DATA ANALYTICS PRACTICE COMMITTEE
Technical Paper: Automated Decision-Making Systems
October 2020

Contents

A. Purpose and status of Technical Paper

This Technical Paper has been prepared by the Data Analytics Practice Committee of the Actuaries Institute, for the purpose of assisting Members involved in the design, creation, management or assessment of automated decision-making systems.

This Technical Paper does not represent a Professional Standard or Practice Guideline of the Institute. The information contained in this Technical Paper is commentary and general information only. It is not mandatory for Members to consider this Technical Paper in their work. This Technical Paper does not constitute legal advice. Any interpretation or commentary within the Technical Paper regarding specific legislative or regulatory requirements reflects the expectations of the Institute but does not guarantee compliance under applicable legislation or regulations. Accordingly, Members should seek clarification from the relevant regulator and/or seek legal advice in the event they are unsure or require specific guidance regarding their legal or regulatory obligations.

A draft version of this Technical Paper was circulated to Members in June 2020, soliciting feedback. We thank those individuals and groups who took the time to provide feedback to us. This is the first final version of this Technical Paper, which incorporates feedback received during this Member consultation.

Ongoing feedback from Members is encouraged; any feedback should be directed to the Data Analytics Practice Committee via actuaries@actuaries.asn.au.

B. Introduction and scope

This Technical Paper outlines a set of principles and practices for Members involved in the design, creation, management or assessment of automated decisioning systems to consider in the course of their work. With this, we hope that regulators, companies, governments and society at large may then have greater confidence in the deployment of automated decisioning systems that Members are involved in developing and maintaining.

We hope that this set of recommendations, along with application of the Code of Conduct, allows Members to have confidence that they have identified, to the extent reasonably possible, that the development process, outcomes and ongoing management of an automated decisioning system are appropriate, in light of common societal norms and expectations. This includes the profession's traditional role of taking responsibility for equity and fairness of outcomes, balancing the competing needs of various stakeholder groups, and acting in the public interest.

Governance and ethics of automated decisions is a rapidly developing area of academic and policy interest across the globe, with practitioners from a range of

backgrounds actively contributing to the discussion, including (but not limited to) engineers, mathematicians, physicists, data scientists, computer scientists, actuaries, accountants, lawyers, philosophers, political scientists, economists and behavioural scientists.

With such a diverse field and evolving research base, we expect it may be challenging for Members involved in the design, creation, management or assessment of automated decisioning systems to stay abreast of the contemporary literature and modern best practices. This Technical Paper seeks to be a central reference point to assist Members and will be regularly reviewed for relevance, and in response to Member feedback.

The principles and practices set out here are deliberately high level, and not prescriptive nor tailored to any specific context or industry. This is to provide breadth of applicability and relevance but does mean some effort may be required from the reader to apply, adapt or augment the principles and practices to their specific context or industry.

For the purposes of this Technical Paper, an automated decision is defined to be one which occurs in near to real-time, without direct human input at that time. Usually, data is used as an input, and the automated decisioning process consists of software which maps this input data to an output: typically, an item from a set of potential decisions, or some form of prediction or probabilistic estimate. Within this definition, we also include automated systems which provide guidance, prompts, or advice to a final human decision maker. An automated decisioning process may be static (i.e. fixed until deliberately modified) or may be dynamic (i.e. automatically adjusting in line with some specification for adjustment) – both are captured within our definition.

Automated decision-making systems may include predictions from complex data and algorithmic processes, but this need not necessarily be required – simple mappings of data to decisions are commonplace and equally worthy of scrutiny.

It is a matter of professional judgement in which circumstances this Technical Paper is considered relevant, noting it is part of the Professional Governance Material available to a Member as outlined in the Code of Conduct (clause 3.1). In all cases, where conflict may exist between this Technical Paper and either the Code of Conduct or a Professional Standard, the Code of Conduct or the Professional Standard takes precedence.

We note there has been a great deal of global discussion on the need for enhanced governance of automated decisions, and many frameworks of principles published in recent years, including in Australia. Common areas of discussion include purpose or intent, asymmetries of power and/or information, privacy, fairness, vulnerability, discrimination, human rights and autonomy, unintended consequences, transparency and interpretability, accountability, and redress. Our goal here is to take these published frameworks as inspiration, synthesising many of the core concepts for Member

consumption. This forms Section C. We have added a practical set of steps for Members to consider taking as a result of these principles and what we consider to be generally good risk management practices, in Section D. It is this latter step which we hope is most valuable for members, it being absent from many published frameworks today.

B.1 Definitions

In this Technical Paper, certain terms are used consistent with their general meaning under other aspects of actuarial standards, including:

Actuaries' Code: refers to the in-force Code of Conduct issued by the Institute of Actuaries of Australia.

Member: refers to any member of the Institute of Actuaries of Australia.

C. Principles

In this section, we describe several high-level principles for Members to consider in the design, creation, management or assessment of automated decisioning systems. This is intended to assist Members in determining whether automated systems are delivering appropriate outcomes – for individuals, for businesses, and for the public at large. This is fairly high level and principles based; practical guidance is left to Section D.

The principles set out below reflect a synthesis of various principles frameworks for AI and/or automated decisioning processes put forward around the world in recent years. They are not intended to be exhaustive. There are likely to be situations where some of the principles are not deemed relevant, and there are likely to be situations where some of the principles conflict. The role of a Member in this situation is to determine the appropriate balance of each of the principles and any other considerations they deem relevant. Clearly, this is not to be used as a checklist, but as a prompt for deep questions and thinking about the situation at hand.

There are often inherent conflicts between the objectives of the developers and managers of automated decisioning systems, usually governments or corporate entities, and the public. For example, the public may desire a product which is free, but a business requires a profit. Frequently, there may also be segments of each group with competing interests (notably, between different groups of public users). This means the determination of an appropriate balance across and within the principles set out below is a non-trivial task which will generally be domain and context specific.

C.1 Improve Wellbeing

Automated decisioning systems have the potential to create and amplify both positive and negative outcomes for large groups of people. Recognition of this has led to general inclusion amongst many ethical frameworks of the importance of "promoting wellbeing", or "doing good" as a foundational principle. This suggests a definition of "wellbeing" should be constructed, and some monitoring or assurance that the system is actually achieving this aim, once live, should occur.

Wellbeing may be contemplated at both the individual and aggregate level, and these levels may, at times, conflict. Notably, some decisions may seek to reduce the wellbeing of a certain individual or set of individuals in order to improve the overall wellbeing of other individuals in broader society. A simplistic view of incarceration provides an example: jailing a person likely reduces their individual wellbeing, but this may improve the wellbeing of many others, if it is viewed that they are a danger to the community. This form of trade-off between outcomes for different groups in society may require significant contemplation. Such outcomes could be considered both when the system is designed and during its operation.

C.2 Consider Fairness

Wherever a decision is being made, the fairness of that decision could be questioned. Automated systems are no different. Hence, careful consideration of fairness is an essential part of the design of an automated decisioning system.

Consideration of fairness of automated decisioning systems is often an extremely complex task which has spawned a large field of academic research. Whilst the recent literature on the subject has primarily focussed on issues of discrimination, fairness is a substantially broader construct of which discrimination is but one theme. Members are encouraged to take a broad rather than a narrow view of the term "fairness", in considering its application. It is notable that trade-offs amongst individually appealing fairness concepts must almost always be made, and that some notions of fairness may conflict with the accuracy of the decision procedure itself, or other desirable properties of the system.

C.3 Respect Autonomy of Individuals

The right to autonomy, or self-determination, is central in many ethics frameworks. In designing an automated decisioning system, this suggests we should evaluate the degree to which the autonomy of an individual is respected and seek to promote the ability of people to exercise that autonomy. On a practical level, generally good practice is to promote free choices where this is reasonable and possible, and to not seek to remove choices unnecessarily. This may, in some circumstances, require transparency about how and why a system makes decisions, or what its objectives are, in order that a person using the system can either know how to achieve a different outcome by modifying their choices, choose not to engage with an automated system if that is an option, or know how to challenge the basis of a decision they believe to be inappropriate.

C.4 Responsible and Appropriate Use of Data

An automated system usually requires input data, both in the design or training of an algorithm and in the live deployment of the decisioning system. Good practice suggests that data be sourced and used in a manner that actively considers what people would reasonably expect. At a minimum this requires compliance with privacy and related laws, though there may be situations where reasonable customer expectations exceed such standards and this should also be considered. Depending on the situation, this may require a Member to contemplate a wide variety of aspects such as data accuracy or completeness, transparency, the ability to edit or modify data, the intuitive relationship of data to the decision process itself, and considerations of ownership, power or intellectual property.

The system may also generate new data (for example a record of the decision taken), which should be stored and used in a responsible and appropriate manner.

C.5 Accountability, Contestability and Redress

The accountability for any issues arising from the operation of an automated decisioning system may require careful consideration, as the correct approach to accountability may not be immediately apparent. Good practice is to clearly identify a person who takes responsibility for the decisions made by the system; though a decision may be made automatically, this does not absolve human beings of the responsibility to ensure the decision is made appropriately.

Accountability considerations could also include consideration of avenues for explanation, complaint, contest or redress for any real or perceived issues caused by an automated decisioning system, as well as appropriate monitoring for potential issues in order that problems can be identified quickly and suitable adjustments made.

C.6 Professionalism

Any Member working in the area of automated decisioning systems is required to act in line with the Actuaries Code as with any other area of endeavour. Notably, in such a rapidly emerging field, a Member should be aware of contemporary technical, ethical, legislative and other developments in the area, take active steps to maintain their knowledge, and should seek expert advice where they do not have such knowledge.

D. Good Practices

Most proposals in this area stop at statements of principles. This can create confusion for practitioners, who need to apply high level ideals in a practical setting, with little guidance as to how to go about doing so.

Here, we identify a series of practical steps that Members can consider taking to apply the Principles in their work, using the Actuarial Control Cycle as a framework.

Like the Principles themselves, these steps are not an exhaustive list, nor will they all be relevant or appropriate in every situation. They should not be used as a checklist. Whilst the steps outlined below are generally described as "good practices", none of them should be considered as required for Members to consider. Instead, we encourage Members to take these as general suggestions to build upon in the context or situation in which they find themselves.

D.1 Defining the Problem

The first step of the Actuarial Control Cycle is to define the problem to be solved. Good practice at this stage involves having the problem definition documented and understood by all involved in the design, construction, operation and monitoring of the system. The principles above suggest several good practices to follow, during this step:

D.1.1 Clearly define and document the objective

Often the most challenging aspect of a project, not solving for the correct objective is a frequent cause of analytics project failure. It can also be the cause of poor outcomes for consumers affected by the project, who may receive a decision which is not aligned to a valid objective or goal. Once the objective is well-specified, the actuary can begin to assess it in light of the principles above. Specifying the objective precisely can, in some cases, elicit important and challenging questions to be considered by the project sponsors: is wellbeing actually being improved, or autonomy respected, or are there inherent challenges of fairness caused by the objective itself?

We suggest the objective is the first thing which is agreed between the person responsible for the outcomes of the automated decisioning system, and the team building and operating the system.

Often this step requires a great deal more thought than it might have been given in a traditional decisioning context. For example, an executive may declare an objective of "maximising sales". However, the real objective may be more subtle: "incrementally sell more business via a marketing intervention than we would have otherwise sold". A Member may spend considerable time determining the true objective, and agreeing with their client how it is expressed, in order to avoid issues down the track.

D.1.2 Elicit Constraints

In many cases the agreed objective may only represent the primary objective, with many unspecified constraints or conditions to be elicited. Traditionally, executives are not required to specify such constraints: humans will implicitly apply or understand the environment they operate within and apply "common sense". Automated decisioning systems will not do this.

For example, if a client asks for a system to "maximise sales", this is unlikely to literally mean "maximise sales at all costs". There will be competing objectives: budgets, profitability, resourcing, customer experience, etc. Without specifying such constraints, poor outcomes could occur (such as mis-selling) which may violate the principles articulated above. Members may need to be particularly careful in situations where primary objectives are focussed on traditional business objectives (e.g. sales, profitability), and seek to ensure that other ideals such-as wellbeing and fairness are also considered. This could form part of the overall goals of a system, or exist as constraints.

D.1.3 Ensure the domain is well specified

In defining the problem, it is equally important to specify when a solution will be used, and when not. There may be situations where an automated decisioning system is not considered appropriate, or areas of the population in which any system may be considered unreliable (for example if it is a segment of the population rarely encountered). In such situations, it could be appropriate to revert to a human decisioning process. Considerations here can form part of the problem definition.

D.2 Designing the Solution

Once the problem has been defined, the solution can then be designed, which may include contemplating principles such as those in Section C. Practical steps to take may include the following.

D.2.1 Ensure the problem is accurately translated

Good practice in problem translation includes being able to clearly articulate how the problem and business objectives have been translated into the analytical setting, and why that translation is appropriate. Good practice also involves identifying any assumptions, simplifications or risks of mistranslation and confirming their reasonableness in light of the nature of the decision being made and the expectations of those who might be affected by such decisions. Such errors of translation carry the risk of undermining any determinations made when the problem was specified, hence why they are particularly worthy of attention.

For example, consider a simplistic predictive policing algorithm. The goal is to predict where crime is occurring, and weight police activity towards areas where more crime is predicted. If historic arrest data is used to predict where crime is more likely, without adjustment, this mistranslates the true goal, since this historic data does not include those historic crimes which went undetected. This could lead to underprediction of crime in areas which were historically sparsely policed. If this translation issue is not recognised, it may lead to policing being targeted in a way which merely reinforces historic practices and fails to solve the actual problem at hand. This could create a self-reinforcing cycle, with potential social costs that are challenging under several of the principles articulated in Section C above.

Recognition of translation errors at the design stage may allow any risks and issues caused by mistranslation to be managed as the models and decisioning systems are designed and built.

D.2.2 Collect and use data appropriately

Automated decisioning systems require data. There are typically two categories of data to be considered:

- Training data: Historic data that is used to train/calibrate the decision-making system, and
- Scoring data: Data that is input into the system for decision making, at the time of decision. This could be in the form of inputs provided directly by the customer, or data collected via some other means.

Both forms of data may be analysed for appropriateness of use.

Good practice includes evaluation of whether data is being used in an appropriate manner. Considerations may include:

a. How does the data relate to the defined objective?

b. How was the data collected?

c. Is the form of data used to train the automated decisioning system the same as the data which will be used for scoring at the time of decision?

d. Would customers expect, and are they aware of, the data being used in the manner being proposed?

e. Is the amount of data being used in excess of that reasonably required for the system to operate to a suitable standard?

f. Has privacy been adequately considered? Has personal or sensitive data been adequately secured or deidentified? Are there any residual risks of re-identification?

g. How was the data processed and manipulated and is this appropriate?

h. If the data processing involves combining several distinct datasets, would reasonable customers expect this to occur, and does the combined dataset result in unreasonable knowledge or power over individuals?

i. What inherent biases are present in the dataset, and how do they impact the outcomes?

j. Is there potential for the resultant model to unfairly discriminate, whether directly or indirectly, against a group or individuals?

k. Is the data of suitable quality and reliability?

Analysis of the data using items such as those listed above is good practice to adopt in construction of a system in order to help avoid poor outcomes.

D.2.3 Design, Modelling and Constraints

In designing an automated system, good practice is to explicitly consider aspects of judgement in the modelling process (including traditional statistical modelling, machine learning, and related concepts), and more general design of the algorithm itself.

Being able to justify the modelling approach taken, including selection of the model and any validation tests that were performed, is an important consideration, as is clearly documenting any trade-offs made in the model design and assumption selections.

An assessment of fairness could be conducted during this phase, having regard to any definition(s) of fairness or constraints determined during the problem specification stage, and any trade-offs required. Constraints or other adjustments to data, models or outcomes could be utilised if required in order to promote desired outcomes.

In considering fairness, it is common to consider:

a. What expectations may individuals reasonably hold about the decision-making process and how they will be treated? This incorporates themes of transparency, autonomy, communication and procedural fairness, as well as discrimination.

b. Who may be harmed through the automated decision-making system? This commonly involves an analysis of harms caused by decision errors, but need

not be limited to this. Harms may also be relative, not absolute – for example the giving of cash to one person could be considered a relative harm to another person who does not receive the cash, though it has not harmed them in absolute terms.

c. How, or to what degree, will they be harmed?

d. Are harmful outcomes spread evenly across the population, or are they biased against particular subsets of the population?

e. What redress is available to those harmed?

f. What trade-offs have been made in the design of the algorithm, to account for issues arising from the considerations above?

In some cases, the modelling process may involve multiple independent parties, whose separate contributions are combined into a complete model or algorithm. Good practice involves taking steps to ensure that any analysis of fairness considers the whole process, not just each part in isolation.

D.2.4 Transparency

Transparency may be considered internally to an organisation, and externally.

Within an organisation, accountable persons may require some control and visibility of the decisions made by an automated system to assure themselves of the reasonableness of those decisions. Members should ensure communication with such algorithm owners is designed to promote such understanding.

Members may need to consider audit requirements for algorithms, both from within and external to an organisation, which may require upfront consideration of transparency.

Externally, good practice involves considering the desirability of transparency and explicability as strategies to promote both the autonomy of individuals and general trust in the system, and promote such strategies where relevant. This may involve notifying customers that an automated decision has been used (particularly if it has a material impact on them), giving an option to elect for a human decision where this is appropriate, or providing a suitably understandable explanation of how a decision was arrived at.

D.3 Monitoring the Solution

The final stage of the control cycle is monitoring. In this section we also incorporate considerations of deployment and ongoing model refinement.

D.3.1 Deployment and Accountability

It is good practice that automated decisioning systems are deployed and maintained using a clearly defined control cycle with appropriate governance procedures around design, development, sign-off, deployment, monitoring and updates.

It would, therefore, usually be appropriate that final sign off for model deployment is given by the individual accountable for the outcomes of the system. Additional sign off processes may also be deemed appropriate.

Prior to deployment in a production environment, it is good practice to ensure that the system is performing as expected from the development stage.

D.3.2 Performance triggers for manual recalibration

Good practice includes careful consideration as to how the performance of the automated decision-making system will be monitored. This may include references to the initial objective specified, any constraints, and any defined metrics relating to considerations such as fairness. Good practice also involves specifying in advance the frequency and form of monitoring and any performance triggers or thresholds indicating that the model needs re-calibration or refinement, and accountable individuals to oversee this task. This could include an assessment of the potential for harms caused by model drift and/or failure when making these decisions.

If performance triggers or thresholds are breached, good practice is to trigger a decision as to whether the system requires changes to monitoring, a simple re-calibration, or a more thorough review.

If model performance deteriorates below critical thresholds it may be necessary to turn-off the decision-making system until the issue can be rectified. Considerations of fairness and potential harms could influence this decision, as might any considerations of natural volatility or variation in the system's performance.

D.3.3 Monitoring for systems which autonomously recalibrate

Systems which autonomously recalibrate in light of emerging experience may require additional scrutiny over and above the performance monitoring identified above.

If the system can autonomously adapt, good practice is to specify the boundaries of allowable adaptation in advance, having regard to the potential for a system to evolve into something which could give poor outcomes.

If the system evolves in a manner which aims to move past this prespecified threshold, good practice is to trigger a suitable review of the system.

Notwithstanding the additional triggers above, good practice involves scheduling a manual review periodically, with the general goal of testing that the system is still operating in accordance with the overall intent.

D.3.4 Record keeping

During deployment, good practice involves ensuring a system generates records of material aspects of its operation – for example decisions taken and individuals impacted.

Good record keeping practice involves contemplating any downstream requirements for explanation, challenge or audit of the decisions, either as individual decisions or in aggregate, and involves ensuring the records created are sufficient to allow such inspections to occur.

Records should be stored and used in a responsible and appropriate manner. If they represent personal and/or sensitive data, this needs to be compliant with relevant legislation regarding these forms of data.

D.4 Considerations of Professionalism

In this section, we consider concepts under the Actuaries Code and how they might apply to this setting.

D.4.1 Integrity

The Code requires Members to be respectful and truthful in delivering their services. A high level of transparency may assist a Member to demonstrate being truthful and to identify security, privacy and ethical issues before they cause harm. Many of the concepts of fairness and equity discussed above relate to the Code's requirement that a Member act with integrity and show respect for others. The Code's requirement to respect confidentiality also aligns with the need to protect personal or sensitive information by keeping it secure.

D.4.2 Compliance and Speaking Up

The Code requires Members to comply with all relevant laws, regulations and Professional Standards. Notably, there are many security and privacy laws that a Member must be aware of, particularly when working with personal and/or sensitive information. Also of note is anti-discrimination legislation, which the design and operation of an automated decisioning systems should contemplate, particularly if the decision could harm an individual, either in absolute or relative terms.

A Member has a responsibility to respond appropriately to non-compliance by others. This includes raising issues and concerns about any aspects of a decisioning system as appropriate within their organisation. If concerns are not addressed within the organisation, a Member needs to consider what further escalation of the issue may need to occur.

D.4.3 Competence and care

Under the Code's competence and care requirements, a Member must have due regard to Professional Governance Material and Regulatory Guidance. While many of the international ethical guidelines in this area (and, similarly, this Technical Paper) are not compulsory, a Member could take these into account in endeavouring to demonstrate acting with integrity and meeting the competence and care obligations of the Code.

In addition, the competence and care obligations require a Member to have due regard to others whose interests may be affected by the services provided. This aligns closely with the ethical principles of improving wellbeing, respecting autonomy, and considering fairness, outlined in Section C.

To act with care, a Member should take personal accountability for the elements of a decision-making process in which they have been involved. A Member could encourage the organisation to establish a suitably defined, accountable and aware owner of the entire decision-making system, as well as appropriate systems of governance and accountability.

D.4.4 Objectivity

An important requirement of the objectivity principle of the Code is that a Member provides objective advice that is free from bias. This requires a Member to consider any bias that might be contained in their work, including inherent biases that might exist in the historical data used to train a model.

D.4.5 Communication and Documentation

Relevant considerations including those discussed above should be communicated and documented in line with a Member's responsibilities under the Actuaries' Code.

Notably, many ethical considerations such-as those outlined here require careful judgement and trade-offs between different idealised outcomes. Consideration should be given to how such judgements and trade-offs should be clearly documented and communicated to appropriate individuals, particularly those in decision making or responsible roles.

A Member might consider whether written documentation for the system as a whole is sufficiently comprehensive that a suitably qualified individual could understand the design, construction and operation of the system and offer comment or critique on its appropriateness.

Automated decisioning systems can be complex and technical in nature, which can be particularly difficult for non-experts to understand. A Member needs to take care to ensure that communication is tailored to the audience, and that any technical detail does not confuse the audience or obfuscate salient facts from being understood.