NIASRA

NATIONAL INSTITUTE FOR APPLIED STATISTICS RESEARCH AUSTRALIA

National Institute for Applied Statistics Research Australia

University of Wollongong, Australia

Working Paper

16-23

The Robodebt Tragedy

Dennis Trewin, Nicholas Fisher, and Noel Creesie

National Institute for Applied Statistics Research Australia, University of Wollongong, Wollongong NSW 2522, Australia T: +61 2 42215076. E: karink@uow.edu.au

The Robodebt tragedy

Dennis Trewin, **Nicholas Fisher,** and **Noel Cressie** describe how a lack of statistical thinking led to mind-less government technology and unimaginable stress for half a million Australians – some of whom took their own lives.

A government program, designed to save billions of dollars, instead costs billions. It generated debt notices without human oversight resulting in untold misery for hundreds of thousands of Australian citizens. Cabinet Ministers and heads of government departments are facing possible prosecution for what is now dubbed the "Robodebt" scheme. The July 2023 Report of the Royal Commission into the Robodebt Scheme (1) called it "amateurish, rushed and disastrous."

In 2016, the Australian Government introduced an automated system that was supposed to improve the integrity of the welfare-benefits scheme by identifying incorrect or fraudulent claims. For over three years, an algorithm calculated debts owed to the government and automatically sent debtors letters with threats to pay up, or else. In fact, most of the debts were false.

But all this was entirely preventable had Human Intelligence, especially in the form of statistical thinking, been present in Robodebt's development phase. This debacle serves as a timely reminder of the hazards associated with automated decision-making, particularly applications involving Artificial Intelligence (AI).

How did this come about?

The Australian social services agency, Centrelink, stores data on the earnings of Australians receiving various types of welfare payments. Its records are not perfect, and it looks for data from other sources, such as the Australian Tax Office (ATO), to provide a cross-check. Until the first half of 2016, if there was a discrepancy, after reviewing a please-explain notice was issued by a Centrelink staff member. In the middle of that year, a new scheme called "Centrelink Online Compliance Intervention" and later dubbed "Robodebt" by the media was introduced to massively reduce the overpayment of welfare and cut administrative costs. It did this by comparing welfare recipients' Centrelink-reported **fortnightly income** with their ATO-reported

yearly income. The yearly figure was used by an income-matching algorithm to **estimate** fortnightly figures from ATO data (the "gold standard") that could then be matched with Centrelink data used for calculating welfare payments. If the difference showed an overpayment by Centrelink, a red flag was raised. In contrast to Centrelink's old compliance scheme, the new AI-based scheme then automatically generated debt notices and placed the onus on the recipients to prove they were not welfare cheats.

In the mindset of the government of the day, AI and "big data" would save vast amounts of human toil and, of course, produce billions of dollars of savings now and into the future. Computers don't make mistakes!

During the few years of its operation, some 700,000 debts were raised. Thus, for some 700,000 people, the presumption of innocence was suspended, using a computer algorithm that turned out to be deeply flawed and a scheme that had a presumption of guilt at its core. The consequences were disastrous, leading to a vast amount of misery for hundreds of thousands and, according to the Robodebt Royal Commission, suicide for some.

By way of comparison, the previous compliance scheme had relied on **Human**, not **Artificial**, Intelligence and resulted in about 20,000 cases each year. The dramatic increase in cases detected by AI seemed to indicate to the scheme's proponents that welfare fraud was far more prevalent than thought previously. There was no official suggestion that the increase might have been caused by a flawed algorithm that produced a high proportion of false positives. Indeed, the apparent success of the scheme led the government to consider extending the approach to other areas.

However, a deluge of complaints ensued, from recipients of the notices, welfare groups, political parties, not to mention fierce media attention. This led to a Senate enquiry, legal challenges, and the scheme being declared illegal by the Australian Federal Court in late 2019. Ultimately, the scheme was scrapped in 2020, and well over a billion Australian dollars repaid for 470,000 incorrectly issued notices. In 2022, the government was voted out of office, and the newly elected government established a Royal Commission to investigate the scheme.

Where did things go wrong?

As statisticians, we made a technical submission[1] to the enquiry, focusing on what we saw as incredibly poor statistical practice and what good practice might look like in similar situations.

One core assumption was that income was earned evenly throughout the year, a flawed assumption. Many in the gig economy had an uneven-wages profile, which prejudices casual and under-employed workers. The government minister responsible quoted an error rate of approximately 1% without clarifying exactly what this error rate applied to. Subsequently, it was revealed that these "errors" were clerical data-entry errors, and the false-positive rate of the scheme was much, much higher (see below). In fact, Robodebt failed because its underlying algorithm was flawed, and its error metric was ill-defined.

The Royal Commission reported[2] on 7 July 2023 and made a number of findings, especially on the illegality of the scheme. Of particular interest to us was the recommendation that

> "*The evidence before the Commission ….. indicates the need for an Office with a broad remit to improve the use of automation and AI in public administration".*

Our submission was referenced in conjunction with this recommendation. There was also a recommendation that spoke to another part of our submission:

> *"…business rules and algorithms should be made available to enable independent expert scrutiny".*

What were the main statistical flaws?

In our Royal Commission submission, we distinguished two classes of algorithms for tackling large, complex data sets such as the linked data used to support the Robodebt scheme.

The first group of algorithms have been developed somewhat independently in the Statistics and Computer Science communities over the last sixty or so years[3]. They are fundamentally statistical in nature and include so-called machine learning algorithms, decision trees, and so forth. To varying extents, it is possible to form an understanding of which factors are the most influential in the predictions and decisions made by the algorithms. For convenience, we refer to them collectively as *machine learning (ML)* algorithms. They are often included in the AI family of algorithms, and we have done so in this article.

The second group of algorithms derive from the application of so-called Artificial Neural Nets (ANNs), which are an important tool for the rapidly expanding area of AI. These can be characterised colloquially as “black boxes”: interpreting which factors are playing the principal roles in the outputs is challenging. Although ANNs did not play a part in the Robodebt Scheme, we mentioned it in our submission in anticipation of its appearance elsewhere in future government activities.

We then identified the three basic phases that the Robodebt scheme should have used in its development:

- **Design:** Developing a clear description of the task, including required outcomes, key quality requirements, and a description of how risks were to be identified and managed.
- **Data acquisition and pilot experimentation:** Assembling from diverse sources the data required for analysis. Quantifying the risks where possible.
- **Implementation:** Selecting and applying algorithms to identify individuals who may have been overpaid (or underpaid) by Centrelink, validating the results, and quantifying associated uncertainties.

For each of these phases, we identified what we viewed as good statistical practice. Because of space limitations, we only offer a few examples:

Good practice versus Bad practice

Design Phase example

Best practice: After identifying the data sources (ATO annual incomes, Centrelink fortnightly incomes), document their limitations, their potential biases, their impact on the algorithms, the accuracy of any data linkage, and so forth.

What actually happened: If "best practice" was attempted, it was not done competently. Often, annual income does not flow evenly throughout the year. An ATO annual income averaged into Centrelink fortnights could not be compared accurately with fortnightly Centrelink income declarations[4].

Data Acquisition and Pilot Experimentation example

Best practice: Assemble a pilot data set and check on data-quality issues, especially those resulting if an initial data transformation is required.

What actually happened: A 2015 pilot study of 2,600 cases was conducted, but its design was clearly unsound as it did not detect obvious flaws.

Implementation example

Best practice: Monitor the outputs, especially the misclassification error rates[3].

What actually happened: This was not done until after the scheme was implemented and not with great accuracy. The very high proportion of persons for whom the algorithm failed (false-positive rate) must have been a strong signal that something was wrong.

For such an important project as the Robodebt scheme, there should have been independent peer review of each of the three phases, with at least one accredited statistician involved. In fact, professional statisticians do not appear to have been consulted at any stage, and the consequences were disastrous.

The people responsible for the Robodebt scheme should have had a strong interest in keeping low error rates – both false positives and false negatives – front and centre of their design work. It's straightforward to estimate error rates for an AI scheme. Experts can do this by running simulations inside a virtual model called a "digital twin". These can be used to carry out statistical evaluations and expose conscious and unconscious biases in bad algorithms[4].

Sensitivity analysis could also be undertaken to understand the robustness of assumptions such as the income-averaging assumption. This does not appear to have been done despite multiple warnings from the Australian Taxation Office that the assumption was flawed.

In summary, the main statistical flaws were:

- using a variable (annual tax income) that was inappropriate for the intended purpose, despite warnings that this was the case;
- no documented understanding of other error sources associated with the variable or the linking process;
- no sensitivity analysis around the use of this variable or around other assumptions;
- inadequate testing of the algorithms prior to use;
- no understanding of the error rates;

and crucially,

- no involvement of professional statisticians to guard against these issues.

So, what was the error rate?

The Minister responsible for Centrelink in the new government has stated that the false-positive rate was at least 27%. We believe it was much higher.

During the scheme, one million reviews were performed, of which **81%** led to a debt being raised. Of these, about **70%** (567,000 debts) were raised through the use of income averaging in the Robodebt algorithm. In 2020, the government conceded that of these 567,000 debts, about 470,000 of them, or around **80%**, had been falsely raised. Compared to the usual target of a few percent, this is an eye-wateringly large error rate[4].

As a footnote, Robodebt also broke a law of mathematics. There's also an obvious question hanging over the whole system: why did it only seem to generate demands for payments, and never credits? Some might suspect the algorithm was deliberately biased that way. In fact, Noel Cressie [an author of this paper] has shown that the bias is a predictable consequence of something called Jensen's Inequality - which, of course, is well-known to statisticians[4].

Conclusions

In our Royal Commission submission, we noted the following observation by Prof Terry Carney[5], a long-term member of the Australian Administrative Review Council who presided over many of the early judgments on Robodebt:

> "Machine learning decision making systems are surely the way of the future. *Properly designed and monitored*, they offer a trifecta of greater accountability, greater accuracy and responsiveness, and greater efficiency of administration." [Emphasis added.]

There is a major opportunity to capture the generic learnings from the Robodebt experience in processes that are readily promulgated across different jurisdictions of government. Additionally, there is an opportunity to spread awareness of the limitations and risks associated with automating the *Big Data* $\rightarrow$ *Information* $\rightarrow$ *Business-decision* pipeline, especially if AI is being deployed. This led to our recommendations to the Royal Commission:

1. We recommend a Manual of Good Practice, in establishing a trusted pipeline, be developed and promulgated throughout government and the public service.
2. Because of the ubiquitous need throughout government for high-level data-scientific oversight of actual or potential decision-making based on complex data, and the need for an independent source of advice, we recommend the establishment of the position of Chief Data Scientist with strong parallels to that of Australia's Chief Scientist.

At the time of writing, the recommendation to establish this position is being considered by the Minister of Finance.

We hope that the events and outcomes surrounding Robodebt will lead to the incorporation by governments of statistical thinking and the much better application of advanced statistical methods in the future. As this case demonstrates, it may well save lives and, incidentally, the careers of Ministers and public servants.

(1) In Australia, Royal Commissions are the **highest form of inquiry on matters of public importance**. A Royal Commission has broad powers to gather information to assist with its inquiry, including the power to summons witnesses to appear before it and the power to request individuals or organisations to produce documents as evidence.

References

1. Nicholas Fisher, Dennis Trewin & Noel Cressie (2023), *Submission to the Royal Commission Enquiry into the Robodebt Scheme*.
2. Royal Commission into the Robodebt Scheme
3. David Donoho (2017), "50 Years of Data Science". *Journal of Computational and Graphical Statistics* 26(4), 745–766. https://doi.org/10.1080/10618600.2017.1384734
4. Noel Cressie (2023), "Robodebt not only broke the laws of the land – it also broke laws of mathematics. *The Conversation*, March 17, 2023. https://theconversation.com/robodebt-not-only-broke-the-laws-of-the-land-it-also-broke-laws-of-mathematics-201299
5. Terry Carney (2019) "Robo-debt illegality: the seven veils of failed guarantees of the rule of law. *Alternative Law Journal* **44**(1) 4-10. https://journals.sagepub.com/doi/abs/10.1177/1037969X18815913

Dennis Trewin AO, FASSA is the former Australian Statistician and former President of the International Statistical Institute and the International Association of Survey Statisticians.

Nicholas Fisher is Principal of ValueMetrics Australia and Visiting Professor of Statistics at the University of Sydney. He was formerly a Chief Research Scientist in CSIRO.

Noel Cressie is distinguished professor at the University of Wollongong, Australia and a centre director in its National Institute for Applied Statistics Research Australia (NIASRA); he is also adjunct professor at the University of Missouri, USA.