THE UNIVERSITY OF MELBOURNE

Frameworks

Michael Wildenauer

School of Computing and Information Systems

The University of Melbourne

Learning outcomes

At the end of this topic, you should be able to:

- Compare ethical AI frameworks and the principles on which they may be based
- Understand the need for frameworks and guiding principles
- Apply the exemplar framework discussed in the topic to a real problem

Related reading

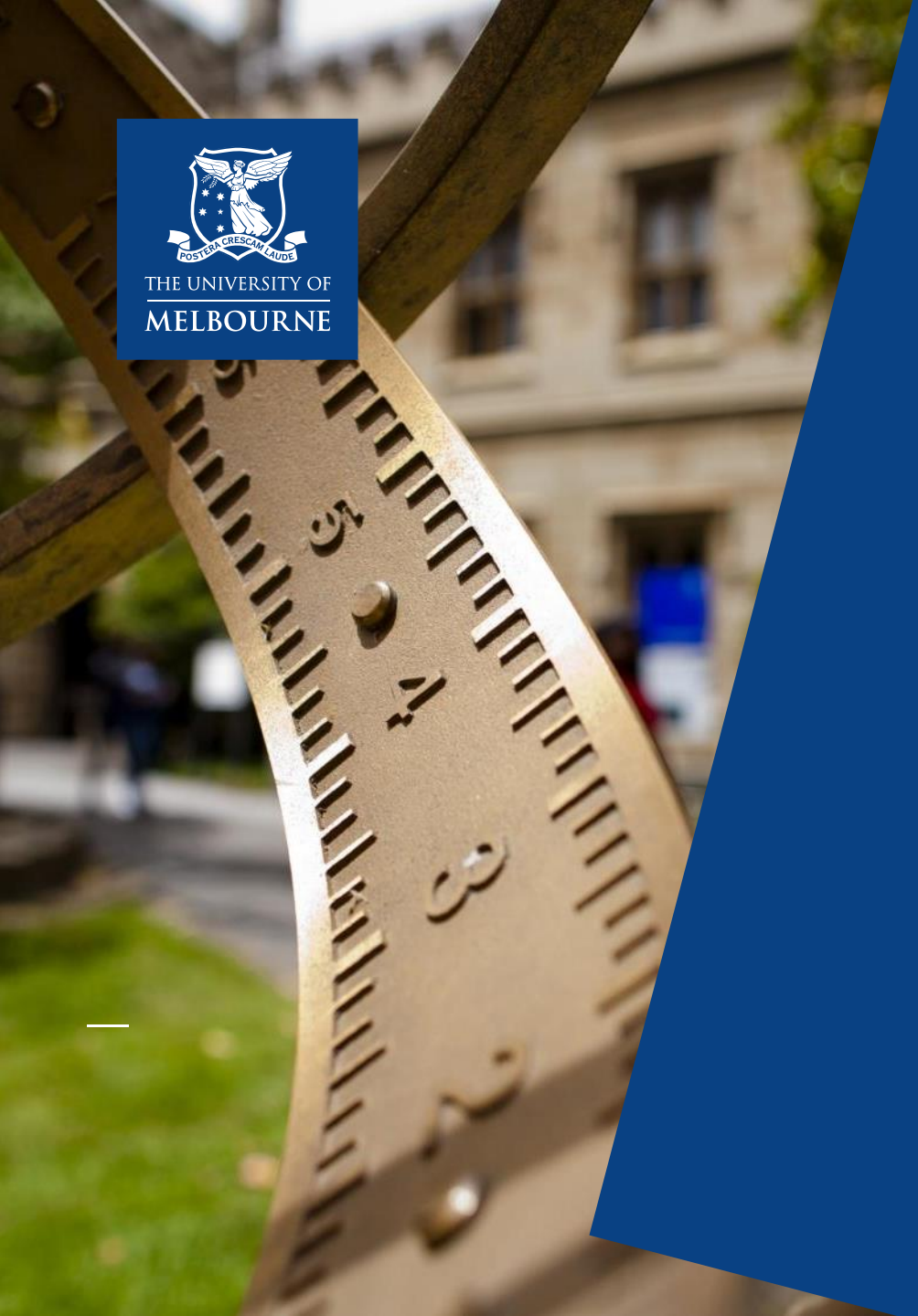
Reading

Beard, M and Longstaff, S. The Ethics Centre: Principles For Good Technology

https://ethics.org.au/ethical-by-design/#download-copy

Outline

1. Overview
2. AI Ethical Frameworks and their problems
3. Ethical by Design: Principles for Good Technology



THE UNIVERSITY OF MELBOURNE

Overview of AI Ethical Frameworks

Introduction

There are many different AI Ethics frameworks, of varying types and origins

- More than 150 at last count according to algorithmwatch.org
- Published by

Academia	Inter-governmental organisations
Civil society	International organisations
Government	Private sector
Industry associations	Professional associations

- You can explore the frameworks for yourself at https://inventory.algorithmwatch.org

These frameworks range from binding agreements to recommendations. Very few contain any enforcement mechanisms however.

Introduction (cont.)

The origin of these frameworks is, with the exception of those from China, Korea, Japan, and India, largely Western-centric. The US is by far the largest single contributor.

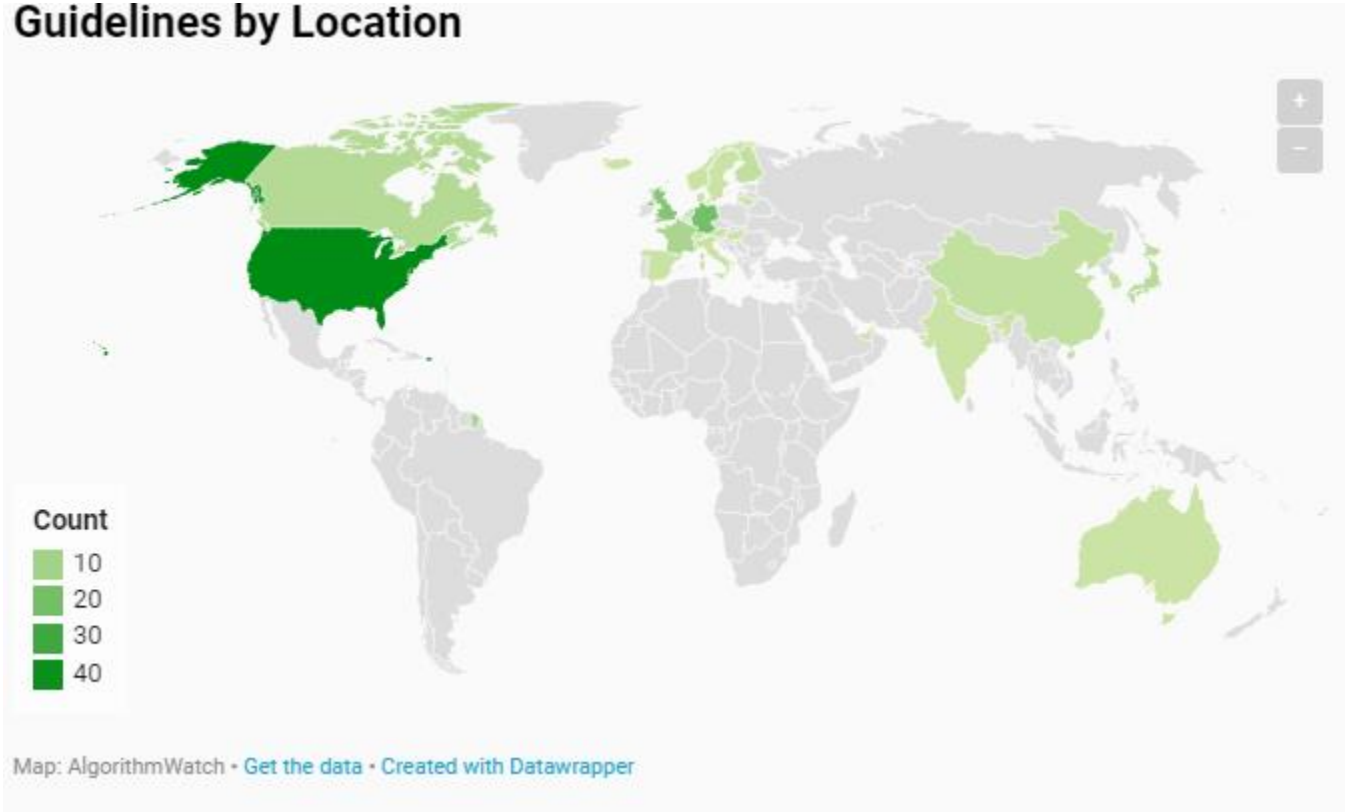


Exhibit A - IEEE

- IEEE framework (Ethically Aligned Design, 300pp)
- Largest engineers association in the world with over 400,000 members
- However companies like Google and Facebook don’t seem to be engaging with them despite having many IEEE members on staff
- OECD did base its principles on the IEEE publication, so maybe starting a socialisation spiral?

The Problem

'Environmental well-being. Human agency. Transparency. These are just a few of the ill-defined principles commonly listed in ethical frameworks for artificial intelligence (AI), hundreds of which have now been released by organizations ranging from Google to the government of Canada to BMW.'[1]

1. Burt, A 2020, 'Ethical Frameworks for AI Aren't Enough', Harvard Business Review Digital

The Problem (cont.)

'The problem? Many AI ethical frameworks cannot be clearly implemented in practice […]' [2]

- Insufficient specificity
- Hard to uphold abstract high-level guidance
- AI ethics frameworks are good marketing
- Fail to be effective

2. Ibid.

The Problem (cont.)

- Tech workers want ethics resources

'Without this more practical guidance, other risks such as "ethics bluewashing'" and "ethics shirking" remain'[3]

Ethics bluewashing – deliberately trying to look more ethical than is the case[4]

Ethics shirking – doing less to respect ethics where this appears not to have a high return[5]

3. Morley, J., Floridi, L., Kinsey, L. and Elhalal, A., 2020. From what to how: an initial review of publicly available AI ethics tools, methods and research to translate principles into practices. Science and engineering ethics, 26(4), p 2147.

4,5. Floridi, L. (2019). Translating principles into practices of digital ethics: Five risks of being unethical. *Philosophy & Technology*. https ://doi.org/10.1007/s1334 7-019-00354 -x.

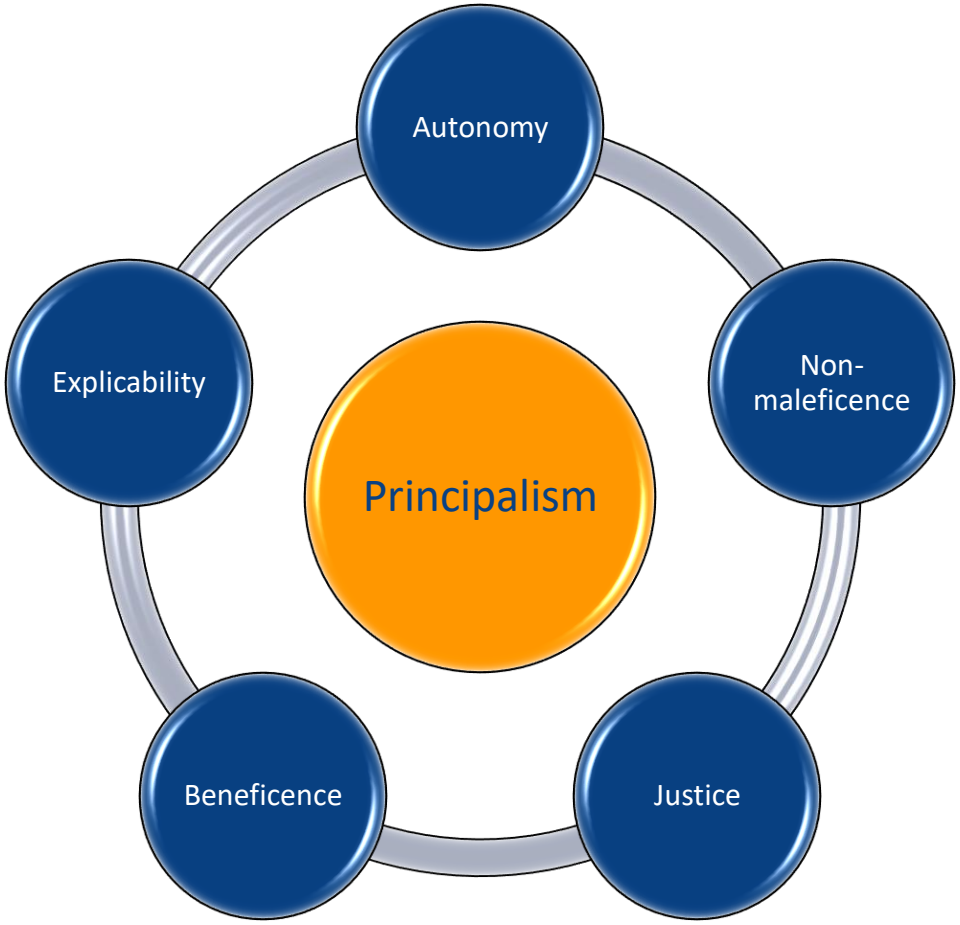
Framework Principles

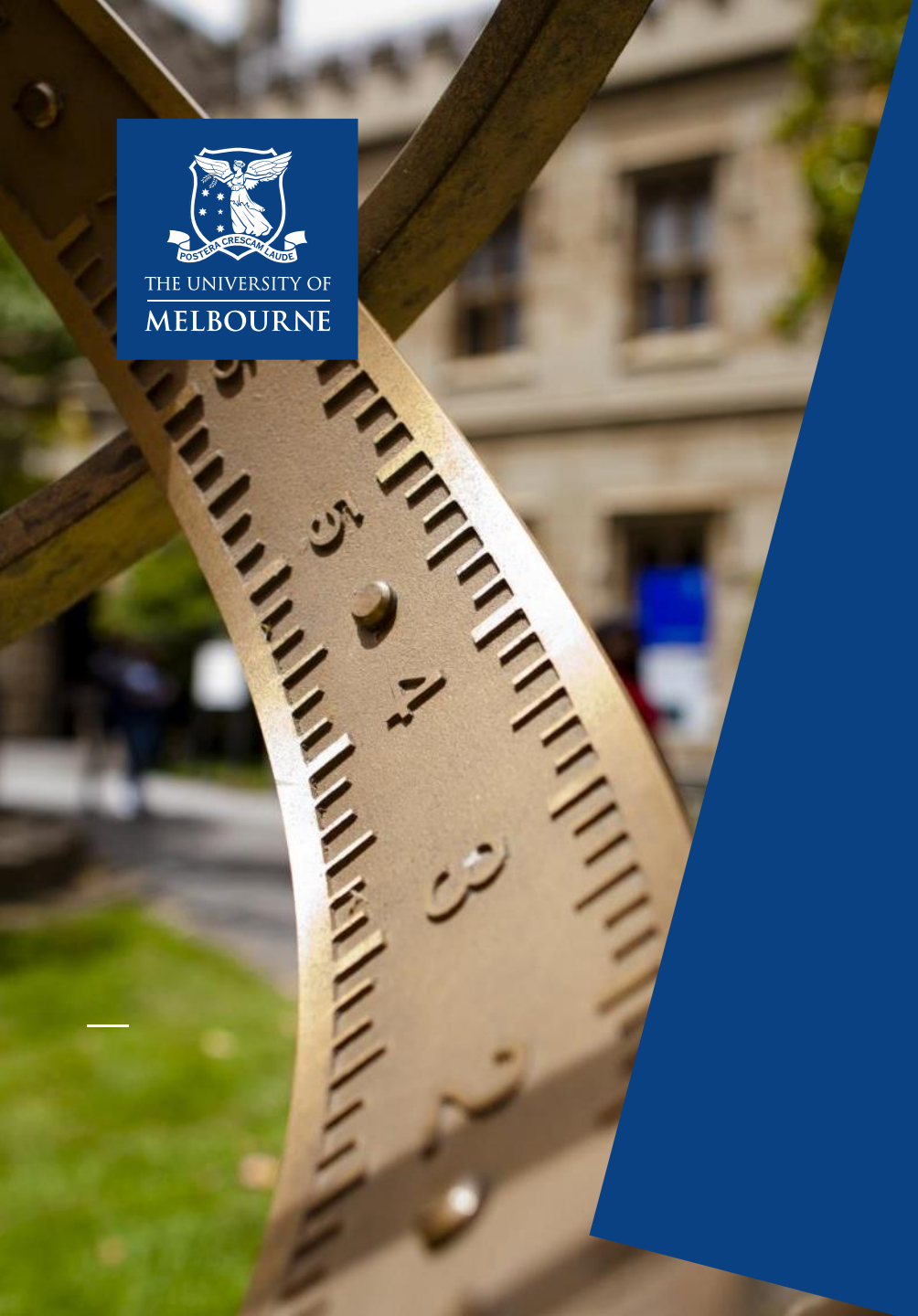
Ethical principles identified in existing AI guidelines[6]

Ethical principle	Number of documents	Included codes
Transparency	73/84	Transparency, explainability, explicability, understandability, interpretability, communication, disclosure, showing
Justice & fairness	68/84	Justice, fairness, consistency, inclusion, equality, equity, (non-)bias, (non-)discrimination, diversity, plurality, accessibility, reversibility, remedy, redress, challenge, access and distribution
Non-maleficence	60/84	Non-maleficence, security, safety, harm, protection, precaution, prevention, integrity (bodily or mental), non-subversion
Responsibility	60/84	Responsibility, accountability, liability, acting with integrity
Privacy	47/84	Privacy, personal or private information
Beneficence	41/84	Benefits, beneficence, well-being, peace, social good, common good
Freedom & autonomy	34/84	Freedom, autonomy, consent, choice, self-determination, liberty, empowerment
Trust	28/84	Trust
Sustainability	14/84	Sustainability, environment (nature), energy, resources (energy)
Dignity	13/84	Dignity
Solidarity	6/84	Solidarity, social security, cohesion

6. Jobin, A., Ienca, M. and Vayena, E., 2019. The global landscape of AI ethics guidelines. Nature Machine Intelligence, 1(9), pp.389-399, p 395.

The Principles



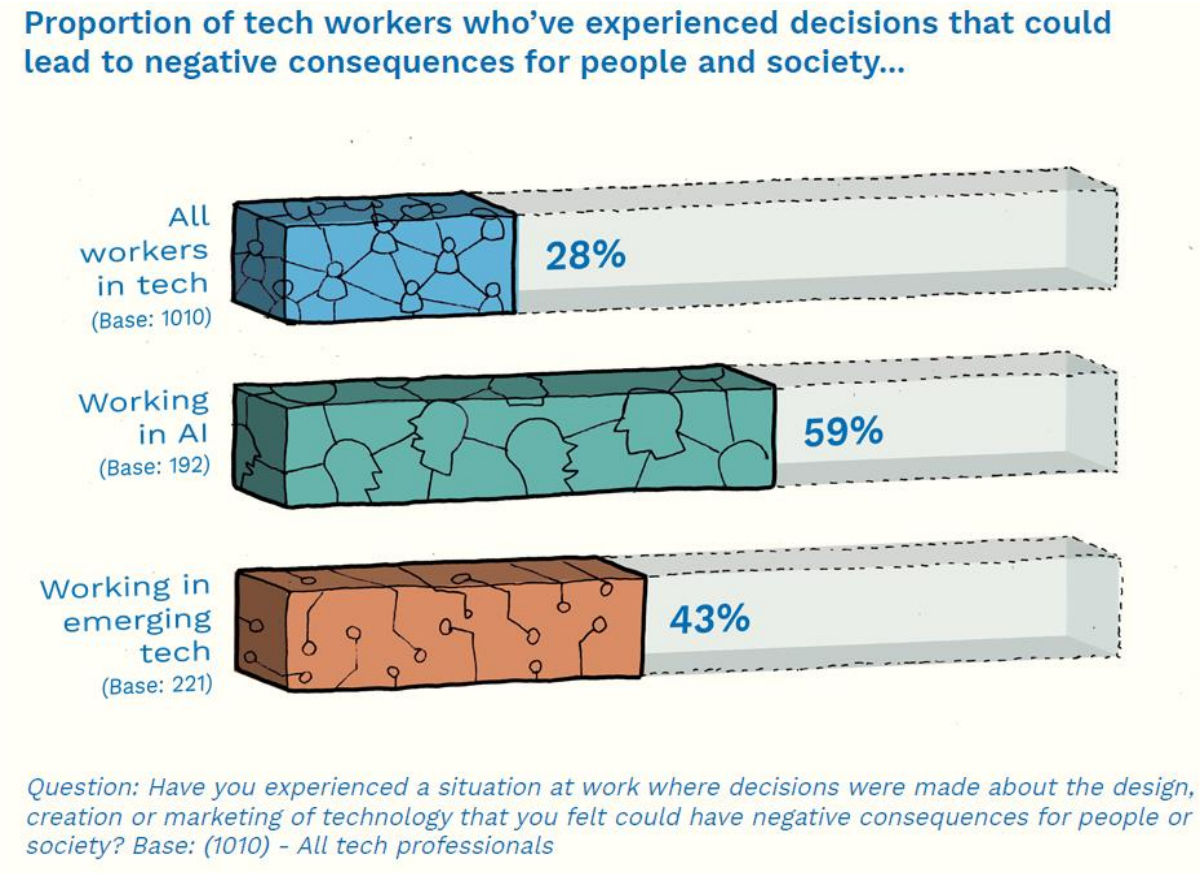


THE UNIVERSITY OF MELBOURNE

Is there a more practical framework?

Need

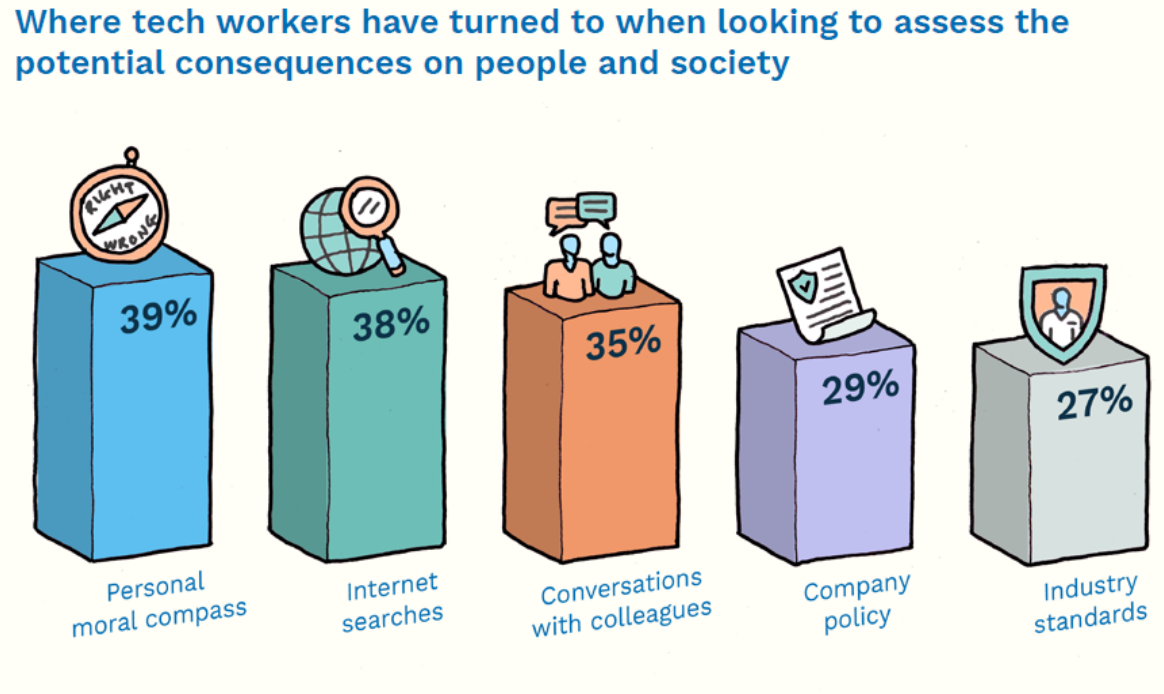
Why do we need ethical frameworks and guidelines in AI development?



Source: Miller C, Coldicutt R. (2019) People, Power and Technology: The Tech Workers' View, London: Doteveryone. https://doteveryone.org.uk/report/workersview

Where the rubber hits the road

How do developers actually go about ensuring that the AI they are developing is ethical?



Source: Miller C, Coldicutt R. (2019) People, Power and Technology: The Tech Workers' View, London: Doteveryone. https://doteveryone.org.uk/report/workersview

Reading - Principles for Good Technology

PR00.

OUGHT BEFORE CAN

The fact that we can do something does not mean that we should.

There are lots of possible worlds out there – lots of things that could be made or built. Ethical design is about ensuring what we build helps create the best possible world. Before we ask whether it's possible to build something, we need to ask why we would want to build it at all.

PR01.

NON-INSTRUMENTALISM

Never design technology in which people are merely a part of the machine.

Some things matter in ways that can't be measured or reduced to their utility value. People, ecosystems, some kinds of animal life and political communities shouldn't be used as tools that can be incorporated into design. They must be the beneficiaries of your design, not elements of a machine or design system.

PR02.

SELF-DETERMINATION

Maximise the freedom of those affected by your design.

Technology is meant to be an extension of human will. It's meant to empower us to achieve goals we otherwise couldn't. Technology can't achieve this goal if it interferes with our freedom. We need to make design choices that support people's ability to make free choices about how they want to live and engage with technology. But remember: maximising freedom doesn't always mean maximising choice – sometimes too much choice can be paralysing.

PR03.

RESPONSIBILITY

Anticipate and design for all possible uses.

Technology is usually designed with a specific use case – or set of use cases in mind. Problems often arise when users deviate from the intended use case. In a lot of cases, it's entirely possible to predict the different ways people will use our designs, if we take the time to think it through. Failing to imagine alternate uses and their implications is risky and unethical. Doing so can alert us to potentially harmful uses we can safeguard against, or potential benefits we can maximise through good design.

PR04.

NET BENEFIT

Maximise good, minimise bad.

The things we build should make a positive contribution to the world – they should make it better. But more than this, we should also be mindful of the potentially harmful side-effects of our technology. Even if it does more good than bad, ethical design requires us to reduce the negative effects as much as possible.

PR05.

FAIRNESS

Treat like cases in a like manner; different cases differently.

Technology designs can carry biases, reflect the status quo or generate blind spots that mean some groups of people are treated negatively on the basis of irrelevant or arbitrary factors such as: race, age, gender, ethnicity or any number of unjustifiable hconsiderations. Fairness requires us to present justifications for any differences in the ways our design treats different user groups. If some groups experience greater harm or less benefit than others, why is this the case? Are our reasons defensible?

PR06.

ACCESSIBILITY

Design to include the most vulnerable user.

Whenever you have intended users and use cases, you also have people who you don't intend to use the technology. This is a risk when design excludes people who might benefit from your design, if you'd only thought of them in the process. Design can reinforce social disadvantage, or it can help people overcome it. But it can only do this if we bear in mind all the possible users, without dismissing some groups as 'edge cases'.

PR07.

PURPOSE

Design with honesty, clarity and fitness of purpose.

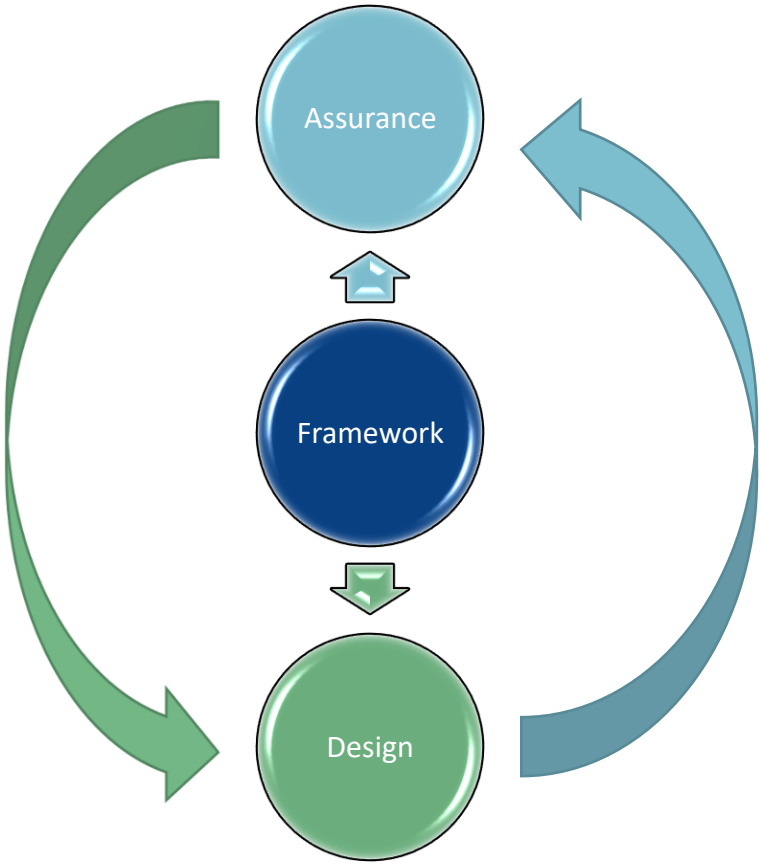
Design is, in one sense, a promise. You are promising to solve a problem your users are experiencing. Like all promises, you should honour this promise. You must be honest and clear about the ability and limitations of your design. Moreover, your design should be tailored to the problem it's trying to solve – and be intended to solve a genuine problem. Good design serves an ethical purpose and does so in efficient and effective ways.

The Framework

Beard and Longstaff talk about

- Framework shaping technology design
- Framework providing standards for assurance processes to use
- Assurance processes feed into design by determining evidence requirements
- Design provides evidence for operational functions

- But AI Ethics auditing has its own problems



Overarching questions

Beard and Longstaff classify questions around the ethics of technology into three primary categories

1. goals
What is the intended goal of building the technology?

2. use

Are the means to achieve a just goal acceptable?

3. unintended effects

Need to account for foreseeable but unintended effects and be able to react to the unforeseen

Intentions

"Although technology is designed with an intention in mind, its use is only ever 'partially bound' by the designer's intention. The innovation and genius that drives technological design also enables (and perhaps encourages) the novel use of artefacts. In part, this is because the best uses for technology aren't always obvious at the moment of design" Beard and Longstaff p 45

Why Principles?

Why do we need principles?

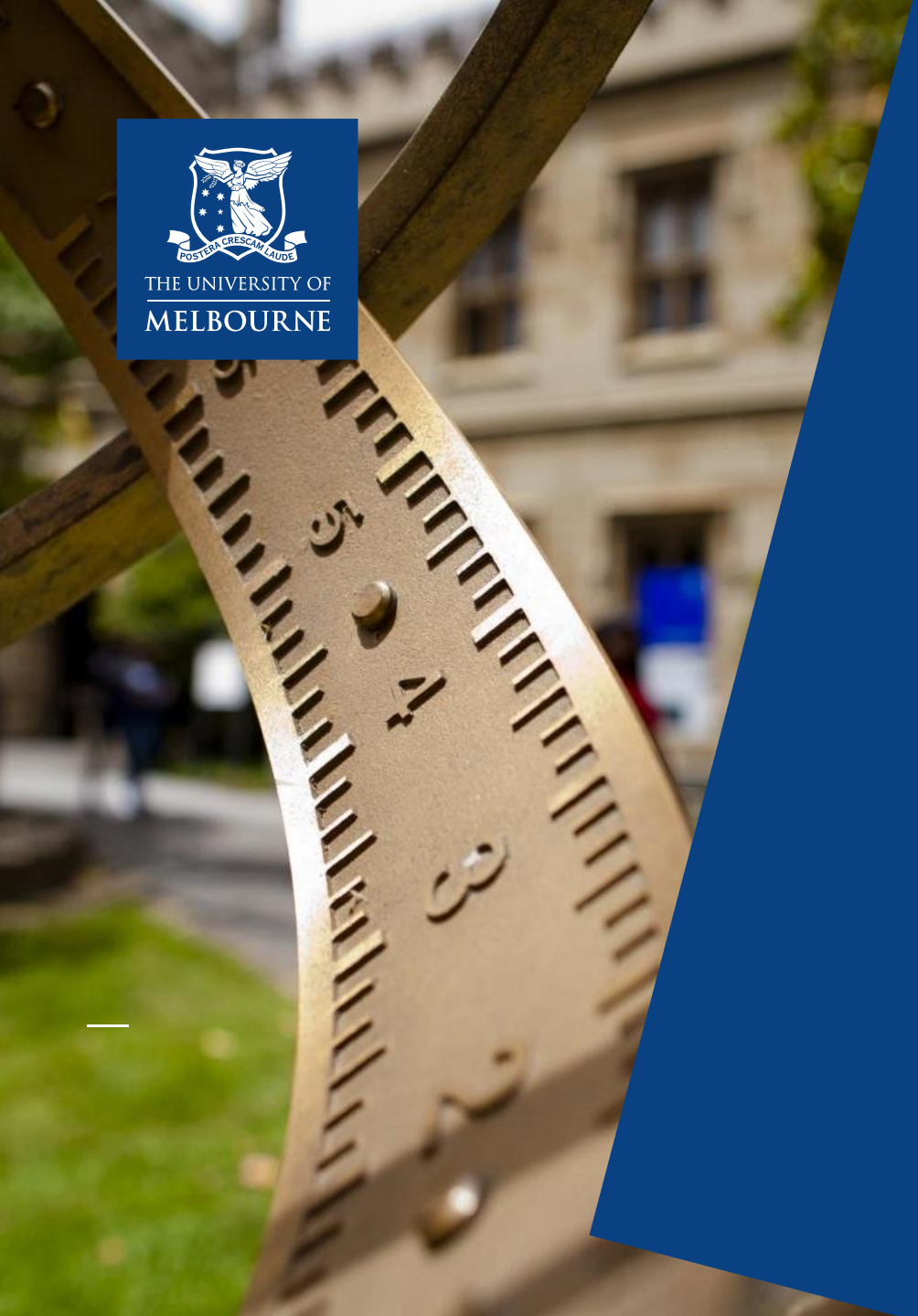
- They are guides to action
  - Allow for a spectrum of values (as opposed to rules)
  - Are somewhat general

Help determine whether technology may be ethically satisfactory

Technology design should not only be compliant with principles, but must be demonstrably so. These demonstrations should be:

- Available (accessible, understandable)
- Capable of being checked by others

Principles 0 - 7



THE UNIVERSITY OF MELBOURNE

Principles 0 - 3

Principles for Good Technology

Principle 0 – Ought before Can

Ought before can.

‘The fact that we can do something does not mean that we should’

Beard and Longstaff take this to be the governing principle of their framework and an understanding of this is also the first learning outcome of this subject

>

Principle 1 – Non-instrumentalism

Non-instrumentalism

- Treat people as ends, not just as means
- Human dignity as its basis
- Not to use people as mere tools
- Do not manipulate people

Design that reduces people to “cogs in the machine” or to just a data source is not ethical design.

Principle 2 – Self-determination

Self-determination

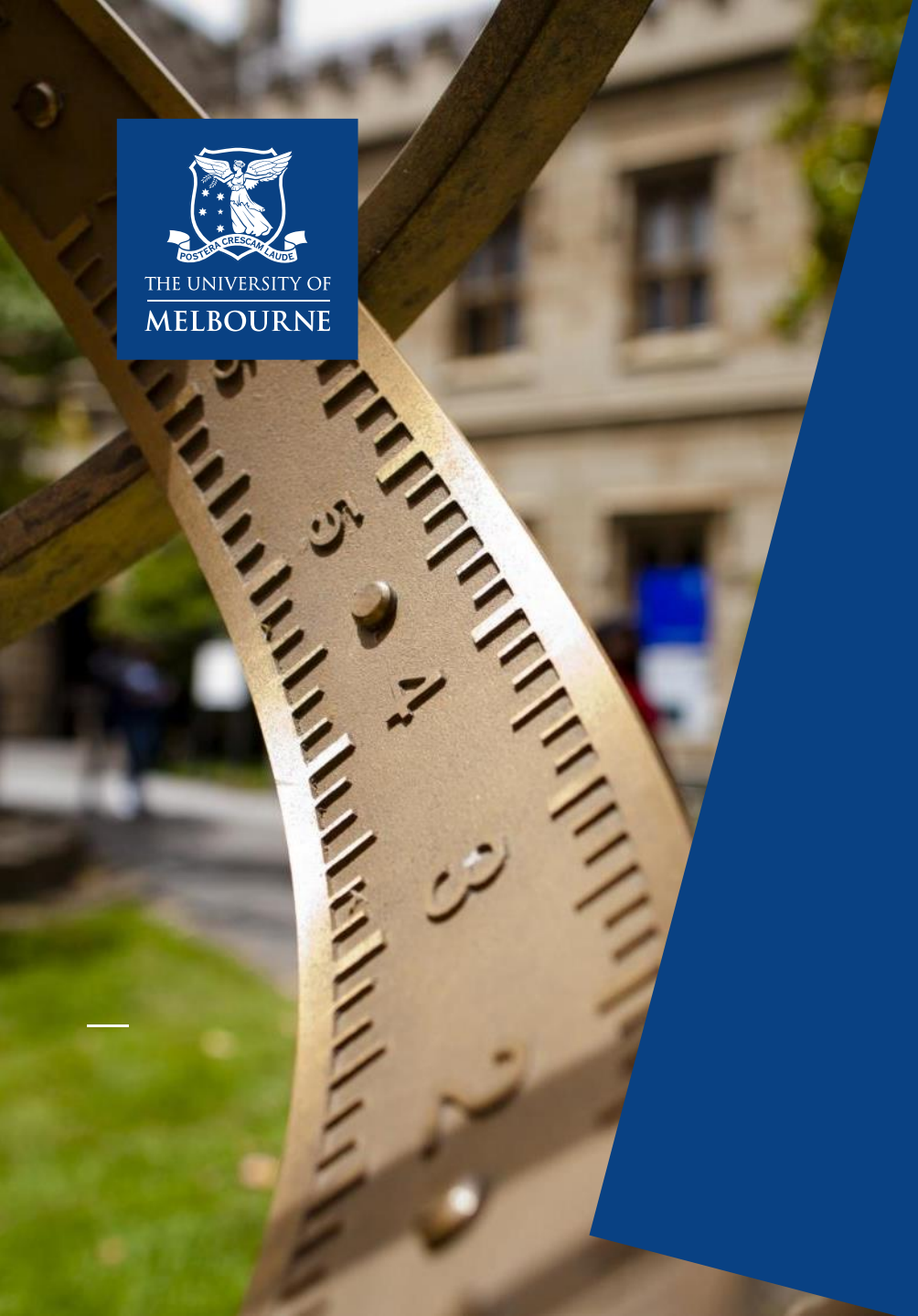
- Autonomy and agency are important
- Beware systems that nudge users or others
- Even good nudges should involve consent.

Q. Who benefits most from the nudging?

Principle 3 - Responsibility

Responsibility

- Minimise ability to repurpose for harm
  - Build this into the design
- Make improvements to mitigate
- Responsibility has limits
- Tools have a bias towards certain uses
- But… alternative uses can be found
- Balancing responsibility requires a utilitarian calculus
- Responsibility is often shared



THE UNIVERSITY OF MELBOURNE

Principles 4 - 7

Principles for Good Technology

Principle 4 – Net Benefit

Net benefit

- There are opportunity costs to consider
  - Prefer those designs that offer more benefits
  - This requires the ability to quantify these…
- Minimise harm
  - At the expense of efficiency
  - At the expense of effectiveness
  - Even where there are large benefits check for harms

A example might be a system that promotes world peace, gained via 24x7 surveillance of the communications of every person. Great benefits, but at what cost to human autonomy and dignity?

Principle 5 - Fairness

Fairness

- Treat like cases alike
  - No arbitrary distinctions
- Biases should be accounted for

- Different treated differently
  - No false equivalence

- Share benefits and burdens
- Most burdens -> most benefit

Limitations of fairness in systems should be transparent to all stakeholders

Principle 6 - Accessibility

Accessibility

- Design to include the most vulnerable user
- cf Rawls' Difference Principle
  - Ensure that the least advantaged receive greater benefits
- You do not always get to pick the user
- Designing (and testing!) for target market could entrench existing inequalities

For example sports events that allow ordering and payment only via smartphones and credit cards excludes many who are poorer or less technologically savvy.

Principle 7 - Purpose

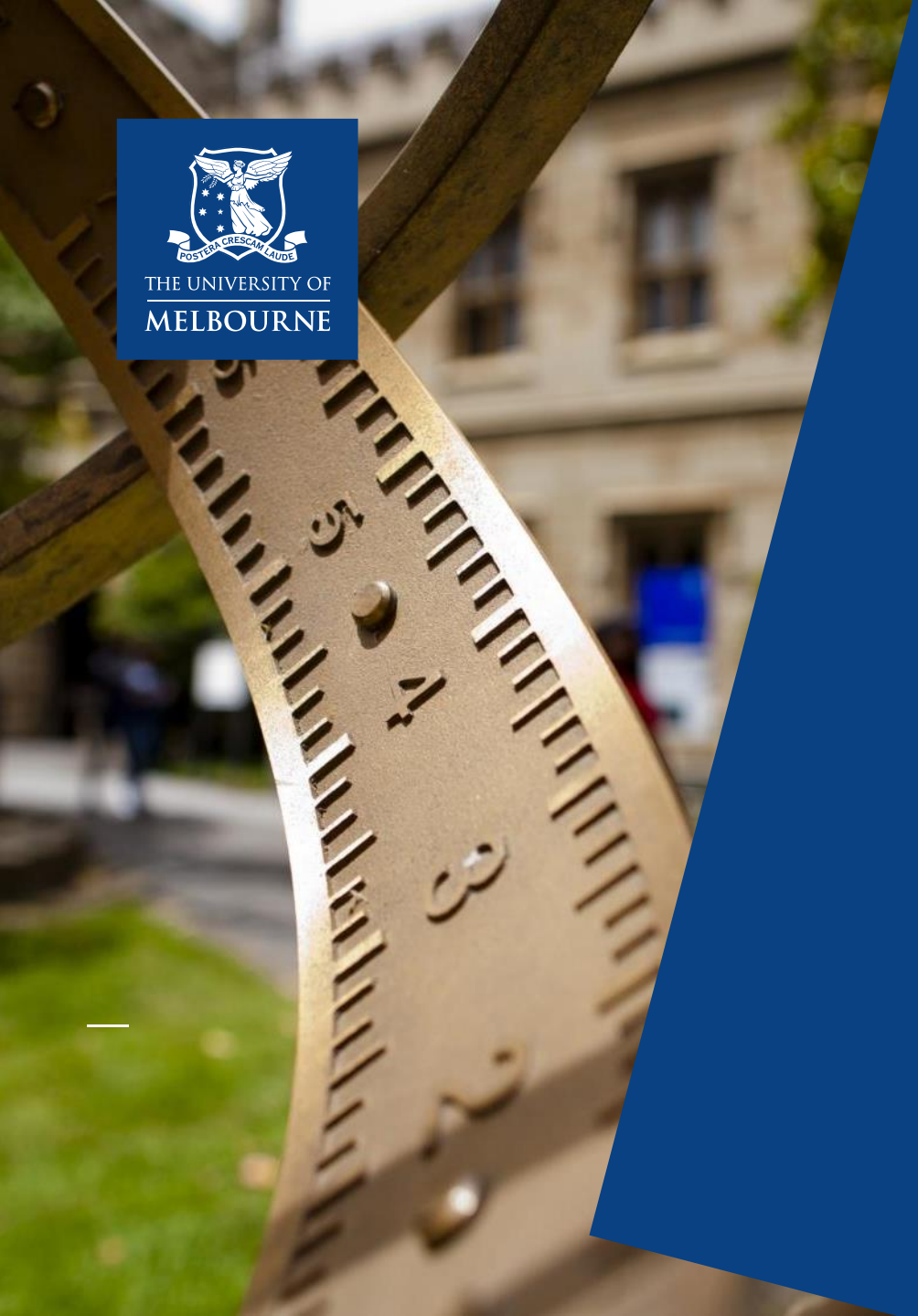
Purpose

- Technology should have purpose
- Be fit-for-purpose
- Be suitable for meeting design goals

That purpose should be

- Legitimate
- Clear
- Honest
- Effective (but still ethical)
- Efficient (but not all costs)

This is a porpoise, not a purpose



To conclude

Final thoughts

Having a framework to guide the design and development of ethical AI is not enough

- You need to care about ethics
- You need to understand something about ethics to use a framework effectively (hopefully this subject has helped)
- You need to think, ponder, reflect
- It’s harder to be ethical in an unethical environment

THE UNIVERSITY OF MELBOURNE

Thank you

Michael WIldenauer

CIS/FEIT/UoM