CHALLENGES FOR AI ETHICS

Dr Jess Whittlestone

Postdoctoral Research Associate at the Leverhulme Centre for the Future of Intelligence, University of Cambridge

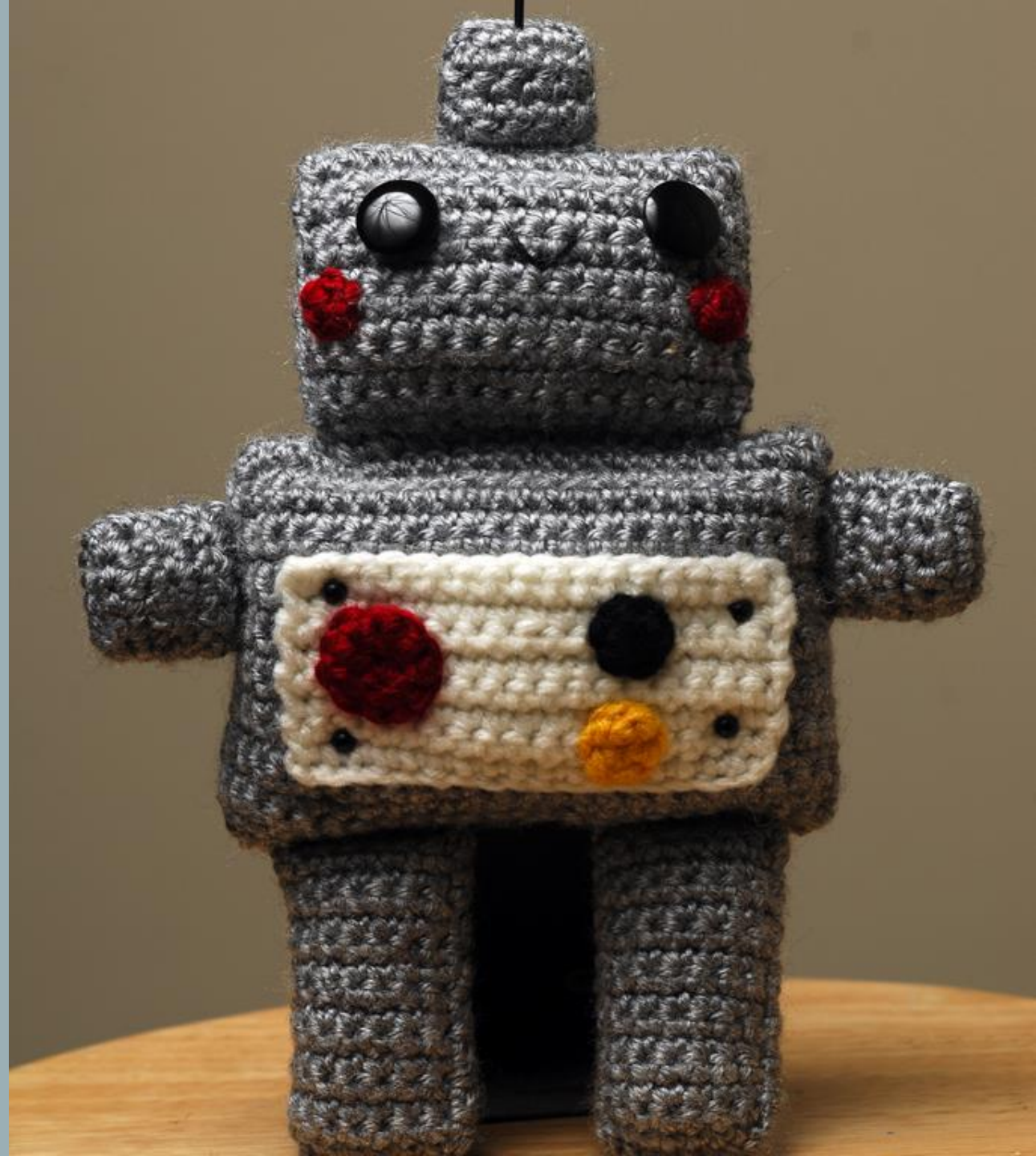
jlw84@cam.ac.uk

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

WHAT IS AI ETHICS?

How can we make sure that AI systems are safe and beneficial, both today and in the long-term?

How can we get the societal benefits of advances in AI while mitigating the risks?



YOU'VE PROBABLY HEARD PEOPLE TALK ABOUT…

- Problems of algorithmic bias and fairness in machine learning
- The importance of making sure machine learning systems are transparent and explainable
- Concerns about personal privacy and autonomy given availability of personal data
- The impact of automation on human capabilities and dignity
- How we can maintain responsibility and accountability as more important tasks and decisions become automated

PRINCIPLES FOR ETHICAL AI

Artificial Intelligence at Google

Our Principles

Objectives for AI Applications

We will assess AI applications in view of the following objectives. We believe that AI should:

1. Be socially beneficial.
2. Avoid creating or reinforcing unfair bias.
3. Be built and tested for safety.
4. Be accountable to people.
5. Incorporate privacy design principles.
6. Uphold high standards of scientific excellence.

ASILOMAR AI PRINCIPLES

usacm ACM US Public Policy Council

Principles for Algorithmic Transparency and Accountability

The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems

An incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies

AI Code

One of the recommendations of the report is for a cross-sector AI Code to be established, which can be adopted nationally, and internationally. The Committee's suggested five principles for such a code are:

1. Artificial intelligence should be developed for the common good and benefit of humanity.
2. Artificial intelligence should operate on principles of intelligibility and fairness.
3. Artificial intelligence should not be used to diminish the data rights or privacy of individuals, families or communities.
4. All citizens should have the right to be educated to enable them to flourish mentally, emotionally and economically alongside artificial intelligence.
5. The autonomous power to hurt, destroy or deceive human beings should never be vested in artificial intelligence.

TOP 10 PRINCIPLES

FOR ETHICAL ARTIFICIAL INTELLIGENCE

UNI global union THE FUTURE WORLD OF WORK

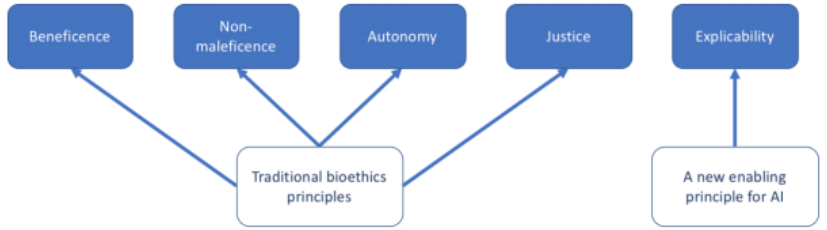


Figure 4.1: an ethical framework for AI, formed of four traditional and one new principle.

*Thanks to Alan Winfield for help identifying recent sets of principles

CLARIFYING KEY CONCEPTS

TRANSPARENCY

For whom?

About what?

For what purpose?

PRIVACY

Is this valued differently across cultures?

Does it mean the same thing across disciplines?

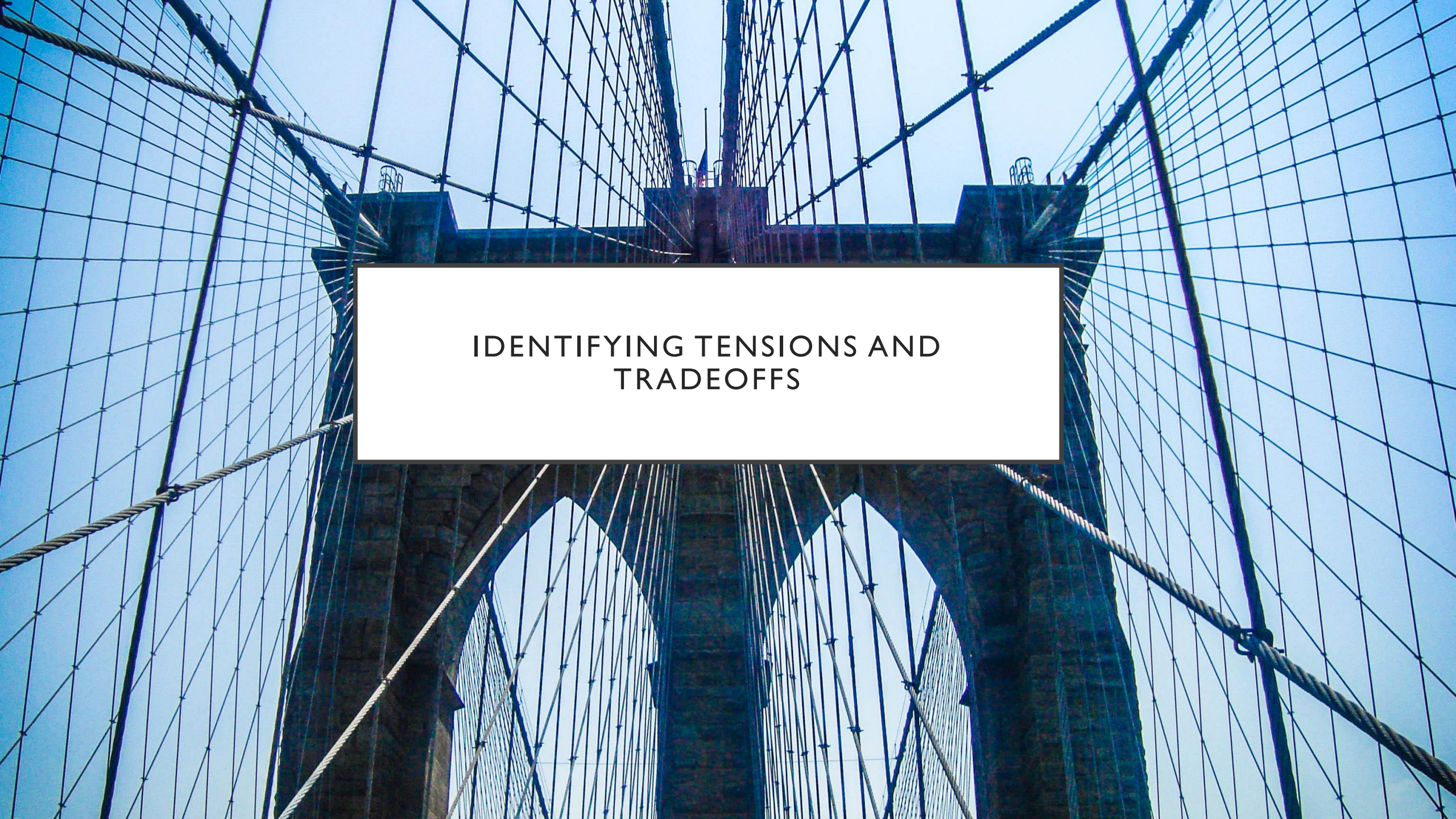
FAIRNESS

Can we agree what this means?

To what extent do disagreements reflect deep underlying political or ethical disagreements?

PROBLEMS WITH KEY CONCEPTS

- **Ambiguity:** the same terms are used to mean many different things, important in different contexts and for different reasons;
- **Differences between disciplines:** what a term means in a technical discipline may not correspond to the legal or popular usage;
- **Differences between cultures:** different cultures may understand and value key concepts quite differently;
- **Complexity:** different interpretations may reflect deeper disagreements between groups who have fundamentally different values or interests.



IDENTIFYING TENSIONS AND TRADEOFFS

"AI should be socially beneficial"

(Google AI principles)

"AI should not be used to diminish the privacy of individuals, communities, or families"

(Lords Select Committee on AI, "AI code")

DeepMind Health

The Information Commissioner, the Royal Free, and what we've learned

3 July 2017

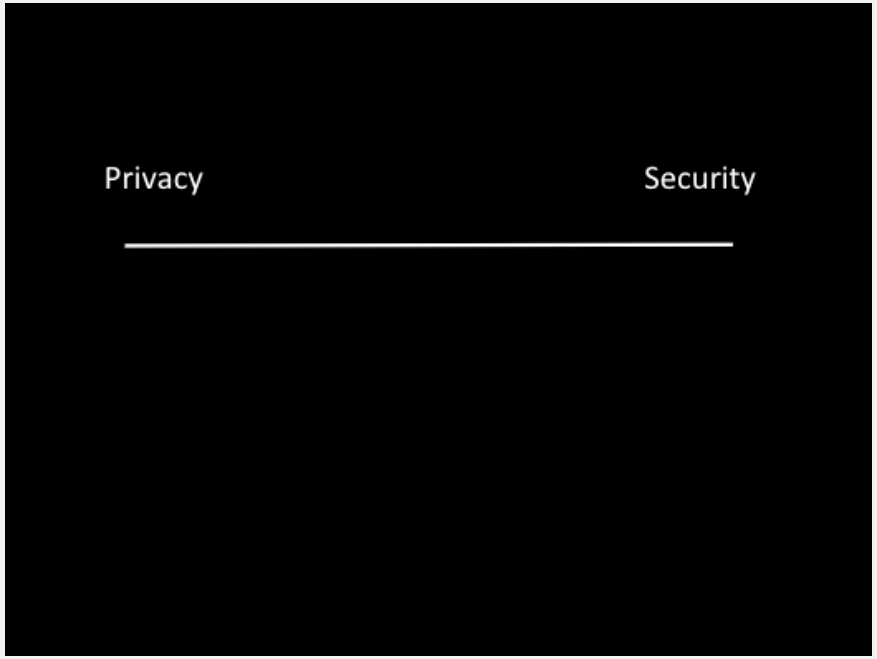
The Information Commissioner's Office has concluded a year-long investigation that focused on the Royal Free's clinical testing of Streams in late 2015 and 2016, which was intended to guarantee that the service could be deployed safely at the hospital. Although the findings are about the decisions made by the Royal Free, we need to reflect on our own actions too.

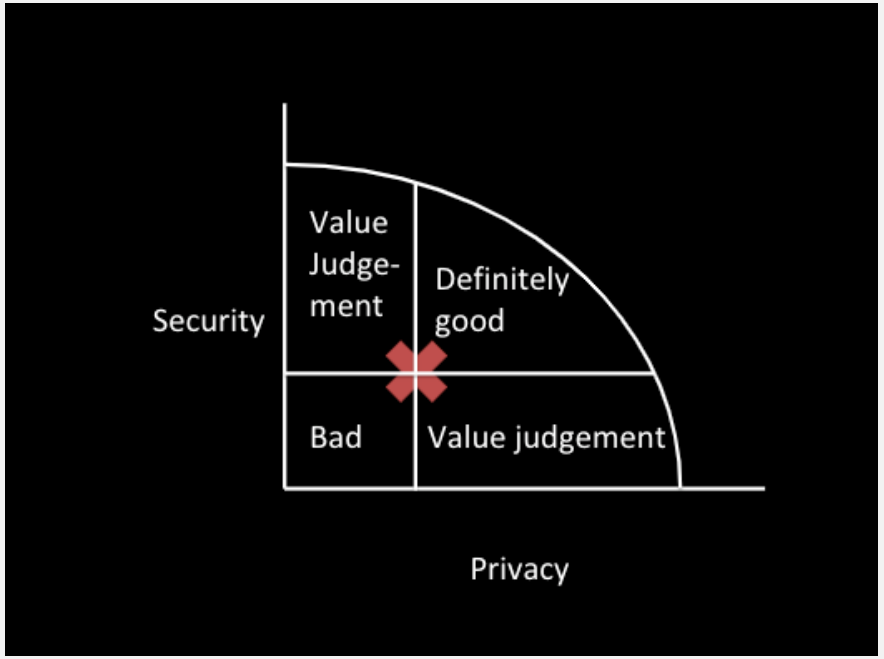
TENSIONS IN AI ETHICS

- Privacy vs. social benefit
- Privacy vs. security
- Accuracy vs. fairness
- Transparency vs. privacy
- Transparency vs. accuracy
- Transparency vs. fairness
- Enhancing vs. devaluing human capabilities
- Enabling vs. threatening human agency
- Short-term efficiency vs long-term safety

How do we get the benefits of AI without threatening important values?

THINKING ABOUT TENSIONS

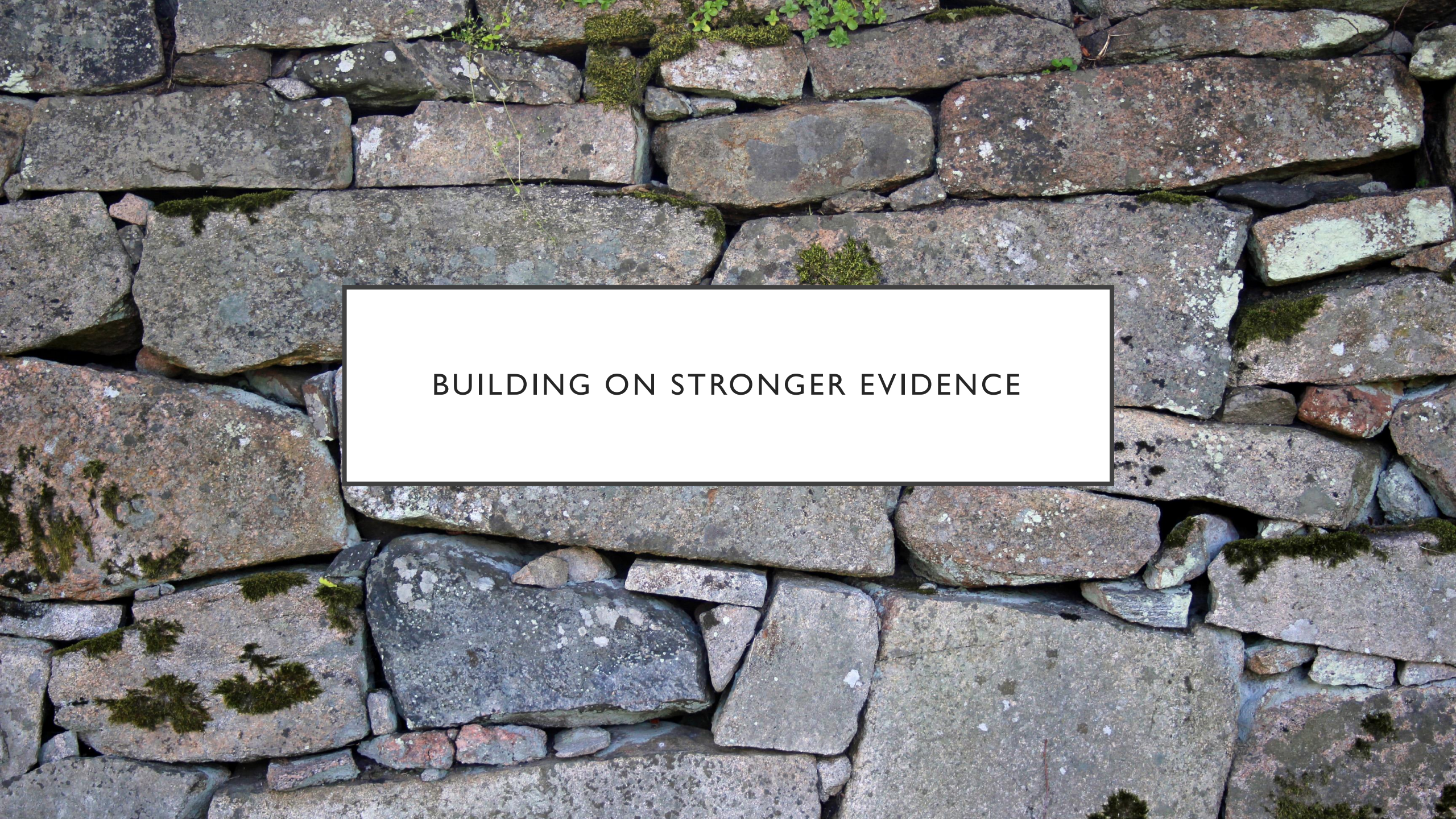




*Thanks to Julian Huppert for these diagrams

MOVING BEYOND PRINCIPLES

- How can we use data to improve the quality and efficiency of services while still respecting the privacy and autonomy of individuals?
- How can we use algorithms to make better decisions while still ensuring fair and equal treatment?
- How can a more granular understanding of society improve individual lives while also enhancing solidarity and citizenship?
- How can we use automation to truly empower people while also promoting their self-actualization and dignity?



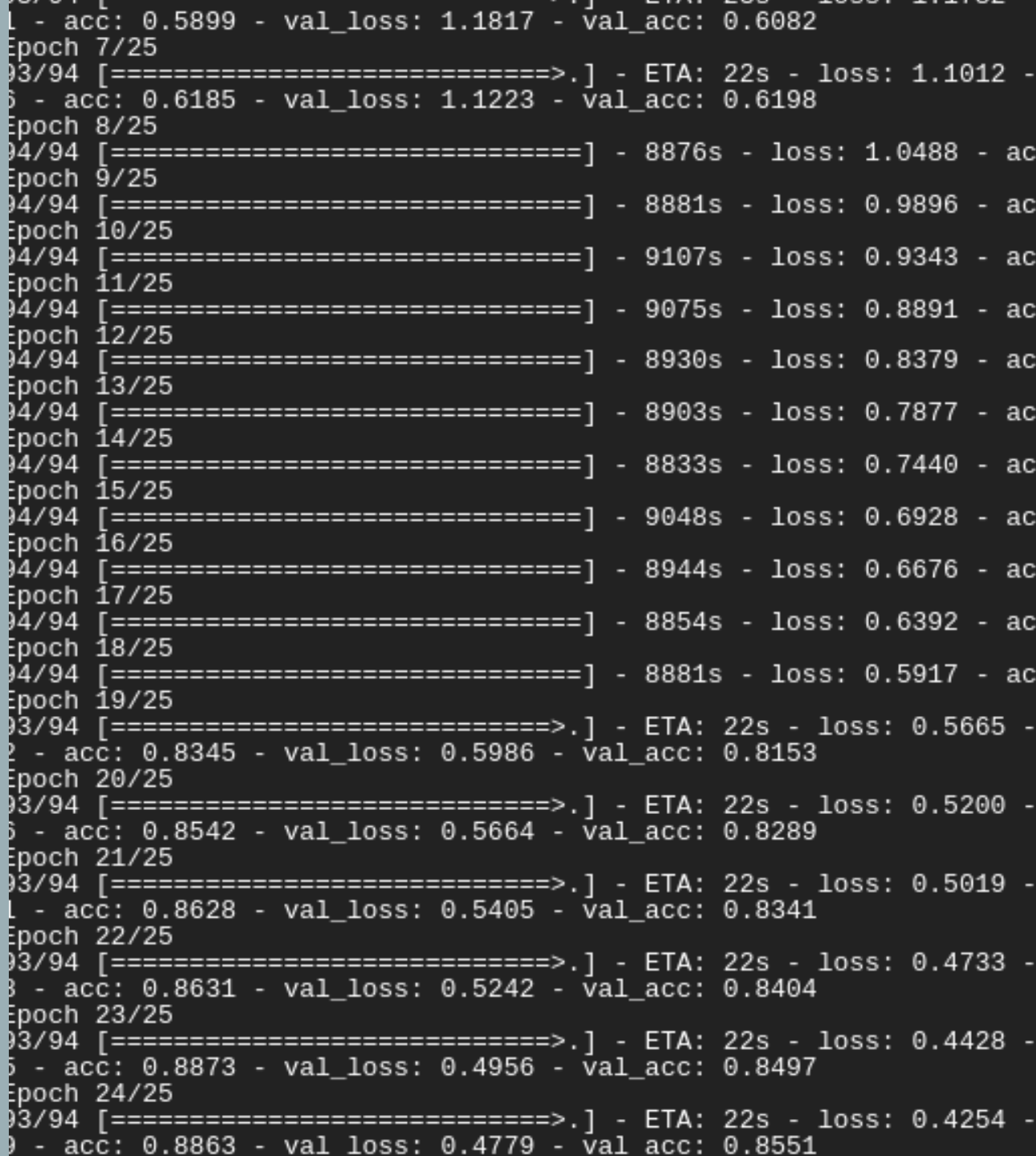
BUILDING ON STRONGER EVIDENCE

UNDERSTANDING TECHNICAL CAPABILITIES

How can knowing what's technically possible help us better understand tensions?

How far can technical methods go to reduce tradeoffs?

Requires more communication with technical experts



UNDERSTANDING SOCIETAL IMPACTS

In what concrete ways are different technologies already impacting society?

To what extent are these tensions reflective of current and potential societal impacts of technology?

Requires more communication with people using AI in practice

UNDERSTANDING DIFFERENT PERSPECTIVES

What are different groups in society most concerned about?

How do different groups think about tradeoffs between values?

Requires more opportunities for the public to input on specific issues

WHAT DO WE NEED NEXT?

- Research to uncover and resolve ambiguities in commonly used terms (e.g. privacy, bias, transparency);
- A focus on identifying key tensions between how technology may both threaten and enhance important values;
- To build on more rigorous evidence about (1) what is technically possible, (2) current uses and impacts of AI in society, and (3) the perspectives of different publics.
- To do all this, we need more collaboration between policymakers and ethicists, technical experts, companies applying AI in practice, and the general public.

THANK YOU!

jlw84@cam.ac.uk

www.lcfi.ac.uk

CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE