AI Ethics: Putting Principles Into Practice

Chris Dolman, FIAA

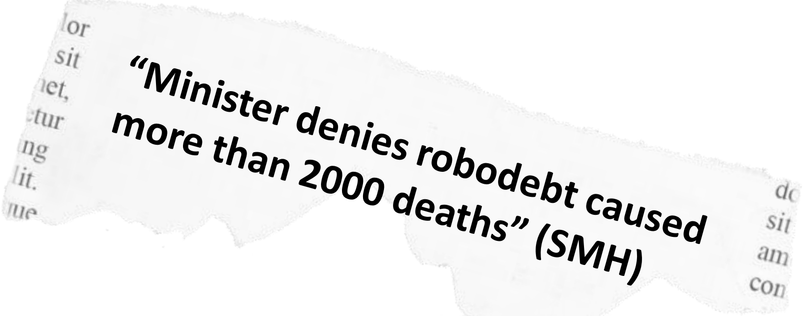
Agenda

- Motivations and Challenges for AI Ethics
- Some Ethical Failure Modes
- Guidance from Australian Institute
- Concluding Thoughts
- Q+A

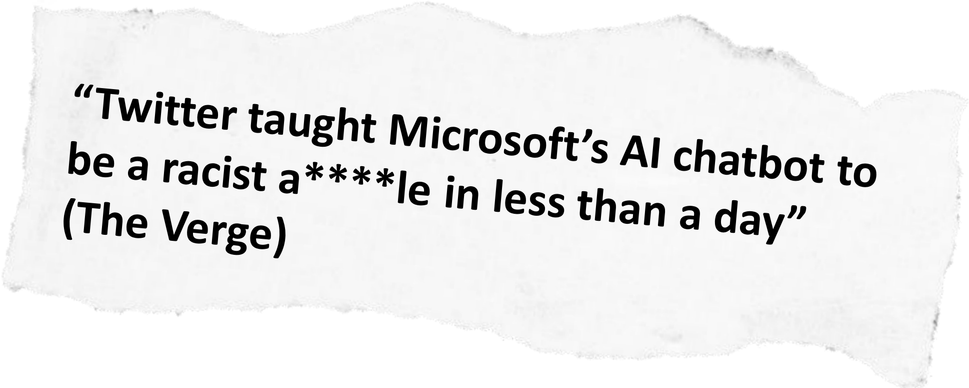
Motivations and Challenges

Headlines You Don’t Want To See

“Amazon used AI to promote diversity. Too bad it’s plagued with gender bias” (Mashable)



“Minister denies robodebt caused more than 2000 deaths” (SMH)



“Twitter taught Microsoft’s AI chatbot to be a racist a****le in less than a day” (The Verge)

“Privacy and profiling fears over secret ACC software” (NZ Herald)

‘AI Ethics’ Worldwide - Snapshot

Government and Public Sector:

- Several countries and regions have published “principles frameworks” via government or supranational bodies. A visual summary comparing some of these can be found here
- Locally in Australia, we have the Australian AI Ethics Framework published by dept of Industry et al
- In Singapore, MAS have published the FEAT principles, and the PDPC has issued the Model AI Governance Framework

Private Sector:

- Various companies have published “AI ethics frameworks” or similar, notably including Google and Microsoft
- Others have sponsored AI ethics research, including Facebook and (at smaller scale!) IAG

Academia and not for profit:

- Various “declarations” and similar for responsible/ethical AI, e.g. Montreal Declaration in 2017 as an early example
- Dedicated conferences for the topic, e.g. ACM FAccT, as well as dedicated sessions at broader events like NeurIPS

Professional Bodies:

- Some large professional bodies have published extensively in this area, notably IEEE
- Various actuarial associations have published in this area, including IFoA and SOA
- Australian Institute recently published an Information Note on the topic – more later in this talk

‘AI Ethics’ – Common Criticisms

Unclear in application

- High level philosophical principles are easy to agree with, but what exactly should you do?

Proliferation of Frameworks

- Which one should I use? Are some better than others?
- Do I have to pick the one my country/industry/profession has issued?

Tradeoffs between Principles

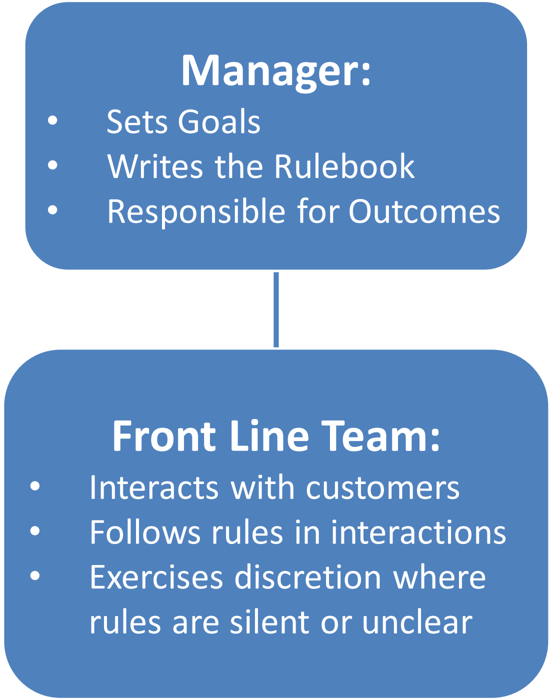
- What should I do if I can’t satisfy all the principles?
- Where should the balance be struck, where a tradeoff is inevitable?
- What if customers disagree about relative importance of principles?

Lack of clarity within Principles:

- E.g. the ‘fairness’ definition debate (as outlined in Dolman & Semenovich 2019), or long running debates over what terms like “explainability” or “interpretability” actually mean in practice

Some Ethical Failure Modes

Traditional Human Decisions at Scale



Manager:

- Sets Goals
- Writes the Rulebook
- Responsible for Outcomes

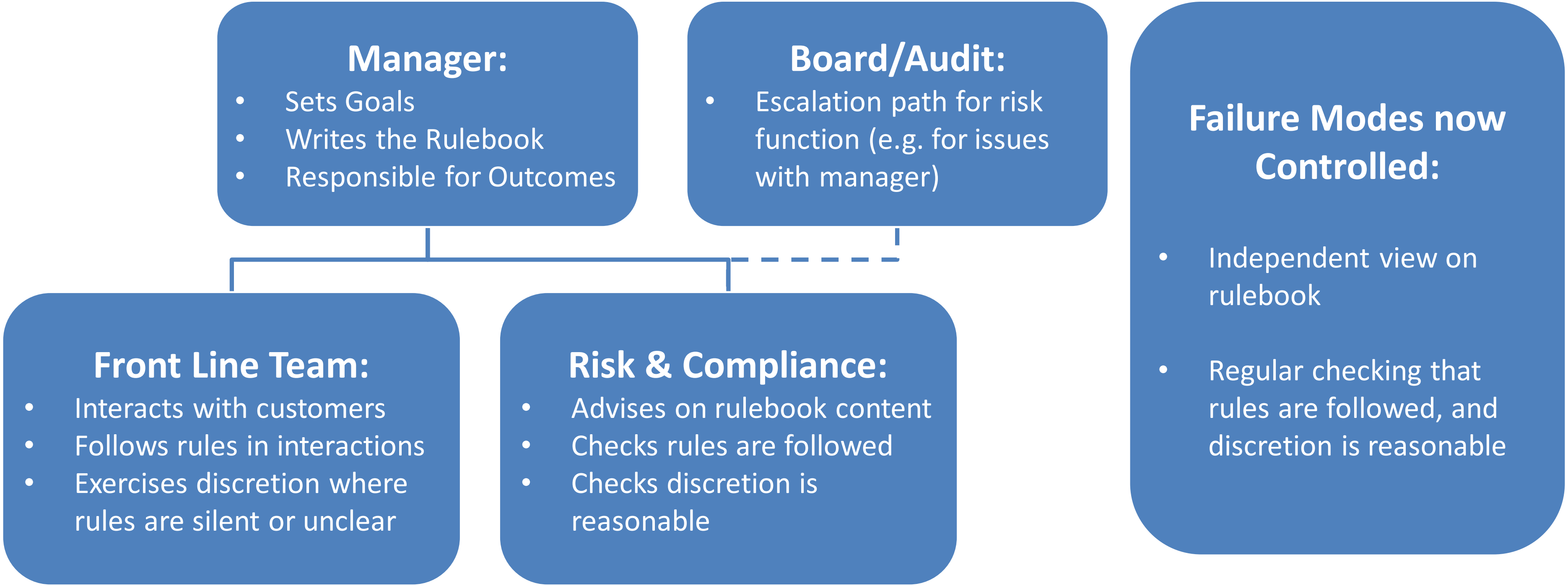
Front Line Team:

- Interacts with customers
- Follows rules in interactions
- Exercises discretion where rules are silent or unclear

Some Obvious Ethical Failure Modes:

- Rulebook is unethical
- Rules (if ethical) are not followed
- Discretion is unethical

Traditional Management of These Human Failure Modes



Manager:

- Sets Goals
- Writes the Rulebook
- Responsible for Outcomes

Board/Audit:

- Escalation path for risk function (e.g. for issues with manager)

Front Line Team:

- Interacts with customers
- Follows rules in interactions
- Exercises discretion where rules are silent or unclear

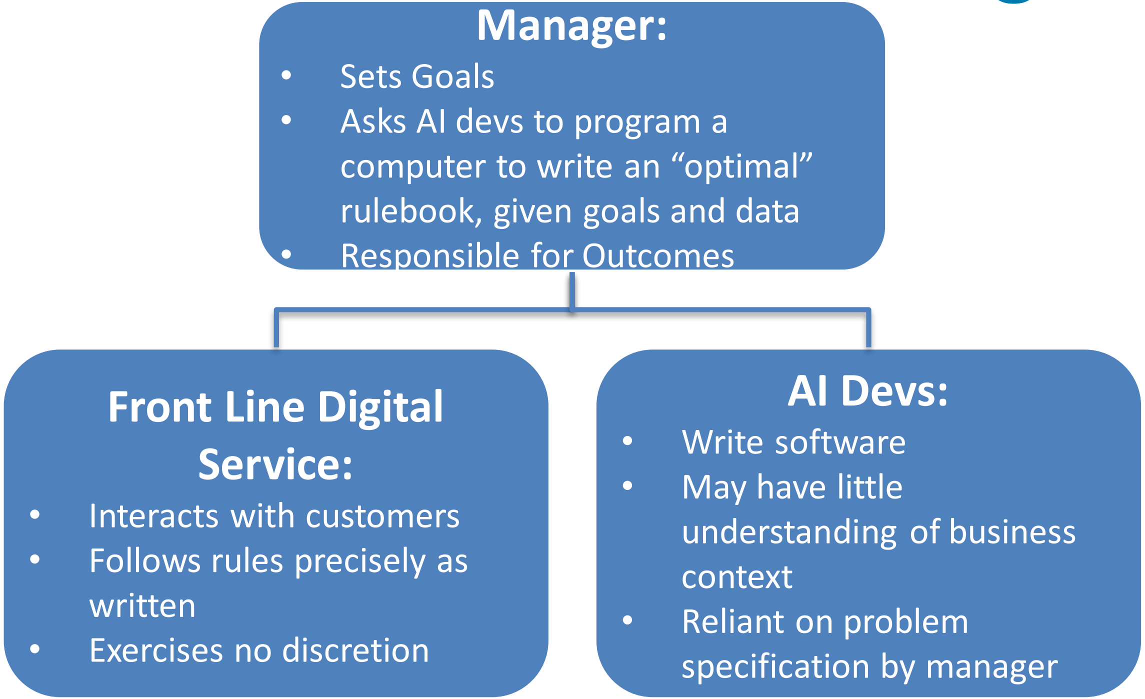
Risk & Compliance:

- Advises on rulebook content
- Checks rules are followed
- Checks discretion is reasonable

Failure Modes now Controlled:

- Independent view on rulebook
- Regular checking that rules are followed, and discretion is reasonable

Automated Decisions Break Traditional Risk Management



Manager:

- Sets Goals
- Asks AI devs to program a computer to write an "optimal" rulebook, given goals and data
- Responsible for Outcomes

Front Line Digital Service:

- Interacts with customers
- Follows rules precisely as written
- Exercises no discretion

AI Devs:

- Write software
- May have little understanding of business context
- Reliant on problem specification by manager

New Failure Modes:

- Goals / specs may be imprecise or unconstrained – "optimal" rulebook goes in "wrong direction" or "too far"
- Rulebook may be unreadable and may change automatically – hard for management to properly challenge
- Human discretion as a "smoothing agent" not present
- Scalability of failures

Traditional compliance problem now "solved"…
…but at what cost?

Australian Guidance for Actuaries

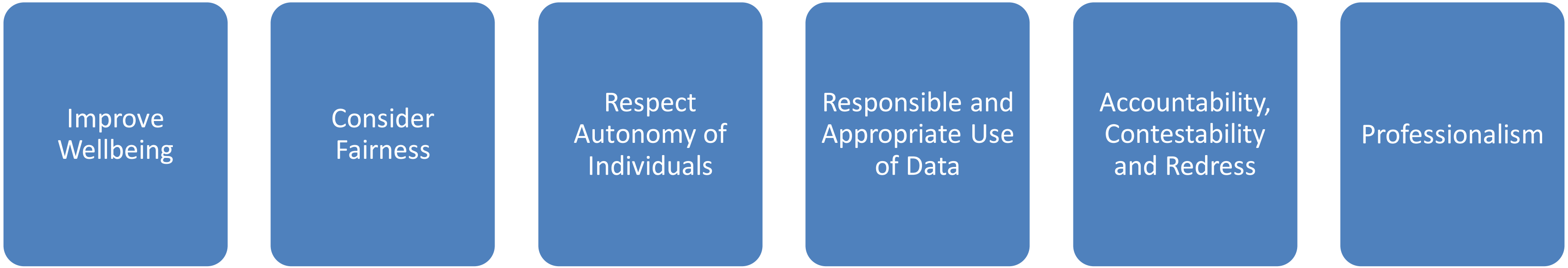
2 Areas of Guidance

- **“Principles” (Section C in IN)**
  - Similar to many other “AI ethics” frameworks published in the last few years
  - Our attempt at a synthesized list – saves people from a long reading list!
  - Likely to be fairly static and uncontroversial (aside from arguable omissions)

- **“Good Practices” (Section D in IN)**
  - Addresses primary flaw of many frameworks – no practical guidance on what to do
  - In our view, likely to be things many Members are doing anyway
  - Aim is to be a helpful list of suggestions to give people a solid starting point, <u>not an exhaustive list of things to do</u>
  - Likely this section will be more changeable, as accepted “good practice” emerges

- **Guidance can be found <u>here</u>**

Principles



Improve Wellbeing

Consider Fairness

Respect Autonomy of Individuals

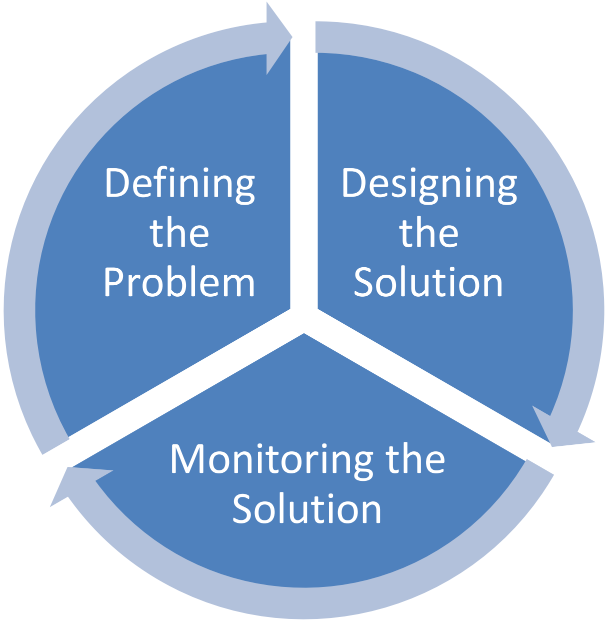
Responsible and Appropriate Use of Data

Accountability, Contestability and Redress

Professionalism

- If you’ve read any other AI ethics frameworks, you’ll recognise lots of these
- High level concepts which most people accept as reasonable without too much debate
- We make no claim that this is the “correct” list of principles, merely:
  - it’s hopefully a good place to start, saving you a big literature review, and
  - these are commonly included in some form in most of the frameworks we have seen

Practices – Why Not Use The Actuarial Control Cycle?

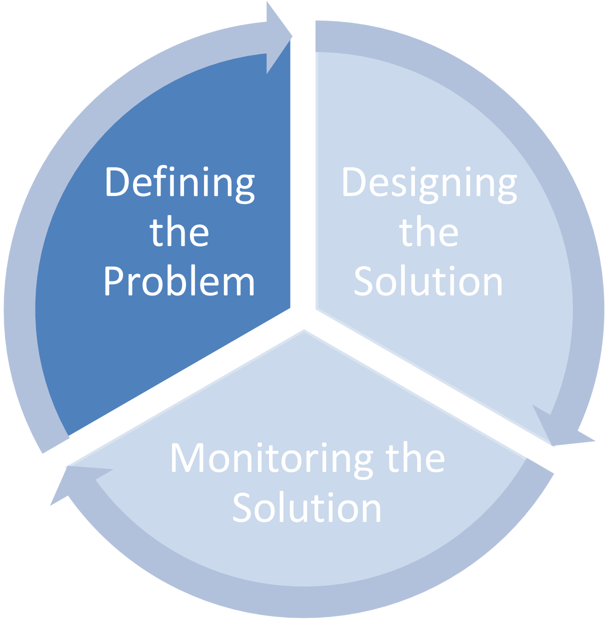


Defining the Problem

Designing the Solution

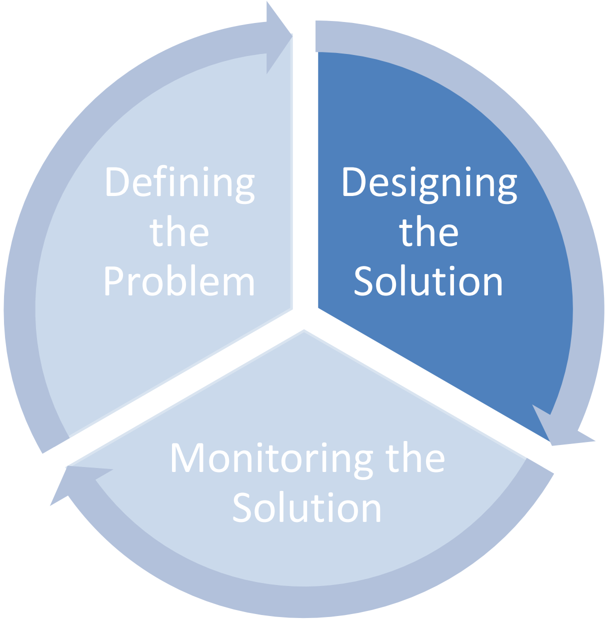
Monitoring the Solution

Defining the Problem



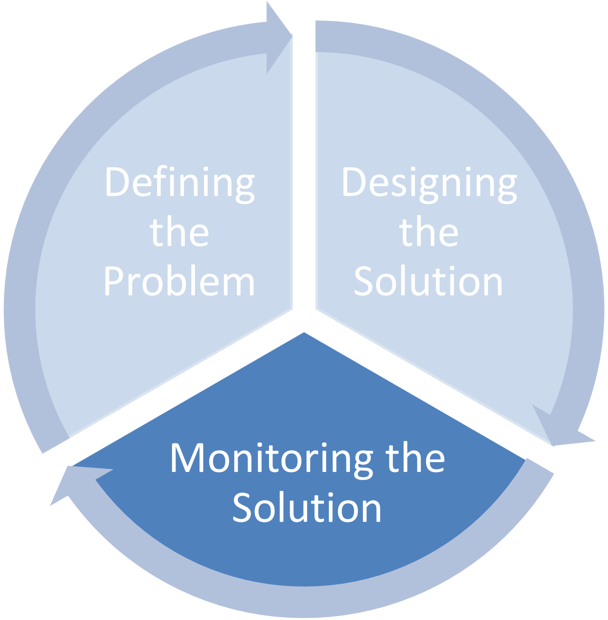
- **Clearly Define and Document the Objective**
  - Watch out for vague “business goals” – things are often more subtle
- **Elicit Constraints**
  - What boundaries should not be crossed?
- **Specify Domain**
  - Where / when should the system not be applied?

Designing the Solution



- **Ensure Accurate Translation**
  - Does your model accurately represent the "goal" as specified, or some proxy?
- **Collect & Use Data Appropriately**
  - Go beyond privacy and security law – what would people reasonably expect?
- **Design, Modelling And Constraints**
  - Document assumption, judgements and assessments of 'fairness'
- **Transparency**
  - Consider both for customers and internally

Monitoring the Solution



- **Deployment and Accountability**
  - Identify responsible person, make sure they properly agree to deployment
- **Performance Triggers for Manual Recalibration**
  - Decide before launch how you will monitor, and what would require a refresh/rethink
- **Monitoring Systems Which Autonomously Recalibrate**
  - Define boundaries of acceptable adaptation, monitor and act accordingly
- **Record Keeping**
  - Sufficient to allow downstream requirements of explanation or audit

Professionalism

Links to and embellishments on aspects of the (Australian) Actuaries Code:

- **Integrity**: Considerations of transparency, equity and fairness
- **Compliance and Speaking Up**: Be particularly aware of laws around data (e.g. privacy) and laws around discrimination. Respond appropriately to any concerns observed
- **Competence and Care**: Use guidance appropriately, be aware of how work affects others, create and encourage governance and accountability within your organisation
- **Objectivity**: Relates to considerations of bias
- **Communication and Documentation**: document and communicate tradeoffs and judgements appropriately; consider whether documentation is sufficient to allow review, audit or challenge; ensure explanations of technical concepts are suitably understood

Some Further Suggestions in a Business Context

Don’t treat AI ethics as an extension of traditional privacy / data risk. Biggest risks are conduct related

Ensure technical competence of 2nd line in this area

Leverage existing governance / decision forums wherever possible

Revisit legacy systems as well as considering new systems being built

Closing Remarks

- AI ethics is a big area of public discussion
- Current frameworks are (generally) very high level – hard to know practical steps to take
- Guidance from Australian Institute tries to help with this. Very much version 1 – we encourage feedback!
- Practical steps look a lot more like good governance and risk management, than sets of abstract philosophical principles
  - Good opportunity for actuaries to play a significant role?

Q & A

...Over to you!