' code.* IFoA. https://www.actuaries.org.uk/system/files/field/document/2019_04_05%20Guidance%20FINAL.pdf.

IFoA. (n.d.). *Professional standards directory*. IFoA. https://www.actuaries.org.uk/upholding-standards/standards-and-guidance/professional-standards-directory.

insight2impact. (2018, October). *Inclusive Insurance enhanced through the use of client data.* i2ifacility. Retrieved November 16, 2020, from https://i2ifacility.org/system/documents/files/000/000/075/original/Inclusive_insurance_enhanced_through_the_use_of_client_data_i2i_FSDA_October_2018_SINGLE.pdf?1540279096.

Institute for Ethical AI & Machine Learning. (n.d.). *The responsible machine learning principles*. Institute for Ethical AI & Machine Learning. Retrieved November 16, 2020, from https://ethical.institute/principles.html#commitment-2.

Institute of Electrical and Electronics Engineers. (2019). *Ethically aligned design, first edition.* Institute of Electrical and Electronics Engineers.

Insurance Europe. (2020). *Artificial Intelligence (AI)—Views of the European insurance industry.* Insurance Europe. https://www.insuranceeurope.eu/sites/default/files/attachments/Views%20of%20EU%20insurance%20industry%20on%20AI.pdf.

Insurance Europe and AMICE. (2019, September 30). *Proposals for making proportionality work in Solvency II*. Insurance Europe. https://www.insuranceeurope.eu/sites/default/files/attachments/Proposals%20for%20making%20proportionality%20work%20in%20Solvency%20II.pdf.

Janković, M., Savić, A., Novičić, M., & Popović, M. (2018). *Deep learning approaches for human activity recognition using wearable technology.* MedPodml.

Johnson, J. M., & Khoshgoftaar, T. M. (2019). Medicare fraud detection using neural networks. *Journal of Big Data*.

Joseph, A. (2019, May 24). *Opening the machine learning black box*. Bank Underground. Retrieved November 16, 2020, from https://bankunderground.co.uk/2019/05/24/opening-the-machine-learning-black-box/.

Joseph, A. (2020, July). *Staff Working Paper No. 784: Parametric inference with universal function approximators.* Bank of England. https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/shapley-regressions-a-framework-for-statistical-inference-on-machine-learning-models.pdf?la=en&hash=C2D404E60A877EC8623240173EA3D93303B26080.

Karpathy, A. (2015, May 21). *The unreasonable effectiveness of recurrent neural networks*. Andrej Karpathy blog. https://karpathy.github.io/2015/05/21/rnn-effectiveness/.

Kipling, R., & Iyer, U. J. (2008). Cultural Differences in Perceptions of Fairness in Organizational Contexts. In C. Wankel, *21st Century Management: A Reference Handbook* (pp. 221–229). SAGE Publications, Inc.

Kirlidog, M., & Asuk, C. (2012). A Fraud Detection Approach with Data Mining in Health Insurance. *Procedia—Social and Behavioural Sciences*, 989–994.

Kopczyk, D. (2019). *Proxy modeling in life insurance companies with the use of machine learning algorithms.* Quantee, Research Department.

Ladder Life. (2020, September 25). Ladder Life. Retrieved November 16, 2020, from https://www.ladderlife.com/.

Le, J. (2019, October 31). *Recommendation System Series Part 2: The 10 Categories of Deep Recommendation Systems That Academic Researchers Should Pay Attention To*. Towards Data Science. https://towardsdatascience.com/recommendation-system-series-part-2-the-10-categories-of-deep-recommendation-systems-that-189d60287b58.

LifeScore Labs. (2020, October 1). *MyLifeScore360*. LifeScore Labs. Retrieved November 16, 2020, from https://www.lifescore360.com/mylifescore360/.

Lindholm, M., Richman, R., Tsanakas, A., & Wüthrich, M. V. (2020). Discrimination-Free Insurance Pricing. *SSRN Electronic Journal*.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., & Puri, a. R. (2019). *Bias Mitigation Post-Processing for Individual and Group Fairness.* IEEE.

Loisel, S., P. P., & Tsai, C.-H. J. (2019). *Applying economic measures to lapse risk management.* ArXiv.org.

Lundberg, S. (2020, October 01). *Shap*. GitHub. Retrieved November 16, 2020, from https://github.com/slundberg/shap.

Madras, D., Pitassi, T., & Zemel, R. (2018). Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. *32nd Conference on Neural Information Processing Systems.* Montréal: NeurIPS.

Maier, M., Carlotto, H., Sanchez, F., Balogun, S., & Merritt, S. (2019). Transforming Underwriting in the Life Insurance Industry. *Proceedings of the AAAI Conference on Artificial Intelligence.* AAAI. Retrieved November 16, 2020, from https://www.aaai.org/ojs/index.php/AAAI/article/view/4985/4858.

Malhan, J. (2020, May 28). *reCaptcha : How we are training Google's AI by proving 'I am not a robot.'* Medium. Retrieved November 16, 2020, from https://medium.com/@jyotimalhan/recaptcha-how-we-are-training-googles-ai-by-proving-i-am-not-a-robot-76651fbbe26a.

Manulife. (2018, June 21). *Manulife announces plan to transform its Canadian business*. Manulife. Retrieved November 16, 2020, from http://manulife.force.com/Master-Article-Detail?content_id=a0Qf200000Jq5SWEAZ&ocmsLang=en_US.

Marsh, S. (2020, August 24). Councils scrapping use of algorithms in benefit and welfare decisions. *The Guardian*. Retrieved November 16, 2020, from https://www.theguardian.com/society/2020/aug/24/councils-scrapping-algorithms-benefit-welfare-decisions-concerns-bias.

Meharabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). *A Survey on Bias and Fairness in Machine Learning.* arXiv.org.

Microsoft. (n.d. a). *AI Principles*. Microsoft. Retrieved November 16, 2020, from https://www.microsoft.com/en-us/ai/responsible-ai?activetab=pivot1%3aprimaryr6.

Microsoft. (n.d. b). *FATE: Fairness, Accountability, Transparency and Ethics in AI*. Microsoft. Retrieved November 16, 2020, from https://www.microsoft.com/en-us/research/theme/fate/.

Milhaud, X., Loisel, S., & Maume-Deschamps, V. (2010). Surrender Triggers in Life Insurance: Classification and Risk Predictions. *Research Papers in Economics*.

MIT Technology Review. (2019, May 31). *Why does Beijing suddenly care about AI ethics? MIT Technology Review*. Retrieved November 16, 2020, from https://www.technologyreview.com/2019/05/31/135129/why-does-china-suddenly-care-about-ai-ethics-and-privacy/.

Moisejevs, I. (2020, October 07). *Evasion attacks on machine learning (or 'adversarial examples')*. *Towards Data Science*. Retrieved November 16, 2020, from https://towardsdatascience.com/evasion-attacks-on-machine-learning-or-adversarial-examples-12f2283e06a1.

Molnar, C. (2019). *Interpretable machine learning—A guide for making black box models explainable.*

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

Morik, K., & Köpcke, H. (2004). Analysing customer churn in insurance data—A case study. *Knowledge Discovery in Databases: PKDD 2004, 8th European Conference on Principles and Practice of Knowledge Discovery in Databases.* Pisa: Lecture Notes in Computer Science.

Murillo, A., & Ricardo, A. (2016). *High-frequency trading strategy based on deep neural networks.*

National Association of Insurance Commissioners. (2020). *National Association of Insurance Commissioners (NAIC) principles on artificial intelligence (AI).* National Association of Insurance Commissioners. https://content.naic.org/sites/default/files/inline-files/AI%20principles%20as%20Adopted%20by%20the%20TF_0807.pdf.

New York State Department of Financial Services. (2019, January 18). *Insurance Circular Letter No. 1 (2019): RE: Use of External Consumer Data and Information Sources in Underwriting for Life Insurance*. New York State Department of Financial Services. Retrieved November 16, 2020, from https://www.dfs.ny.gov/industry_guidance/circular_letters/cl2019_01.

Nielson, M. A. (2015). *Neural networks and deep learning.* Determination Press. Retrieved November 16, 2020, from http://neuralnetworksanddeeplearning.com/.

OECD. (2019). *Recommendation of the Council on Artificial Intelligence.* Retrieved November 16, 2020, from https://legalinstruments.oecd.org/api/print?ids=648&lang=en.

Open Data Institute. (2018, February 12). *ODI survey reveals British consumer attitudes to sharing personal data*. Open Data Institute. Retrieved November 16, 2020, from https://theodi.org/article/odi-survey-reveals-british-consumer-attitudes-to-sharing-personal-data/.

OpenAI. (n.d.). *OpenAI*. OpenAI Charter. Retrieved November 16, 2020, from https://openai.com/charter/.

Palacio, S. M. (2019). Abnormal pattern prediction: Detecting fraudulent insurance property claims with semi-supervised machine-learning. *Data Science Journal*.

Partnership On AI (e.g., Apple, Microsoft, Google). (n.d.). *Partnership on AI—Tenets*. Partnership on AI. Retrieved November 16, 2020, from https://www.partnershiponai.org/tenets/.

Perez, S. (2016, March 24). Microsoft silences its new AI bot Tay, after Twitter users teach it racism. Tech Crunch. Retrieved November 16, 2020, from https://techcrunch.com/2016/03/24/microsoft-silences-its-new-a-i-bot-tay-after-twitter-users-teach-it-racism/?guccounter=1&guce_referrer=aHR0cHM6Ly93d3cuZ29vZ2xlLmNvLnVrLw&guce_referrer_sig=AQAAACG7CYwDi2aPegWWMwwTIkzldttF70y3EARf1xMs_2oNU3dUEph7mBhBUpYwq.

Phelan, E., Murray, K., Tucker, G., & Comerford, E. (2019, January). Milliman data science survey*.* Milliman. Retrieved November 16, 2020, from https://www.milliman.com/-/media/milliman/importedfiles/uploadedfiles/insight/2019/milliman-data-science-survey.ashx.

Proudman, J. (2019, June 4). *Managing machines: the governance of artificial intelligence.* Bank of England. Retrieved November 16, 2020, from https://www.bankofengland.co.uk/-/media/boe/files/speech/2019/managing-machines-the-governance-of-artificial-intelligence-speech-by-james-proudman.pdf?la=en&hash=8052013DC3D6849F91045212445955245003AD7D.

Reuters. (2019, December 22). China should step up regulation of artificial intelligence in finance, think tank says. Reuters. Retrieved November 16, 2020, from https://www.reuters.com/article/us-china-economy-artificial-intelligence/china-should-step-up-regulation-of-artificial-intelligence-in-finance-think-tank-says-idUSKBN1YQ045.

Richman, R. (2020a). AI in actuarial science—a review of recent advances—part 1. *Annals of Actuarial Science*, 1–23.

Richman, R. (2020b). AI in actuarial science—a review of recent advances—part 2. *Annals of Actuarial Science*, 1–29.

Richterich, A. (2018). *The Big Data Agenda.* Critical Digital and Social Media Studies.

Rovatsos, M., Mittelstadt, B., & Koene, A. (2019, July 19). Landscape summary: bias in algorithmic decision-making*.* Centre for Data Ethics and Innovation. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/819055/Landscape_Summary_-_Bias_in_Algorithmic_Decision-Making.pdf.

RSA. (2018). Artificial intelligence: Real public engagement*.* RSA. https://www.thersa.org/globalassets/pdfs/reports/rsa_artificial-intelligence---real-public-engagement.pdf.

RSS. (2020, October 28). *Data science ethics: The role of practitioners [Video]*. YouTube. https://www.youtube.com/watch?v=mJUwN6wtmak&feature=youtu.be.

RSS and IFoA. (2019, October 7). A guide for ethical data science. IFoA. https://www.actuaries.org.uk/system/files/field/document/An%20Ethical%20Charter%20for%20Date%20Science%20WEB%20FINAL.PDF.

Saravia, E. (2018, August 24). Deep learning for NLP: An overview of recent trends. Medium. Retrieved November 16, 2020, from https://medium.com/dair-ai/deep-learning-for-nlp-an-overview-of-recent-trends-d0d8f40a776d.

Saxena, N. A. (2019). Perceptions on fairness. *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics and Society* (pp. 537–538). ACM.

Segal, E. (2020, April 26). AI-based hyper personalisation for enhanced customer experience. IBM. Retrieved November 16, 2020, from https://developer.ibm.com/recipes/tutorials/aibased-hyper-personalisation-for-enhanced-customer-experience/.

Shoham, Y., Perrault, R., Brynjolfsson, E., Clark, J., Manyika, J., Niebles, J. C., & Terah Lyons, J. (2018). *The AI Index 2018 Annual Report.* Stanford University, Stanford, CA. AI Index Steering Committee, Human-Centered AI Initiative.

Siemes, T. (2016). *Churn prediction models tested and evaluated in the Dutch indemnity industry.* Open Universiteit Nederland.

Sigmoidal. (n.d.). *Natural Language Processing Algorithms (NLP AI)*. Sigmoidal. Retrieved November 16, 2020, from https://sigmoidal.io/boosting-your-solutions-with-nlp/.

Sillars, J. (2020, September 22). *Insurance renewals face curbs to reverse £1.2bn 'loyalty penalty.'* Sky News. Retrieved November 16, 2020, from https://news.sky.com/story/automatic-insurance-renewals-face-curbs-to-reverse-1-2bn-loyalty-penalty-12078428.

Solomon, S., & Abelson, J. (2012). *Why and When Should We Use Public Deliberation?* Hastings Cent Rep.

Squires, G. D., & Velez, W. (1988). Insurance redlining and the process of discrimination. *The Review of Black Political Economy*, pp. 63–75.

State of California Department of Justice. (2018). California Consumer Privacy Act of 2018.

Stefano, B., & Gisella, F. (2001). Insurance fraud evaluation: a fuzzy expert system. *10th IEEE International Conference on Fuzzy Systems.* IEEE.

Subramanya, K. B., & Somani, A. (2017). Enhanced feature mining and classifier models to predict customer churn for an e-retailer. *2017 7th International Conference on Cloud Computing, Data Science & Engineering—Confluence.* Noida: IEEE.

Suss, J., & Treitel, H. (2019, October). *Staff Working Paper No. 831: Predicting bank distress in the U.K. with machine learning.* Bank of England. https://www.bankofengland.co.uk/-/media/boe/files/working-paper/2019/predicting-bank-distress-in-the-uk-with-machine-learning.pdf?la=en&hash=463498FF652EB9F61D55A4553D2813E2F8AC8F33.

Taylor, G., & McGuire, G. (2004). *Loss reserving with GLMs: A case study.*

TensorFlow. (2020, October 1). *Model understanding with the What-If Tool Dashboard*. TensorFlow. Retrieved November 16, 2020, from https://www.tensorflow.org/tensorboard/what_if_tool.

The Alan Turing Institute. (2018). *A right to explanation*. The Alan Turing Institute. Retrieved November 16, 2020, from https://www.turing.ac.uk/research/impact-stories/a-right-to-explanation.

The Atlantic. (2018, June 7). People are changing the way they use social media. *The Atlantic*. Retrieved November 16, 2020, from https://www.theatlantic.com/technology/archive/2018/06/did-cambridge-analytica-actually-change-facebook-users-behavior/562154/.

The Beijing Academy of Artificial Intelligence. (2019, May 28). *Beijing AI principles*. The Beijing Academy of Artificial Intelligence. Retrieved November 16, 2020, from https://www.baai.ac.cn/news/beijing-ai-principles-en.html.

The Consumer Financial Protection Bureau. (2019, September 10). *CFPB issues policies to facilitate compliance and promote innovation*. The Consumer Financial Protection Bureau. Retrieved November 16, 2020, from https://www.consumerfinance.gov/about-us/newsroom/bureau-issues-policies-facilitate-compliance-promote-innovation/.

The Forum for Ethical AI. (2019). *Democratising decisions about technology: A toolkit.* RSA. Retrieved November 16, 2020, from https://www.thersa.org/reports/democratising-decisions-technology-toolkit.

The Geneva Association. (2020). *Promoting responsible artificial intelligence in insurance.* Retrieved November 16, 2020, from https://www.genevaassociation.org/research-topics/digitalization/promoting-responsible-artificial-intelligence-insurance-research.

The Guardian. (2018, March 17). *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*. *The Guardian*. Retrieved November 16, 2020, from https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election.

The Guardian. (2019, December 30). *U.K. ethical consumer spending hits record high, report shows*. *The Guardian*. Retrieved November 16, 2020, from https://www.theguardian.com/environment/2019/dec/30/uk-ethical-consumer-spending-hits-record-high-report-shows.

The Royal Society. (2017, April). *Public views of machine learning—findings from public research and engagement conducted on behalf of the Royal Society.* The Royal Society. https://royalsociety.org/~/media/policy/projects/machine-learning/publications/public-views-of-machine-learning-ipsos-mori.pdf.

The Sun. (2018, January 22). *Motorists fork out £1,000 more to insure their cars if their name is Mohammed*. *The Sun*. Retrieved November 16, 2020, from https://www.thesun.co.uk/motors/5393978/insurance-race-row-john-mohammed/.

The Telegraph. (2016, November 2). *Facebook blocks admiral from using profiles to price car insurance*. The *Telegraph*. Retrieved November 16, 2020, from https://www.telegraph.co.uk/technology/2016/11/02/facebook-blocks-admiral-app-that-priced-car-insurance-based-on-s/.

The White House. (2018, May 10). *AI for the American people*. The White House. Retrieved November 16, 2020, from https://www.whitehouse.gov/ai/.

The White House Office of Science and Technology Policy. (2018, May 10). *Summary of the 2018 White House Summit on AI for American Industry.* The White House. Retrieved November 16, 2020, from https://www.whitehouse.gov/wp-content/uploads/2018/05/Summary-Report-of-White-House-AI-Summit.pdf?latest.

The White House Office of Science and Technology Policy. (2019, January 7). *Memorandum for the heads of executive departments and agencies.* The White House. Retrieved November 16, 2020, from https://www.whitehouse.gov/wp-content/uploads/2020/01/Draft-OMB-Memo-on-Regulation-of-AI-1-7-19.pdf.

Théate, T., & Ernst, D. (2020). *An Application of Deep Reinforcement Learning to Algorithmic Trading.* ArXiv.org.

Tomas, J., & Planchet, F. (2014). Constructing entity specific projected mortality table: adjustment to a reference. *European Actuarial Journal*, 247–279.

U.K. Government. (2018, November 20). *Stellar new board appointed to lead world-first Centre for Data Ethics and Innovation*. U.K. Government. Retrieved November 16, 2020, from https://www.gov.uk/government/news/stellar-new-board-appointed-to-lead-world-first-centre-for-data-ethics-and-innovation.

UNESCO. (2020, September 7). *First draft of the recommendation on the ethics of artificial intelligence.* UNESDOC Digital Library. Retrieved November 16, 2020, from https://unesdoc.unesco.org/ark:/48223/pf0000373434.

University of Montreal. (2018). *Montreal declaration for a responsible development of artificial intelligence.* Retrieved November 16, 2020, from https://www.montrealdeclaration-responsibleai.com/the-declaration.

Vanguard. (2020). *Vanguard Digital Advisor*. Retrieved November 16, 2020, from https://investor.vanguard.com/financial-advisor/digital-advisor.

Vartak, P., & Jain, D. (2019). *AI-driven transformation in the insurance Industry.* A Larsen & Toubro Group Company. Retrieved November 16, 2020, from https://www.lntinfotech.com/wp-content/uploads/2019/01/AI-driven-Transformation-in-the-Insurance-Industry.pdf.

Veale, M., Binns, R., & Edwards, L. (2018, October 15). Algorithms that remember: model inversion attacks and data protection law. *Philosophical Transactions of the Royal Society A 376*.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep learning for computer vision: A brief review. *Computational Intelligence and Neuroscience*.

WHO. (2020, October 9). *Track 2: Health literacy and health behaviour*. World Health Organization. Retrieved November 16, 2020, from https://www.who.int/healthpromotion/conferences/7gchp/track2/en/#:~:text=Health%20Literacy%20has%20been%20defined,pamphlets%20and%20successfully%20make%20appointments.

World Economic Forum. (2018). *How to prevent discriminatory outcomes in machine learning.*

Xu, W., Wang, S., Zhang, D., & Yang, B. (2011). Random rough subspace based neural network ensemble for insurance fraud detection. *2011 Fourth International Joint Conference on Computational Sciences and Optimization.* Yunnan, China: IEEE.

Yao, J. (2013). *Generalized Linear Models for Non-Life Pricing—Overlooked Facts and Implications (A Report from GIRO Advanced Pricing Techniques).* Institute and Faculty of Actuaries. https://www.actuaries.org.uk/system/files/documents/pdf/c6-paper.pdf.

YouGov. (2019). *Better safe than sorry—YouGov analysis of Brits' attitudes to insurance.* YouGov.

Youyou, W., Kosinski, M., & Stillwell, D. (2015, January 12). Computer-based personality judgments are more accurate than those made by humans. *PNAS*, pp. 1036–1040.

Yu, D., & Deng, L. (2015). *Automatic Speech Recognition.* Signals and Communication Technology.

Zarifis, A., Holland, C. P., & Milne, A. (2019). *Evaluating the impact of AI on insurance: The four emerging AI-and data-driven business models.* Emerald Open Research.

Zhang, B. H., Lemoine, B., & Mitchell, M. (2018). *Mitigating Unwanted Biases with Adversarial Learning.* arXiv.org.

Zhang, Z., Zohren, S., & Roberts, S. (2019). *Deep Reinforcement Learning for Trading.* arXiv.org.

Zurich. (2020, March 16). *Zurich U.K. partners with Carpe Data to combat fraud*. Zurich. Retrieved November 16, 2020, fromhttps://www.zurich.co.uk/en/about-us/media-centre/general-insurance-news/2020/zurich-uk-partners-with-carpe-data-to-combat-fraud.

Milliman is among the world's largest providers of actuarial and related products and services. The firm has consulting practices in life insurance and financial services, property & casualty insurance, healthcare, and employee benefits. Founded in 1947, Milliman is an independent firm with offices in major cities around the globe.

milliman.com

CONTACT

David Burston
david.burston@milliman.com

Charlie Howell
charlie.howell@milliman.com

Andrew Gilchrist
andrew.gilchrist@milliman.com

Philip Simpson
philip.simpson@milliman.com

The materials in this document represent the opinion of the authors and are not representative of the views of Milliman, Inc. Milliman does not certify the information, nor does it guarantee the accuracy and completeness of such information. Use of such information is voluntary and should not be relied upon unless an independent review of its accuracy and completeness has been performed. Materials may not be reproduced without the express consent of Milliman.