CFI LEVERHULME CENTRE FOR THE FUTURE OF INTELLIGENCE

AN INTRODUCTION TO ETHICS AND THE ETHICS OF TECHNOLOGY

DR CLAIRE BENN
CMAB3@CAM.AC.UK


(THE WORLD'S BRIEFEST)
INTRO TO ETHICS

Facts and Values

"The atomic number of hydrogen is 1"

"is"

vs.

"Murder is wrong"

"ought"

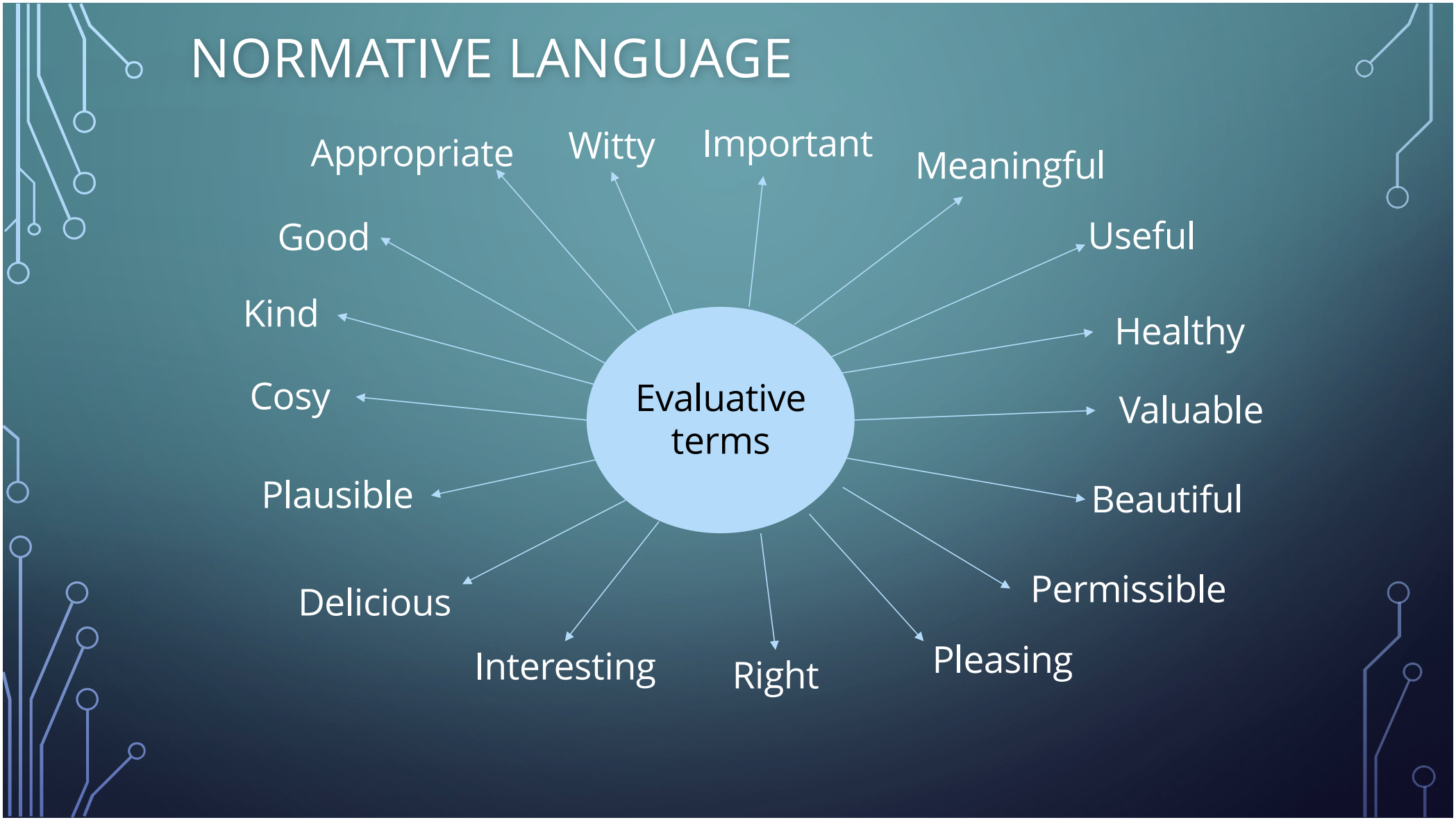
NORMATIVITY

- Leaving aside purely descriptive claims, then:
  - Let's focus on the normative

- Where we evaluate something
  - This might be: an action, a person, an object, an event etc.

- In English, there are lots of evaluative terms

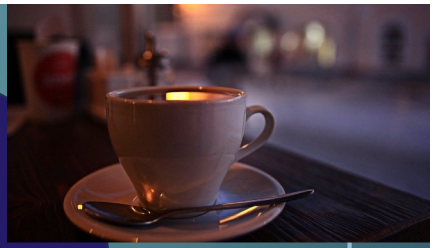
NORMATIVE LANGUAGE



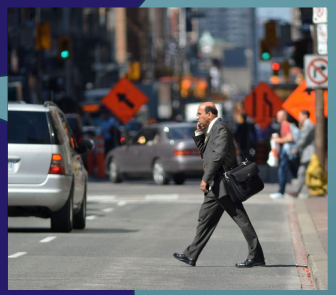
Moral
(Duty, virtue, the good life, wellbeing, etc)

Epistemic (knowledge, rationality, justification)

FACT
FAKE



Practical (goals, effective, efficient)



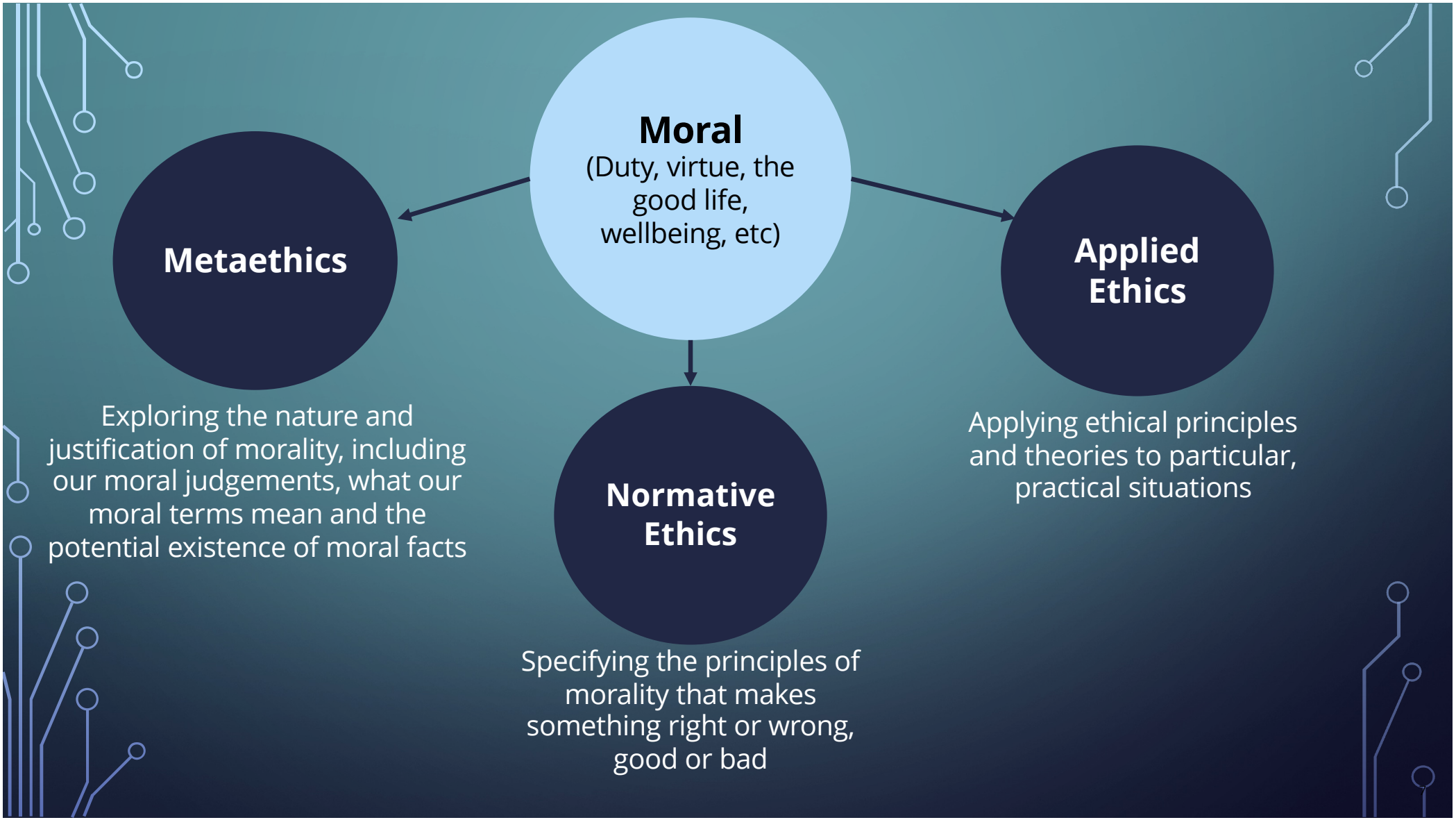
Aesthetic (beauty, art's value)

Legal (law, policy, precedent)

Social (informal norms of acceptable, appropriate behaviour)

Functional (Proper purpose or function)

Etiquette (protocol, politeness, the 'done thing')


Moral
(Duty, virtue, the good life, wellbeing, etc)
Metaethics
Exploring the nature and justification of morality, including our moral judgements, what our moral terms mean and the potential existence of moral facts
Normative Ethics
Specifying the principles of morality that makes something right or wrong, good or bad
Applied Ethics
Applying ethical principles and theories to particular, practical situations

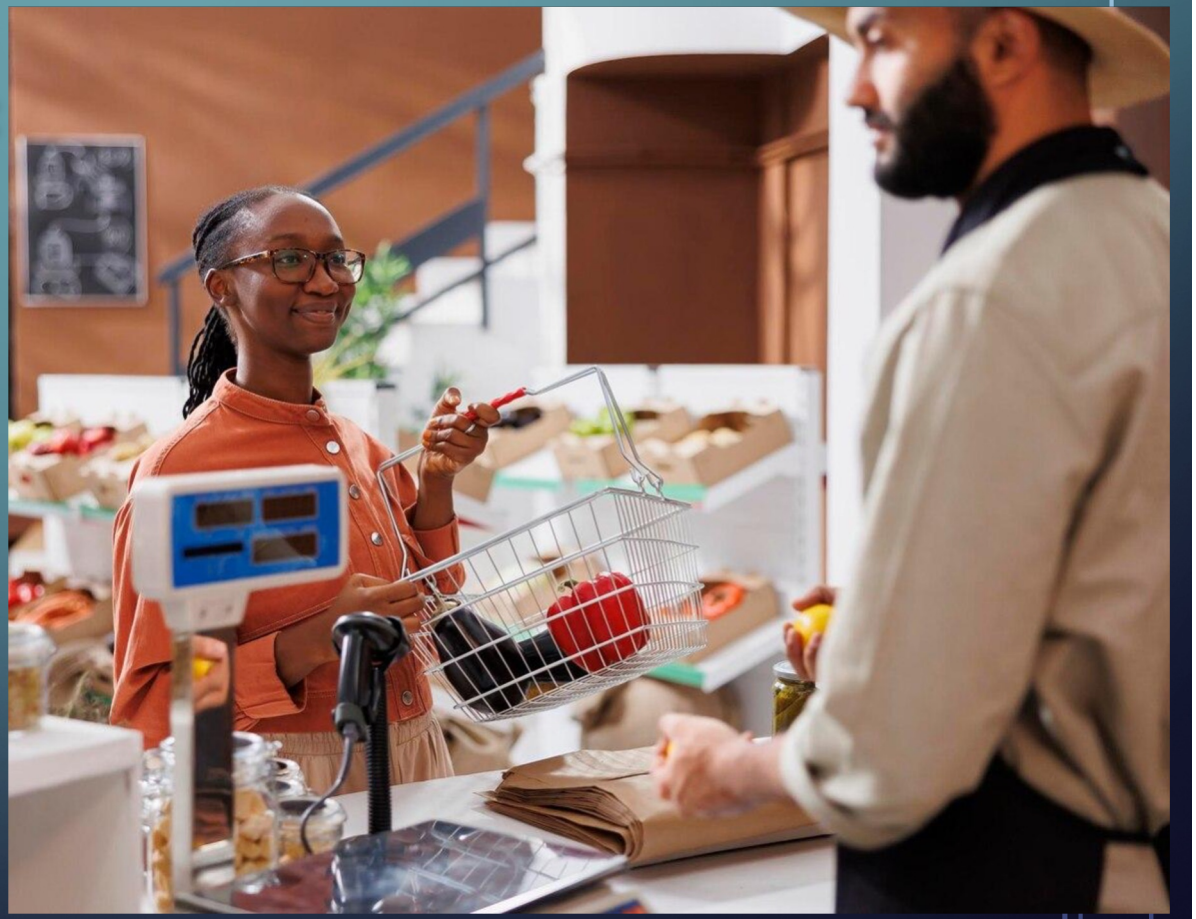
METAETHICS

"MURDER IS WRONG"

- **Semantics**
  - Can the statement be true or false? Or is it more like an emotive expression of disapproval?
- **Metaphysics**
  - What do the normative terms (like 'wrong') refer to? Are there moral facts like there are scientific facts? Can we derive moral facts from non-moral ones?
- **Epistemology**
  - Can we know what is right and what is wrong?
- **Psychological**
  - How do ethical judgements relate to motivation?

NORMATIVE ETHICS

- **Supermarket shopping:**
  - What should you do?
- **My verdict:**
  - You ought not to lie
- **Question:**
  - Why?

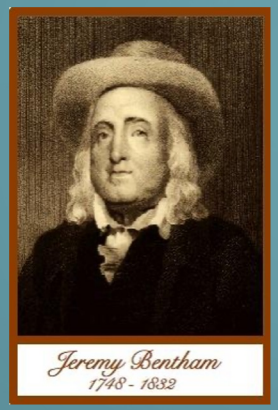


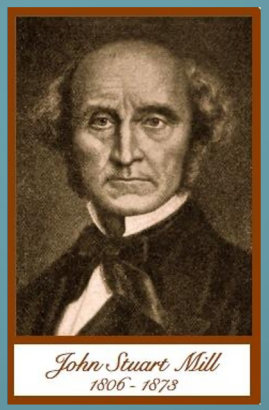
Deontology

Duty, rights, requirements, constraints

Focus: actions

Supermarket: you ought to decide not to lie because there is a principle that 'Lying is wrong'



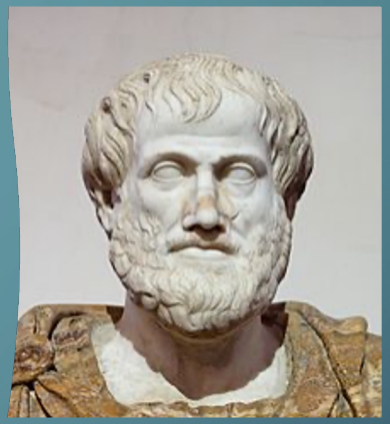


Consequentialism

Good, better, best

Focus: states of affairs/effects of choices

Supermarket: you ought not to lie because lying would have worse consequences compared to telling the truth



Virtue ethics

Praiseworthy, virtuous, responsible

Focus: agent

Supermarket: you ought not to lie because a good person would not lie and you ought to be (like) a good person

Other approaches

- Different **geographic/cultural** approaches

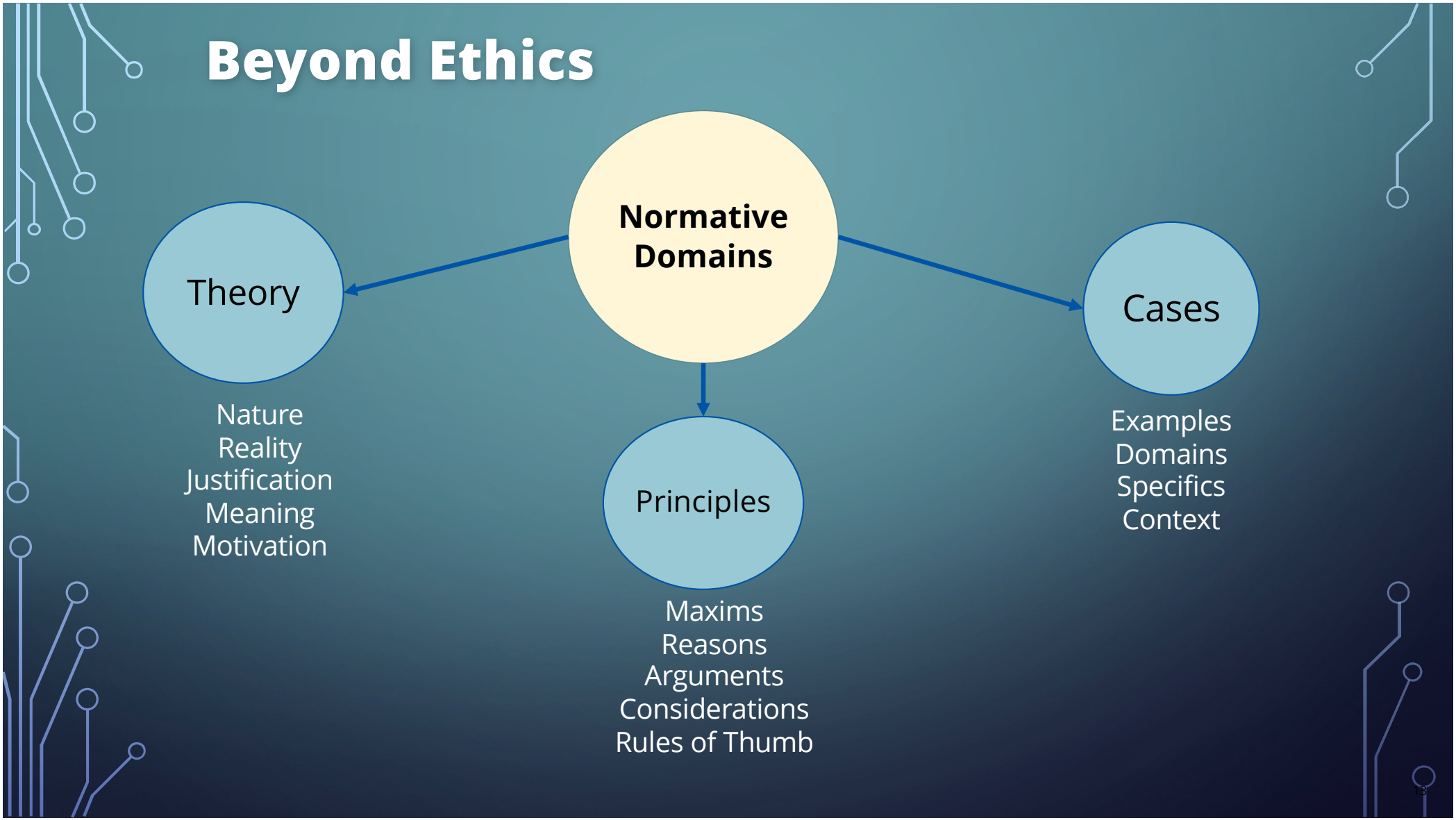
- Eastern ethics, e.g.
  - Buddhism
  - Confucianism
  - Taoism

- Global South ethics, e.g.
  - Indian ethics
  - African ethics
  - Indigenous ethics

- Different sources of **moral authority**, e.g. Religious ethics (including those outside of the Abrahamic tradition)


QUESTIONS IN APPLIED ETHICS
• Are corporations morally responsible for their actions?
• Is testing on animals permissible if it will save the lives of many human beings?
• What are our duties to refugees?
• Should people be required to vote?
• Should all pornography be banned?
• What is wrong with the NSA spying on US and UK citizens?

Beyond Ethics



REFLECTIVE EQUILIBRIUM

Theories & Principles

Particular Judgements

SCIENCE, TECHNOLOGY, AND ORNITHOLOGY

PHILOSOPHY OF SCIENCE

- About scientists
- the method of generating scientific knowledge
- When we ought to accept scientific theories as true
- Nature of evidence


Philosophy of science is to scientists
what ornithology is to birds

PHILOSOPHY OF TECHNOLOGY

- BUT philosophy of science
  - Key to understanding practices (rejecting singular, monolithic conception of 'The Scientific Method')
  - Highlighting the role of values in scientific practice
- Similarly for philosophy of technology:
  - Addresses many key questions
  - My focus: how **normative values** affect
    - What is a technology
    - How we should understand seemingly factual statements about models like accuracy

WHAT IS A TECHNOLOGY

ARTEFACTS

- There is an important question in philosophy of technology:
  - The status and characteristics of artefacts
  - (Similar to question in philosophy of science as to the status and characteristics of scientific models)
- So:
  - What is a technological artefact?

ARTEFACTS AND NATURAL OBJECTS

- What distinguishes artefacts from natural objects?
  - What makes an iPhone different from a tree?

Artefacts have an author or creator (let's focus on *human* creators)

ARTEFACTS AND OTHER MANMADE OBJECTS

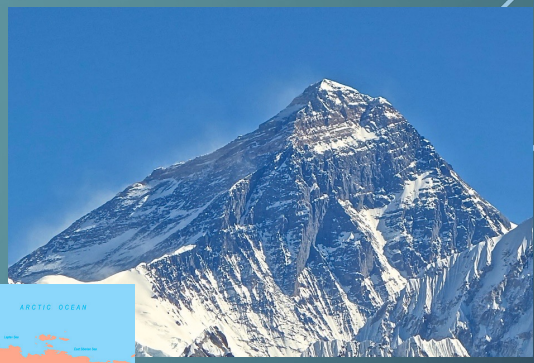
- What distinguishes artefacts from art and waste-products?

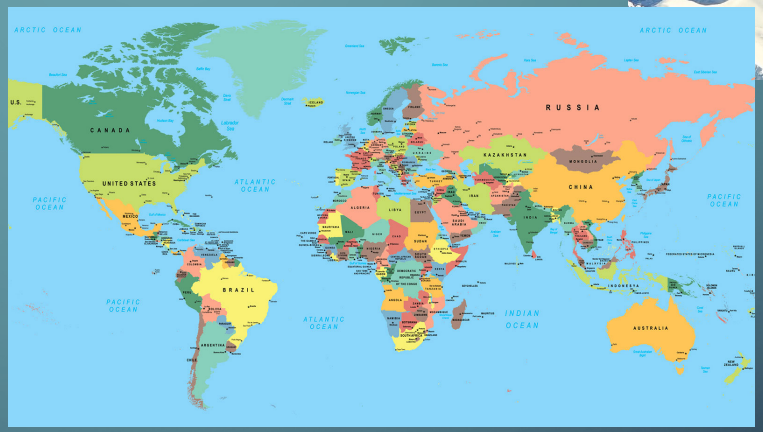
An artefact is made to serve a purpose

FUNCTION

- So one way of articulating what an artefact *is*: to define it in terms of its *function*
- So what is the function of an artefact?
  - Look to the creator's/designer's intentions
- But
  - Biological functions without intentions
  - And the intention is not enough as it might not generally be able to serve that function at all

SEARLE'S SOCIAL ONTOLOGY





- Physical reality
  - Physically existing objects
  - Physical facts
  - Objective

- But there is also a social reality
  - Socially existing objects
  - Social facts
  - Partly objective, but socially constructed

COLLECTIVE INTENTIONALITY

- Searle: “It is a screwdriver only because people use it as (or made it for the purpose of, or regard it as) a screwdriver”

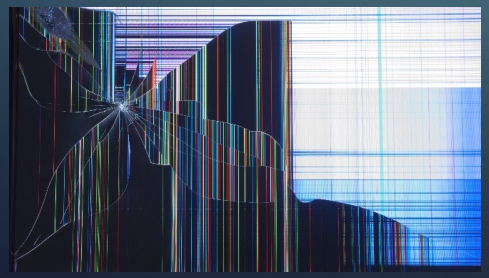
- These are ordinary imposed functions that give rise to ‘ordinary social facts’
  - These have to be able to do the job they are intended to do

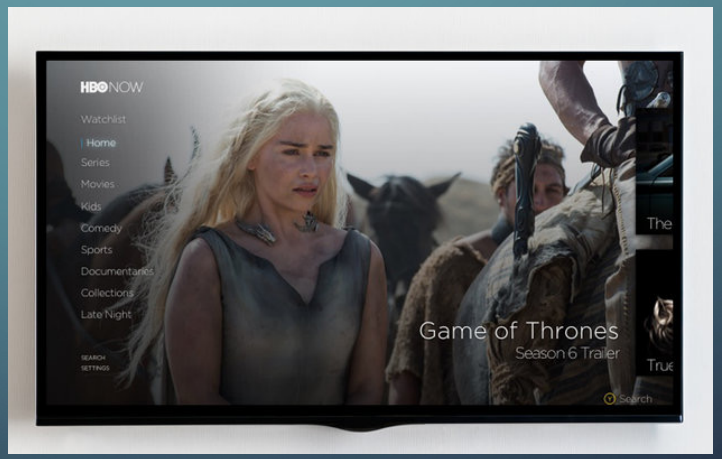
- But then there are status-functions and these give rise to institutional facts
  - E.g. money – these are not based in their physical structure and can be multiply realisable

FUNCTIONS AND NORMATIVITY

- Functions bring with them normative considerations:

- Proper functioning
- Malfunction
- Accidental functions





TECHNOLOGY AND NORMATIVITY

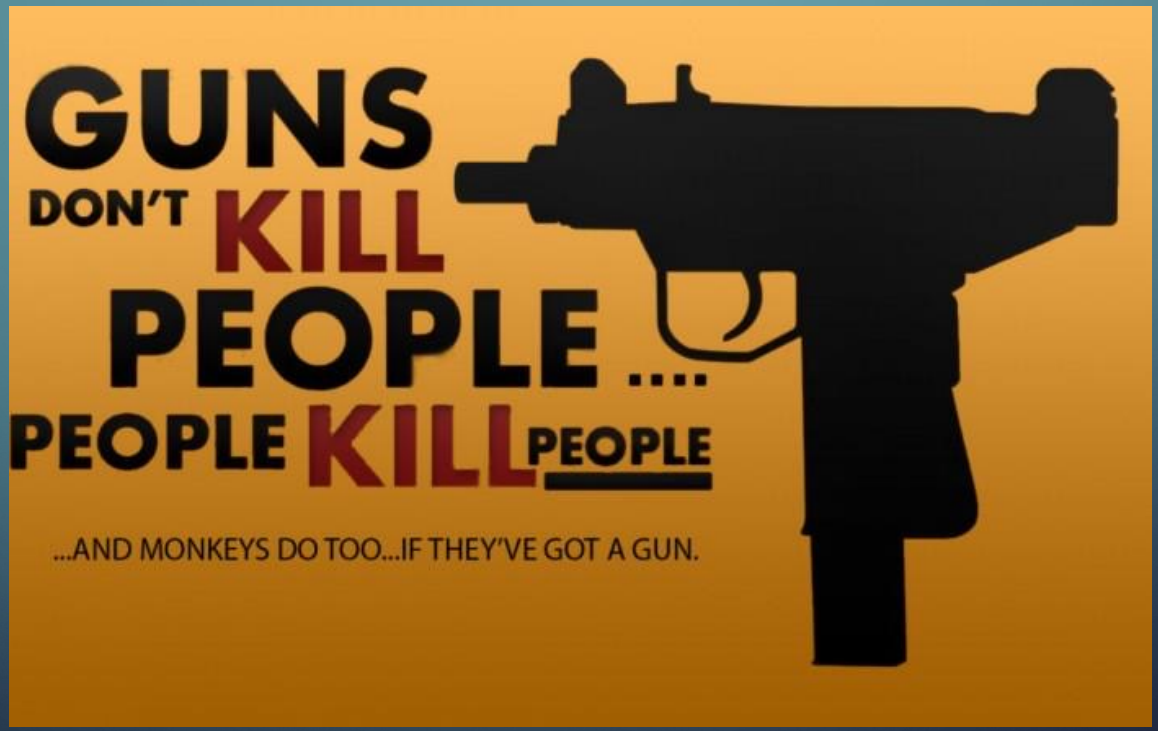
- Thus, what a technological object *is* (whether this is an artefact of this kind)
  - Requires our normative engagement
- What is the *proper functioning* of this object?
- Who made it and for what? What is it treated as? What can it do?

- Is it a good X, a rubbish X or not even an X at all?


THE ETHICS OF
TECHNOLOGY

THE NEUTRALITY THESIS

- Technological artefacts are simply neutral
- They do not have any ethical status in and of themselves though they can be used for good or bad

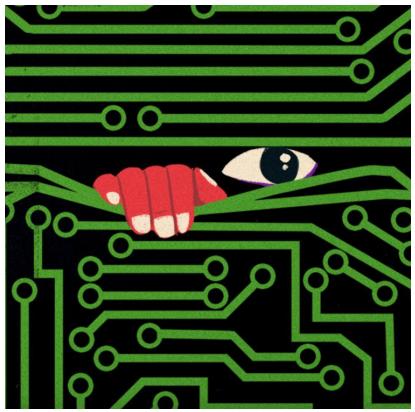

GUNS
DON'T KILL
PEOPLE ....
PEOPLE KILL PEOPLE
...AND MONKEYS DO TOO...IF THEY'VE GOT A GUN.

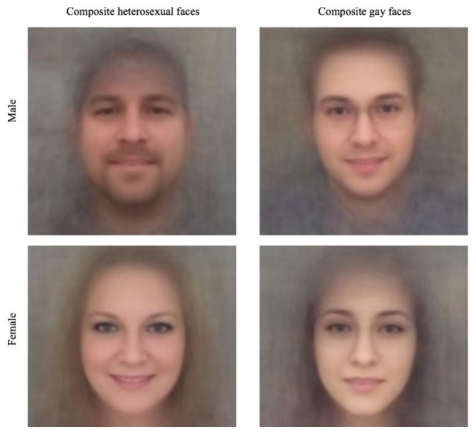
VALUE-FREE IDEAL

- Lots of work has been done debunking this
  - In science
  - About technological objects

CLASSIC ETHICS ISSUES

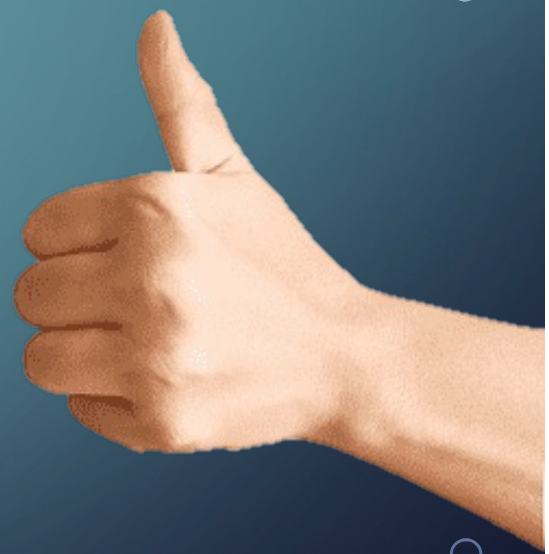
- The initial data collection is bad
  - Privacy violations
- The model is bad
  - Down-ranking the CVs of women
- The deliberate use of the model is bad
  - An authoritarian regime uses a ML tool that classifies people as LGBTQ+ based on their face
- The unintentional use of the model is bad
  - Over-policing





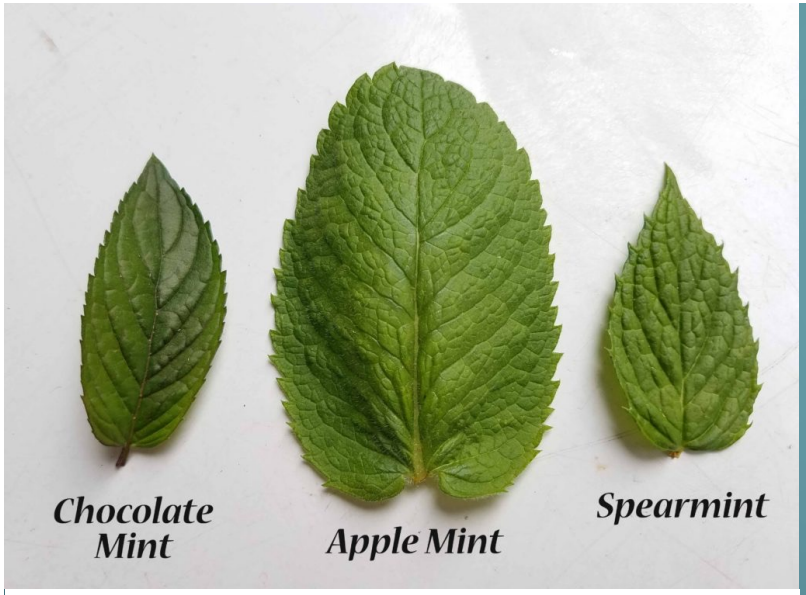
BUT LET US SET ALL OF THAT ASIDE AND ASSUME

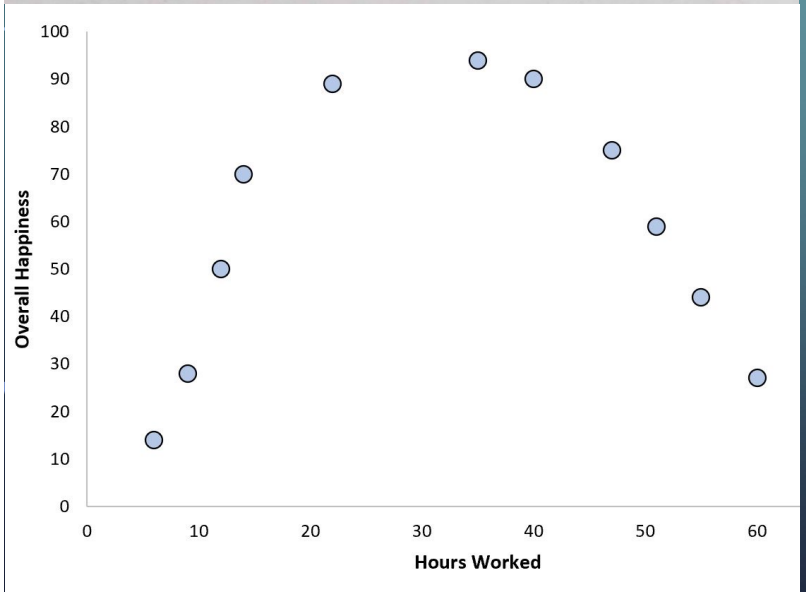
- The data collection done right
- There is no concerning bias
- The use for which the model is made is fine
- The unintended use of the model is fine



- We now just have a series of predictions
- Whether the model is any good is just a matter of **accuracy**
- AND that isn't a matter of ethics, right?


THE ETHICS OF ERRORS





ERRORS

- Models
  - won't have examples of every instant
  - Else wouldn't be very useful (we would already know the answer)
- Therefore, we have to extrapolate (or interpolate) to new instances
- This will involve some errors
- The model is highly unlikely to be 100% accurate all the time (especially if it is to be useful)

CHIHUAHUA CLASSIFICATION

False Negative

Not a Chihuahua

Chihuahua

False Positive

True Negative

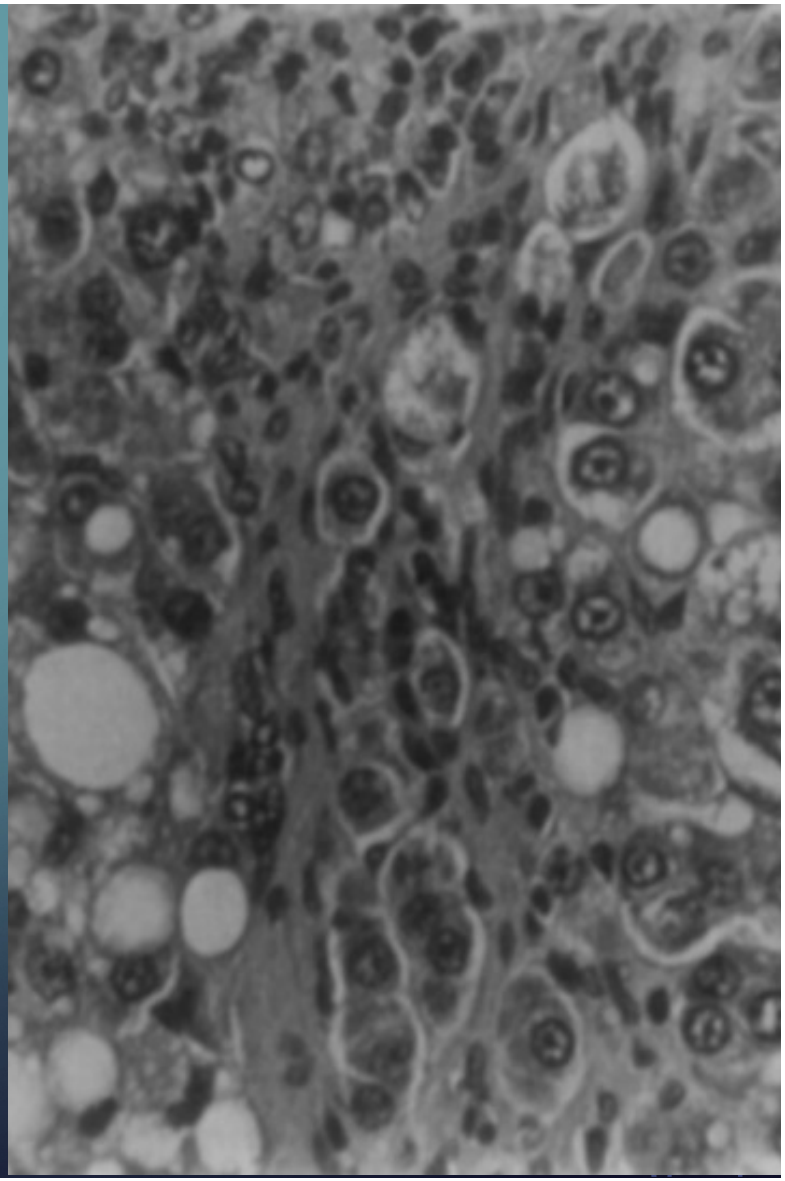
Not a Chihuahua

Chihuahua

True Positive

CHOOSING BETWEEN ERRORS

- Accuracy
  - just choose the one with the least errors!
- BUT equally accurate:
  - 0 False positives + 10% False Negatives
  - 10% False positives + 0 False negatives
- OR ambiguous cases
  - All borderline cases are X
    - more false positives
  - All borderline cases are not-X
    - more false negatives



PRECISION

$$= \frac{True\ Positives}{True\ Positives + False\ Positives}$$

- This measures the proportion of correctly identified positive results (true positives) out of all results that were predicted to be positive.
- How many of the things you identified as True were actually True?

RECALL

$$= \frac{True\ Positives}{True\ Positives + False\ Negatives}$$

- This measures the proportion of correctly identified positive results out of all actual positive instances.
- How many of the things that are True did you identify as True

F_1 SCORES

- We can combine Precision and Recall

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- This is the harmonic mean of Precision and Recall

F-BETA SCORES

- But sometimes we will want to give different weights to Precision and Recall

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}$$

- β determines the weight given to recall relative to precision
- If it equals 1: Precision and recall are weighted equally
- If β > 1: Recall is weighted more heavily
- If β < 1: Precision is weighted more heavily

- The F_2 score for example given four times more weight to recall than the F_1 score

HOW TO WEIGH THEM?

- Choices have to be made

- This depends on
  - the costs of the different kind of errors
  - Whether some errors are much *more* costly than others
  - And who bears those costs in each case

WHICH TO MINIMISE AND WHY?

1. You have some pictures of mushrooms: some are delicious and some are poisonous. Prediction: **delicious mushrooms**
2. You are engaged in an ecological preservation study for endangered birds. The sightings of these birds are from afar and brief as they flit through the trees. It is therefore hard to tell which are the endangered Green Finch and which the more common Siskin. Prediction: **Green Finch**
3. You have some grainy live feeds of possible housefires in a densely populated city and you have to decide if it is warranted to send out one of the city's 5 fire engines. Prediction: **House fire**
4. You have to determine if someone has an otherwise impossible to identify, serious medical condition, the treatment of which is risk-free. Prediction: **Medical condition**

WHY IS THIS IMPORTANT?

- It was easy to think that if we set aside bad data, injustice in the world and bad actors, that we were left with purely epistemic questions:
  - Accuracy
- But answering the very question of accuracy requires non-epistemic values (for all interesting, useful models at least)
- This applies to all moves from data to a model
- And, in fact, all of predictions!

Heather Douglas: ‘Inductive Risk’

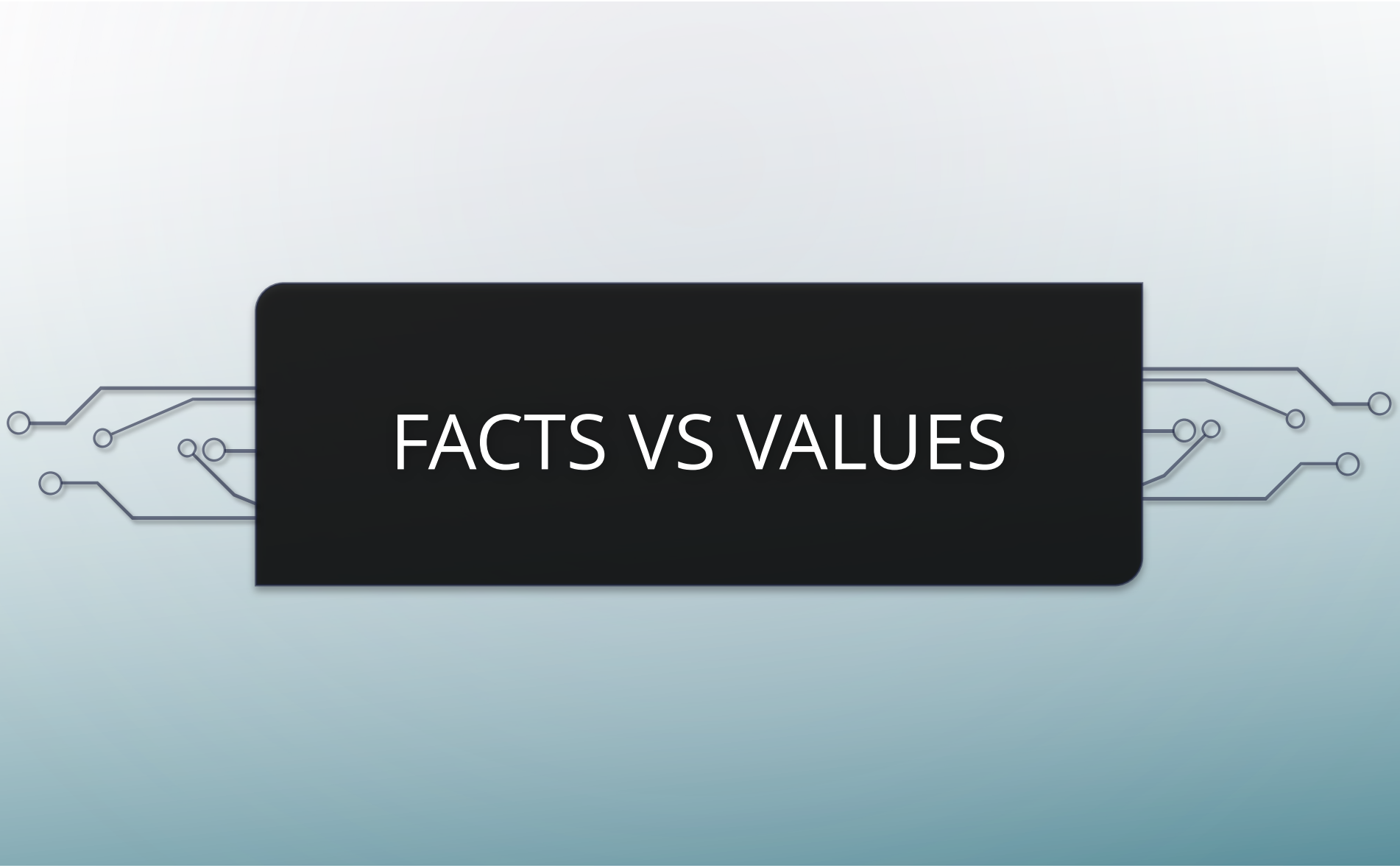
- Heather Douglas revived the discussion of inductive risk to make an argument with huge implications for science:
  - Scientists accept or reject hypotheses.
  - These hypotheses are not deductive but inductive
  - As such, they risk false positives or false negatives
  - The trade-off between these errors can only be made by appeal to the non-epistemic costs associated with each
  - Therefore, scientists must appeal to non-epistemic values in scientific inference

FROM SCIENCE TO TECHNOLOGY

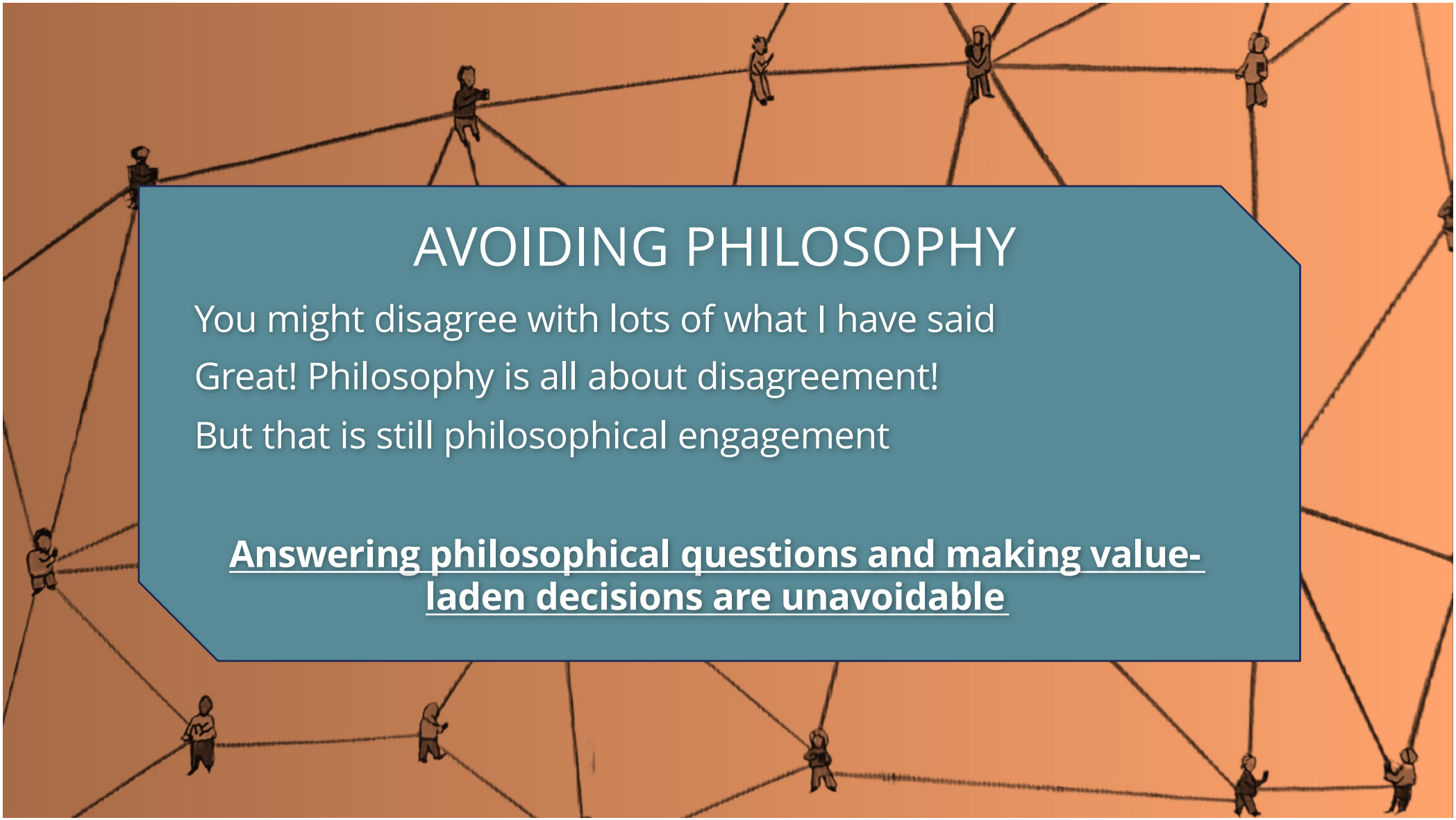
- Douglas's argument was about science
- But it applies also to all model predictions including those used in Computer Science
  - Whenever we are making predictions (i.e. estimating an unknown value on the basis of other, known values) there is the possibility of errors
  - The significance of these errors and their costs determine how we weigh them against each other
  - But this requires us to engage in ethical, social and political questions about who ought to bear the burden of those errors
  - Also: even if we all agree or something seems like commonsense doesn't mean it isn't value-laden. It just means that we all agree on the relevant values or that the values are implicit in the decision

VALUES IN SCIENCE AND TECHNOLOGY

- We might be happy to accept that values matter
  - When it comes to choosing what to work on/create
  - When we evaluate what these models/artefacts are used for
- But what the argument from inductive risk shows
  - The very business of building and evaluating models is inherently value-laden
- This isn't bad!
  - Doesn't mean we shouldn't build them
  - Doesn't mean that we can't work out how to proceed
  - But it does force us to acknowledge the roles that values play in determining the foundations and fundamental practices


FACTS VS VALUES

- We began with a clear split between fact and values
  - Between ‘is’ and ‘ought’ statements
- More intertwined than is often assumed
  - What a technology *is* can be a normative matter
  - How we ought to construe the accuracy of a model requires us to consider normative values


AVOIDING PHILOSOPHY
You might disagree with lots of what I have said
Great! Philosophy is all about disagreement!
But that is still philosophical engagement
Answering philosophical questions and making value-laden decisions are unavoidable

THANKS FOR LISTENING!

Dr Claire Benn

Leverhulme Centre for the Future of Intelligence

University of Cambridge