Lecture 10: AI & Ethics

6CCS3AIN

Stefan Sarkadi

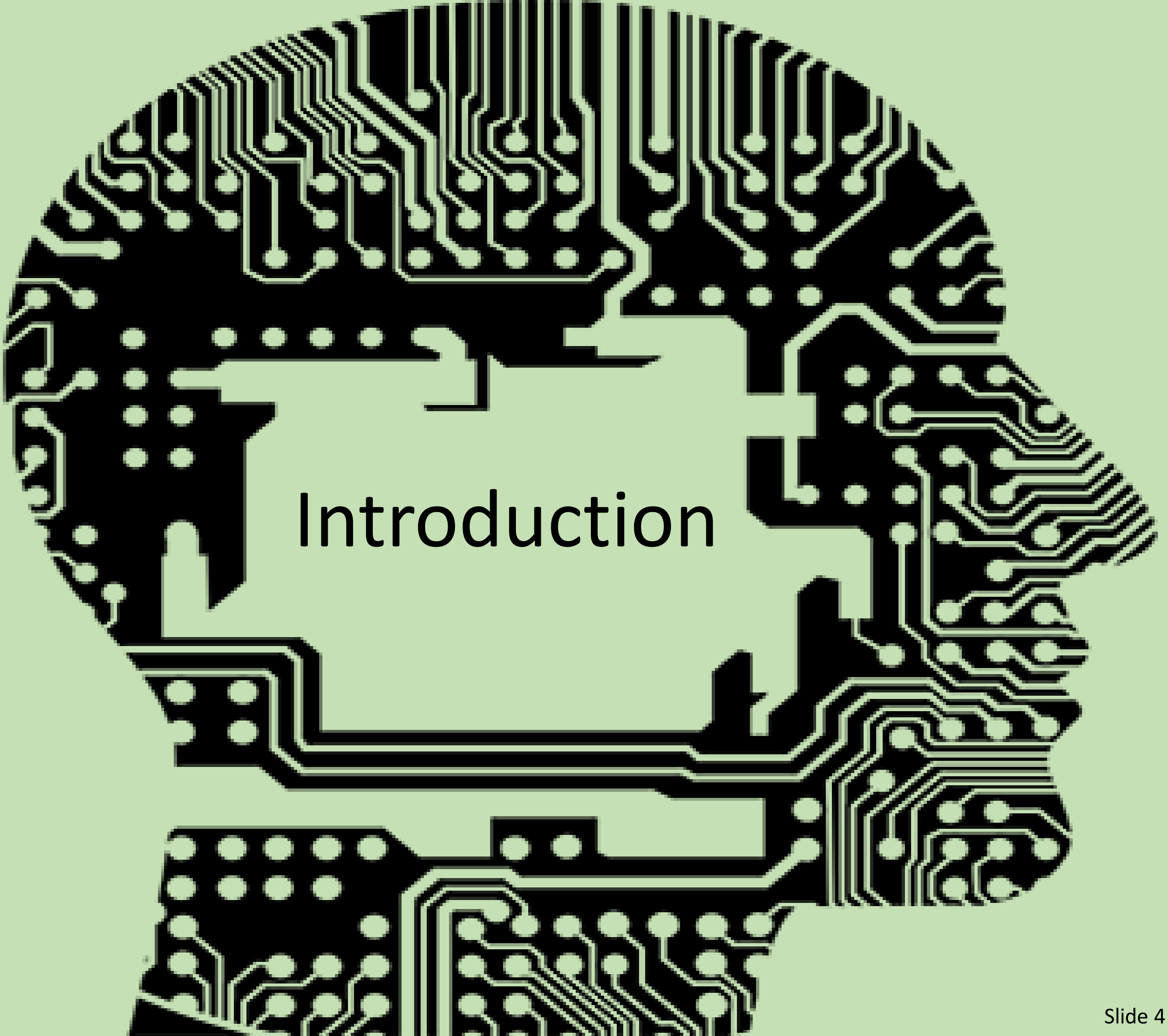
(with thanks to Elizabeth Black and Peter McBurney)

Today

1. Introduction to Artificial Intelligence (FMT)
2. Probabilistic Reasoning I (FMT)
3. Probabilistic Reasoning II (FMT)
4. Sequential Decision Making (FMT)
5. Temporal Probabilistic Reasoning (FMT)
6. Game Theory (FMT)
7. Argumentation I (PMcB)
8. Argumentation II (PMcB)
9. (A peek at) Machine Learning (PMcB)
10. AI & Ethics (SS)

Today

1. Introduction
2. Ethical Theories
3. Principles for guiding AI
4. Research in AI & Ethics
5. Discussion: How worried should we be?



Introduction

Introduction

Premises about machines:

- Are intelligent, interconnected, ubiquitous and autonomous;
- Have access to vast amounts of diverse data;
- Have power to make decisions that can significantly affect both the lives of individuals and society as a whole.

Examples: autonomous vehicles, algorithms that decide whether you can get health insurance, search algorithms, Facebook news feed algorithm.

Introduction

Global research priority: *how to ensure that we reap the benefits from advancing AI technologies, while guarding against the threats they pose for humanity?*

“AI holds the potential to be a major driver of economic growth and social progress, if industry, civil society, government, and the public work together to support development of the technology with thoughtful attention to its potential and to managing its risks.”

U. S. Government’s 2016 report on *Preparing for the Future of AI*.

Introduction

Global research priority: *how to ensure that we reap the benefits from advancing AI technologies, while guarding against the threats they pose for humanity?*

"… need to improve the intelligibility of their AI systems … ensure that our use of AI does not inadvertently prejudice the treatment of particular groups in society."

UK House of Lords Select Committee on Artificial Intelligence report published in April 2018.

Introduction

[Charter of Fundamental Rights of the European Union]

EUROPEAN COMMISSION FOR THE EFFICIENCY OF JUSTICE (CEPEJ) has adopted the first European text setting out ethical principles relating to the use of artificial intelligence (AI) in judicial systems. December 2018

- **respect for human rights and non-discrimination**
- **principle of quality and security**
- **principle of transparency**
- **make the user an enlightened agent**

[See CEPEJ presentation note.]

Introduction

Global research priority: *how to ensure that we reap the benefits from advancing AI technologies, while guarding against the threats they pose for humanity?*

"... form a broad multi-stakeholder platform which will complement and support the work of the AI High Level Expert Group in particular in preparing draft AI ethics guidelines, and ensuring competitiveness of the European Region in the burgeoning field of Artificial Intelligence."

The European AI Alliance on Goals of the AI Alliance.

Introduction

Global research priority: *how to ensure that we reap the benefits from advancing AI technologies, while guarding against the threats they pose for humanity?*

"...It is important to research how to reap its benefits while avoiding potential pitfalls."

Open letter on *Research Priorities for Robust and Beneficial AI* (signed by over 8000 people, including world leading scientists and industry leaders)

future of life INSTITUTE

Introduction

Future of Life Institute :

“**The AI is programmed to do something beneficial, but it develops a destructive method for achieving its goal:** … If you ask an obedient intelligent car to take you to the airport as fast as possible, it might get you there chased by helicopters and covered in vomit, doing not what you wanted but literally what you asked for. If a superintelligent system is tasked with a ambitious geoengineering project, it might wreak havoc with our ecosystem as a side effect, and view human attempts to stop it as a threat to be met.”

future of life INSTITUTE

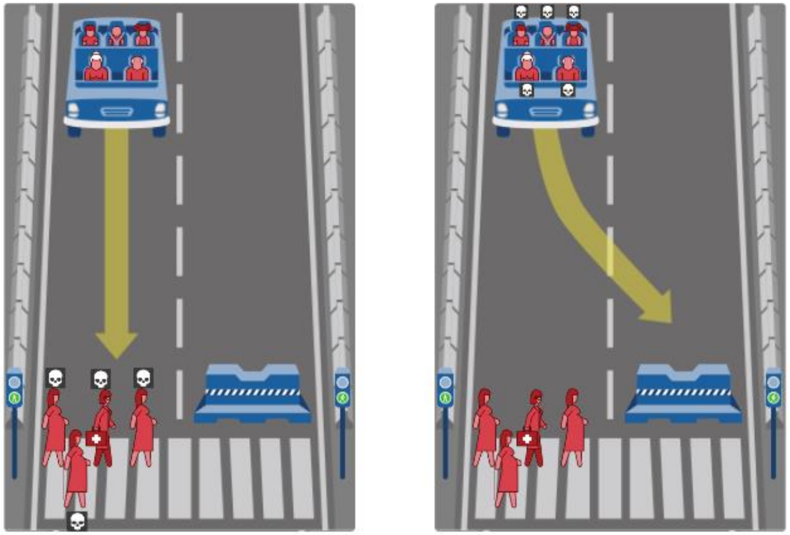
Introduction

Credit: Roo Reynolds, Flikr

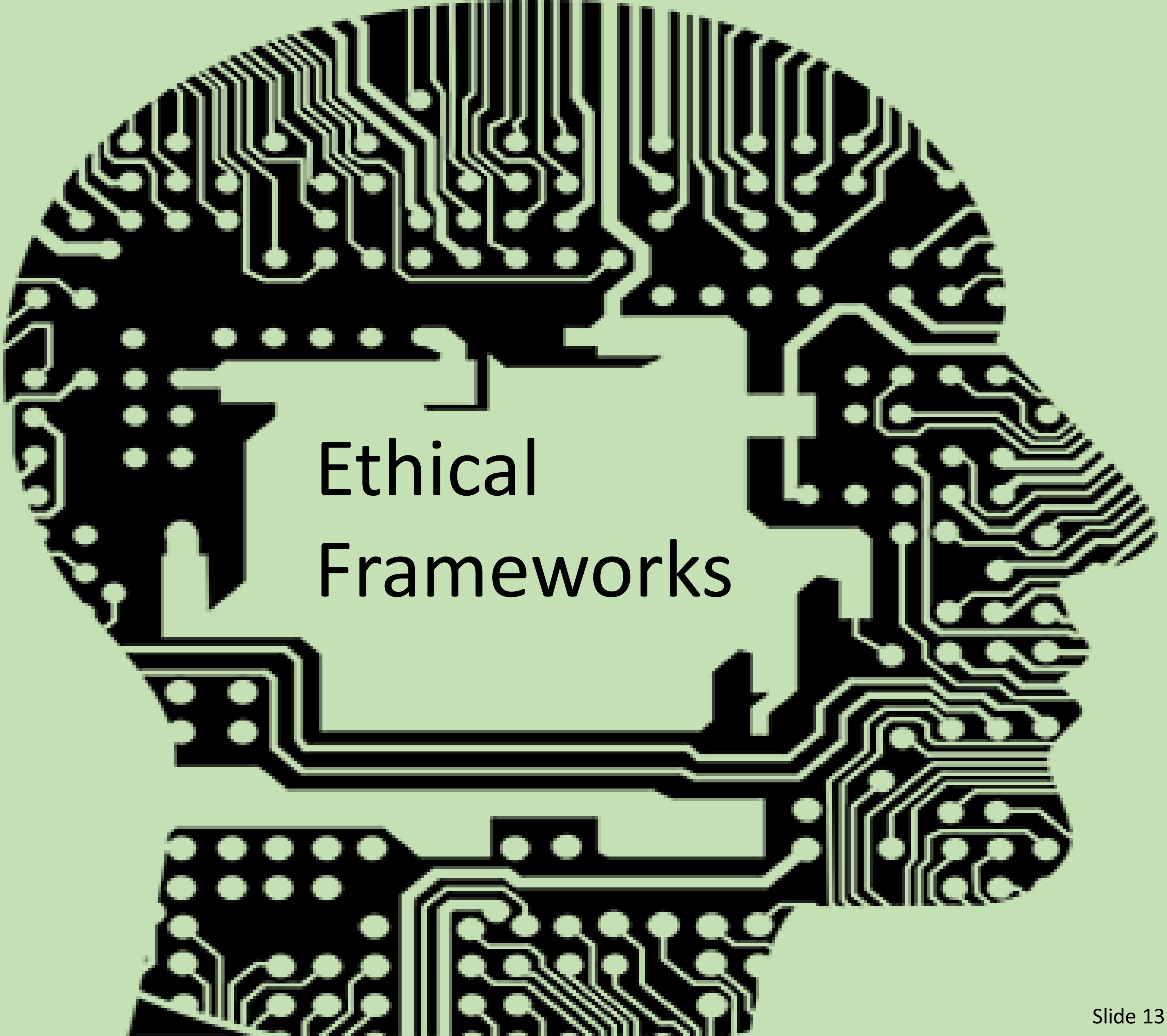
http://medicalfuturist.com/will-robots-take-over-our-jobs-in-healthcare/

Amazon Robotics

U.S. Air Force photo/Lt Col Leslie Pratt - USAF photo via public domain website http://www.af.mil/shared/media/photodb/photos/081131-F-7734Q-001.jpg



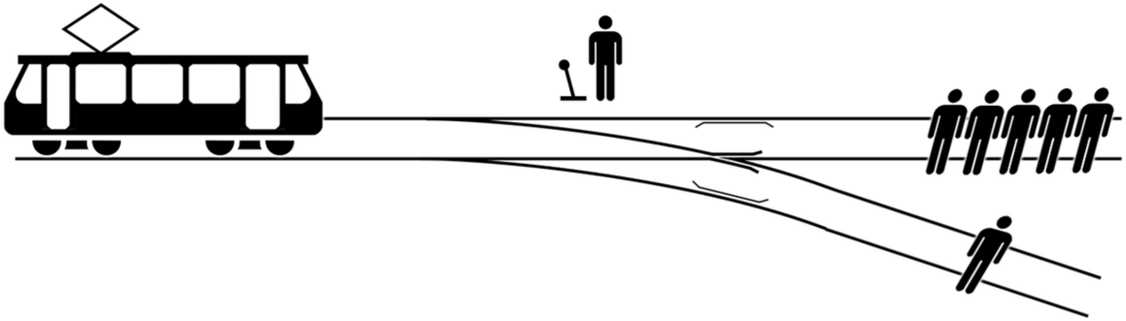
http://moralmachine.mit.edu/



Ethical Frameworks

Ethical Frameworks

1. Normative Frameworks
2. Descriptive Frameworks
3. Applied Frameworks
4. Hybrid Frameworks



This Photo by Unknown author is licensed under CC BY-SA.

Ethical Frameworks – Normative Ethics

Normative ethics are also called **Prescriptive Ethics.**

- Investigate the set of questions that arise when considering how one ought to act from a moral perspective

Deontology (from Greek δέον, *deon*, "obligation, duty")

- Action is more important than the consequences.
- Holds that the morality of an action should be based on whether that action itself is right or wrong under a series of rules, rather than based on the consequences of the action. It is sometimes described as duty-, obligation- or rule-based ethics.

Consequentialism

- Consequences are more important than the actions.
- Holds that the consequences of one's conduct are the ultimate basis for any judgment about the rightness or wrongness of that conduct. Thus, from a consequentialist standpoint, a morally right act (or omission from acting) is one that will produce a good outcome, or consequence.

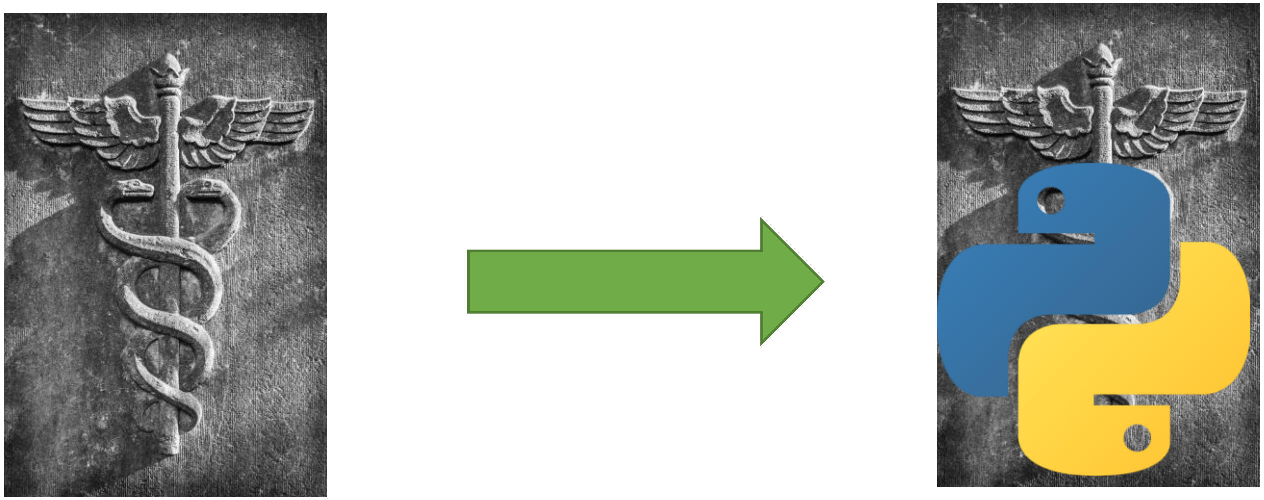
Ethical Frameworks – Descriptive Ethics

- Also known as **comparative ethics**, is the study of people's beliefs about morality.
- What do people think is right?

Ethical Frameworks – Applied Ethics

- Considers the practical application of moral reasoning.
- Expanded beyond the classical philosophical debate.
- Requires specialist understanding of the potential ethical issues in fields like medicine, business or information technology.

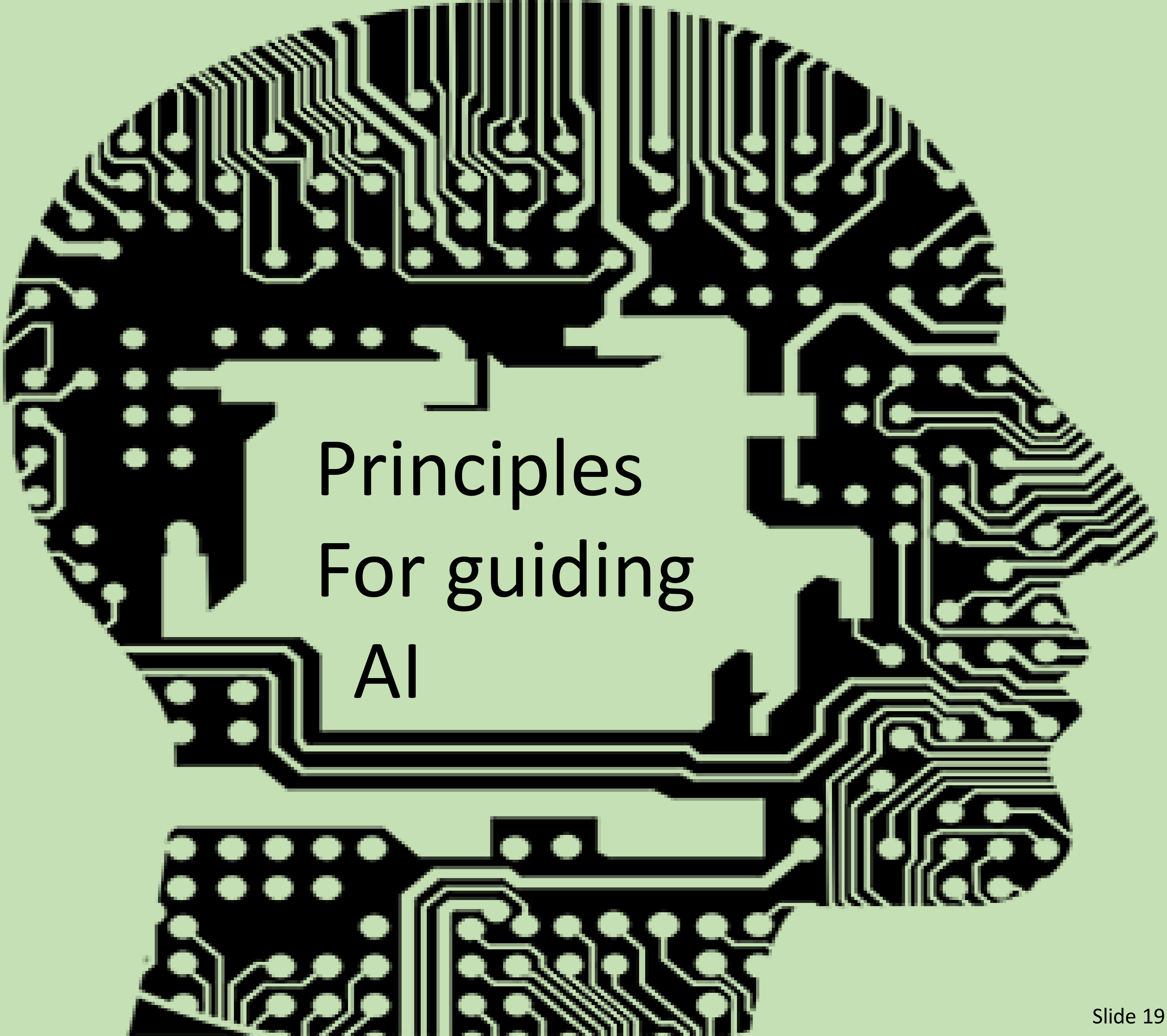
ETHICAL CODE



Ethical Frameworks – Hybrid Frameworks

- **Human Rights**
- Commonly considered a "deontological" concept, can only be justified with reference to the consequences of having those rights.
- See *Contractualism* ~T.M. Scanlon

This Photo by Unknown author is licensed under CC BY-SA.



Principles For guiding AI

Principles for Guiding AI

- Sustainable Development Goals
- Asilomar Principles

Sustainable Development Goals (SDGs)

United Nations General Assembly (UNGA) : overarching concerns over AI given the 17 SDGs.

- SDGs were set in 2015 by the UNGA.
- A collection of 17 global goals designed to be a "blueprint to achieve a better and more sustainable future for all.".
- Intended to be achieved by the year 2030.
- According to the United Nations Activities on Artificial Intelligence 2018 Report, AI has a major impact on most of the 17 SDGs.
- The most impacted SGD is Goal no. 16: Peace, Justice, and Strong Institutions

Sustainable Development Goal no. 16

Aim to “*Promote peaceful and inclusive societies for sustainable development, provide access to justice for all and build effective, accountable and inclusive institutions at all levels*”.

- More details at https://sustainabledevelopment.un.org/sdg16

Asilomar AI principles

Developed by participants at the Beneficial AI 2017 conference, organised by the Future of Life Institute.

Participants (>100) were AI researchers from academia and industry, and thought leaders in economics, law, ethics, and philosophy.

Principles included if >90% of attendees agreed on them.

This Photo by Unknown author is licensed under CC BY-ND.

Asilomar AI principles

Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

2016: Tesla driver killed in collision with tractor trailer while the vehicle was in “autopilot”.

“As the truck turned left, crossing the Tesla's path, neither the human nor the machine could distinguish the white-colored body of the truck from the sky, Tesla said. As a result, the Model S never slowed down….“neither autopilot nor the driver noticed the white side of the tractor trailer against a brightly lit sky,”…. “high ride height of the trailer,”….might have played a role in preventing the radar from reporting correctly.” Source: Washington Post

Asilomar AI principles

Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

What do we mean by “safe”?

If use of technology means less people are harmed, is that good enough?

Does any potential for harm mean technology is unsafe?

Asilomar AI principles

Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

What do we mean by “verifiably”?

- Mathematically proven behaviour. How? Unlikely to always be possible.
- Take some agreed measures in order to show that a good outcome is most likely. What measures? How should we weigh the level of risk against the potential harm?

Asilomar AI principles

Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

"Never mind killer robots – even the good ones are scarily unpredictable" Source: The Conversation.

Study of bots used "to automatically edit Wikipedia articles….designed and exploited by Wikipedia's trusted human editors….underlying software is open-source….all have a common goal of improving the encyclopaedia….found pairs of bots that have been undoing each other's edits for several years without anyone noticing".

Asilomar AI principles

Safety: AI systems should be safe and secure throughout their operational lifetime, and verifiably so where applicable and feasible.

"Never mind killer robots – even the good ones are scarily unpredictable" Source: The Conversation.

What happens when two bots designed to talk to humans talk to one another? "they can quickly start acting in surprising ways, arguing and even insulting each other"
https://www.youtube.com/watch?v=WnzlbyTZsQY

Asilomar AI principles

Failure Transparency: If an AI system causes harm, it should be possible to ascertain why. **Judicial Transparency:** Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

What do we mean by, and how do we identify, harm caused?

Consider for example fake news. Say global warming is presented as false to a large number of people who believe this message, serious harm could be caused to the planet. Can we establish a causal link between the two?

Asilomar AI principles

Failure Transparency: If an AI system causes harm, it should be possible to ascertain why. **Judicial Transparency**: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

- "For **users**, transparency is important because it builds trust in the system, by providing a simple way for the user to understand what the system is doing and why.
- For **safety certification** of an AS, transparency is important because it exposes the system's processes for independent certification against safety standards.
- If accidents occur, AS will need to be transparent to an **accident investigator**; the internal process that led to the accident need to be traceable.
- Following an accident **lawyers or other expert witnesses**, who may be required to give evidence, require transparency to inform their evidence. And
- for disruptive technologies, such as driverless cars, a certain level of transparency to **wider society** is needed in order to build public confidence in the technology."

[Source: Professor Alan Winfield's blog http://alanwinfield.blogspot.co.uk/2017/01/the-infrastructure-of-life-2.html]

Asilomar AI principles

Failure Transparency: If an AI system causes harm, it should be possible to ascertain why. **Judicial Transparency**: Any involvement by an autonomous system in judicial decision-making should provide a satisfactory explanation auditable by a competent human authority.

Challenge of "black box" algorithms.

- Commercially sensitive.
- Too complex for us to understand

"Black Box"
Input, x
Output, F(x)

This Photo by Unknown author is licensed under CC BY.

Asilomar AI principles

Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

Responsible Research and Innovation.

EPSRC: "A Responsible Innovation approach should be one that continuously seeks to:

- **Anticipate** – describing and analysing the impacts, intended or otherwise, (for example economic, social, environmental) that might arise. This does not seek to predict but rather to support an exploration of possible impacts and implications that may otherwise remain uncovered and little discussed.
- **Reflect** – reflecting on the purposes of, motivations for and potential implications of the research, and the associated uncertainties, areas of ignorance, assumptions, framings, questions, dilemmas and social transformations these may bring.
- **Engage** – opening up such visions, impacts and questioning to broader deliberation, dialogue, engagement and debate in an inclusive way.
- **Act** – using these processes to influence the direction and trajectory of the research and innovation process itself."

Asilomar AI principles

Responsibility: Designers and builders of advanced AI systems are stakeholders in the moral implications of their use, misuse, and actions, with a responsibility and opportunity to shape those implications.

Responsible Research and Innovation.

EU: “Responsible Research and Innovation is:

- **Involving society in science and innovation** ‘very upstream' in the processes of R&I to align its outcomes with the values of society.
- **A wide umbrella connecting different aspects of the relationship between R&I and society**: public engagement, open access, gender equality, science education, ethics, and governance.”

Asilomar AI principles

Value Alignment: Highly autonomous AI systems should be designed so that their goals and behaviours can be assured to align with human values throughout their operation. **Human Values**: AI systems should be designed and operated so as to be compatible with ideals of human dignity, rights, freedoms, and cultural diversity.

Asilomar AI principles

Value-sensitive design: an interdisciplinary methodological framework for the design of technologies that account for human values in a principled and comprehensive manner throughout the design process. Consists of iteratively applied conceptual, empirical and technical investigations.

Conceptual investigations, theoretical and literature-based, identify the direct and indirect stakeholders, the values at stake, and potential trade-off processes between competing values.

These are further informed by **empirical investigations**, with stakeholder groups.

Technological investigations focus either on how existing technologies support or hinder ethical values, or on proactive design of technology that supports the values identified in the conceptual and empirical investigations.

Asilomar AI principles

Shared Benefit: AI technologies should benefit and empower as many people as possible. **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

In an interview with the MIT Tech Review, economist Erik Brynjolfsson said, “technology is the main driver of the recent increases in inequality.”

Asilomar AI principles

Shared Benefit: AI technologies should benefit and empower as many people as possible. **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

Concerns about **biased AI**.

AI trained to identify relationships between words "considered the word "programmer" closer to the word "man" than "woman," and that the most similar word for "woman" is "homemaker."" Source: MIT Technology Review.

Asilomar AI principles

Shared Benefit: AI technologies should benefit and empower as many people as possible. **Shared Prosperity:** The economic prosperity created by AI should be shared broadly, to benefit all of humanity.

Concerns about **biased AI**.

"… darker-skinned females are the most misclassified group (with error rates of up to 34.7%). The maximum error rate for lighter-skinned males is 0.8%." Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification, Joy Buolamwini, Timnit Gebru, 2018.

"One widely used facial-recognition data set was estimated to be more than 75 percent male and more than 80 percent white, according to another research study." Source: New York Times.

Asilomar AI principles

Human Control: Humans should choose how and whether to delegate decisions to AI systems, to accomplish human-chosen objectives.

- How do we trust an AI if we can't understand why it has made the decisions it has?
- Should humans always be in control of a machine's decisions?
- When is it appropriate for a machine to take over?

Asilomar AI principles

AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

AUTONOMOUS WEAPONS: AN OPEN LETTER FROM AI & ROBOTICS RESEARCHERS

“ … we believe that AI has great potential to benefit humanity in many ways, and that the goal of the field should be to do so. Starting a military AI arms race is a bad idea, and should be prevented by a ban on offensive autonomous weapons beyond meaningful human control.”

signed by 3722 AI/Robotics researchers and 20467 others

Asilomar AI principles

AI Arms Race: An arms race in lethal autonomous weapons should be avoided.

Response from Evan Ackerman (IEEE Spectrum Journalist): "…no letter, UN declaration, or even a formal ban ratified by multiple nations is going to prevent people from being able to build autonomous, weaponized robots….

Research in AI & Ethics

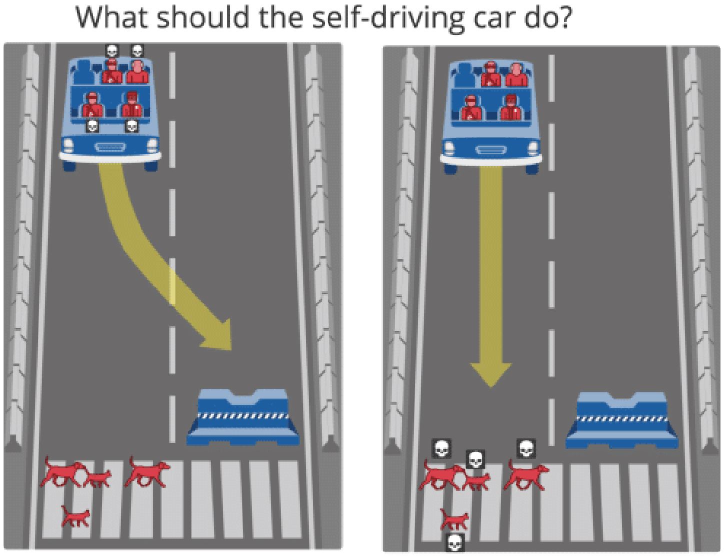
Some Relevant Research in AI

1. Moral Values
2. AI & Future of work
3. Machine Behaviour
4. Bias & Discrimination in AI
5. Legal& Ethical reasoning
6. Explainable & Trustworthy AI
7. Deceptive AI

The Moral Machine experiment

- Awad E, Dsouza S, Kim R, Schulz J, Henrich J, Shariff A, Bonnefon JF, Rahwan I. The moral machine experiment. Nature. 2018 Nov;563(7729):59.

An online experimental platform designed to explore the moral dilemmas faced by autonomous vehicles. This platform gathered 40 million decisions in 10 languages from millions of people in 233 countries and territories.



AI & Future of work

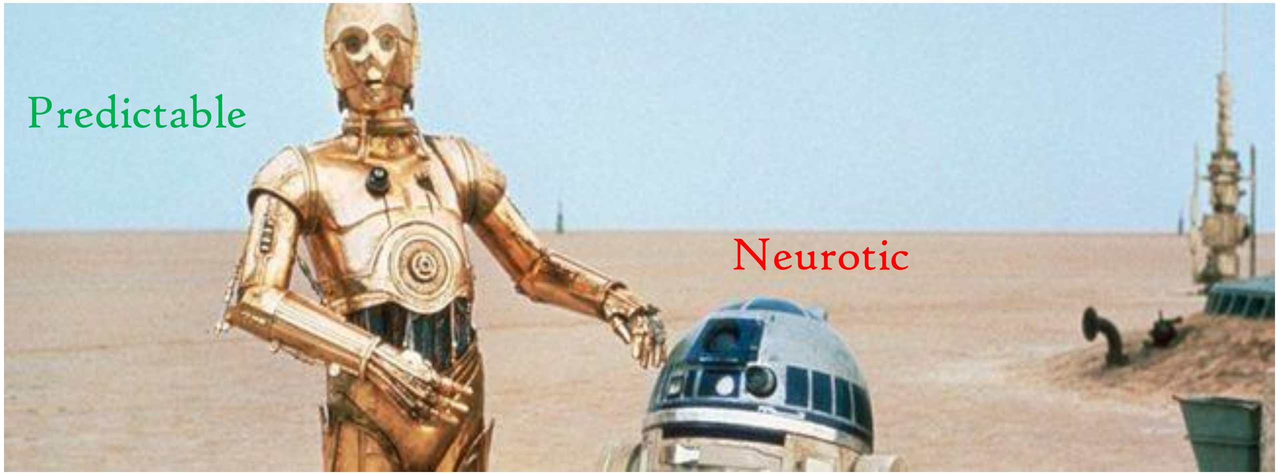
- Frank MR, Autor D, Bessen JE, Brynjolfsson E, Cebrian M, Deming DJ, Feldman M, Groh M, Lobo J, Moro E, Wang D. Toward understanding the impact of artificial intelligence on labor. Proceedings of the National Academy of Sciences. 2019 Apr 2;116(14):6531-9.

Paper that aims to "... discuss the barriers that inhibit scientists from measuring the effects of AI and automation on the future of work. These barriers include the lack of high-quality data about the nature of work (e.g., the dynamic requirements of occupations), lack of empirically informed models of key microlevel processes (e.g., skill substitution and human–machine complementarity), and insufficient understanding of how cognitive technologies interact with broader economic dynamics and institutional mechanisms (e.g., urban migration and international trade policy)..."

Machine Behaviour

- Rahwan, I., Cebrian, M., Obradovich, N. *et al.* Machine behaviour. *Nature* **568,** 477–486 (2019)

Furthering the study of machine behaviour is critical to maximizing the potential benefits of AI for society. The consequential choices that we make regarding the integration of AI agents into human lives must be made with some understanding of the eventual societal implications of these choices. To provide this understanding and anticipation, we need a new interdisciplinary field of scientific study: machine behaviour.



Bias & Discrimination in AI

- DADD Project, King's College London. See https://dadd-project.org/about/

- Aran, X. F., Such, J. M., & Criado, N. (2019). Attesting Biases and Discrimination using Language Semantics. *arXiv preprint arXiv:1909.04386*.

- Bareinboim, E., Tian, J., & Pearl, J. (2014, June). Recovering from selection bias in causal and statistical inference. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

- Leavy, S. (2018, May). Gender bias in artificial intelligence: The need for diversity and gender theory in machine learning. In *Proceedings of the 1st International Workshop on Gender Equality in Software Engineering* (pp. 14-16). ACM.

Legal and Ethical reasoning with Argumentation

AI systems need to be capable of **legal and ethical reasoning**. Argumentation is a candidate for this.

1. Allows reasoning about what to do, that takes into account values and norms.

2. Allows humans and machines to engage in effective joint reasoning and decision-making while ensuring that humans can understand, challenge, influence and benefit from machine reasoning.

- Pagallo, Ugo, Giovanni Sartor, and Gianmaria Ajani. *AI Approaches to the Complexity of Legal Systems*. Springer Berlin Heidelberg, 2010.

Explainable & Trustworthy AI

Understanding Trust in AI

- Castelfranchi, C., & Falcone, R. (2010). *Trust theory: A socio-cognitive and computational model* (Vol. 18). John Wiley & Sons.
- Taddeo, M. (2010). Modelling trust in artificial agents, a first step toward the analysis of e-trust. *Minds and machines*, *20*(2), 243-257.

Understanding Explanations in AI

- Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1-38.

AI reasons about Trust in an explainable way

- Parsons, S., Atkinson, K., Haigh, K. Z., Levitt, K. N., McBurney, P., Rowe, J., Singh, M.P. & Sklar, E. (2012). Argument Schemes for Reasoning about Trust. *COMMA*, *245*, 430.

Deceptive AI

Deception reduces human bias against machines

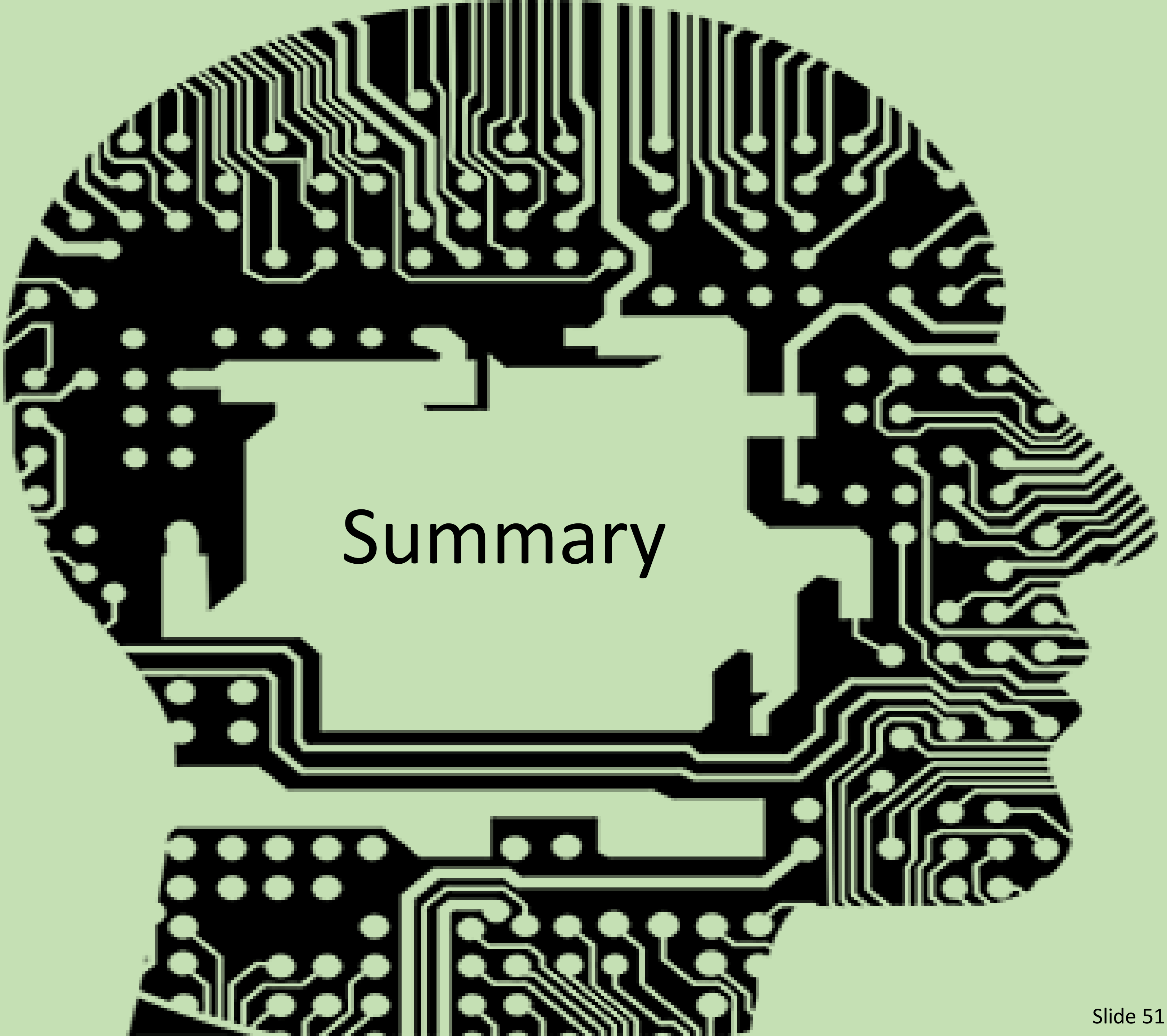
- Ishowo-Oloko, F., Bonnefon, J. F., Soroye, Z., Crandall, J., Rahwan, I., & Rahwan, T. (2019). Behavioural evidence for a transparency–efficiency tradeoff in human–machine cooperation. *Nature Machine Intelligence*, 1-5.

Types of deception in human-robot interaction

- Shim, J., & Arkin, R. C. (2013, October). A taxonomy of robot deception and its benefits in HRI. In *2013 IEEE International Conference on Systems, Man, and Cybernetics* (pp. 2328-2335). IEEE.

Complex reasoning for deception in multi-agent systems

- Sarkadi, Ş., Panisson, A. R., Bordini, R. H., McBurney, P., Parsons, S., & Chapman, M. (2019). Modelling deception using theory of mind in multi-agent systems. *AI Communications*, 32(4), 287-302.



Summary

Summary

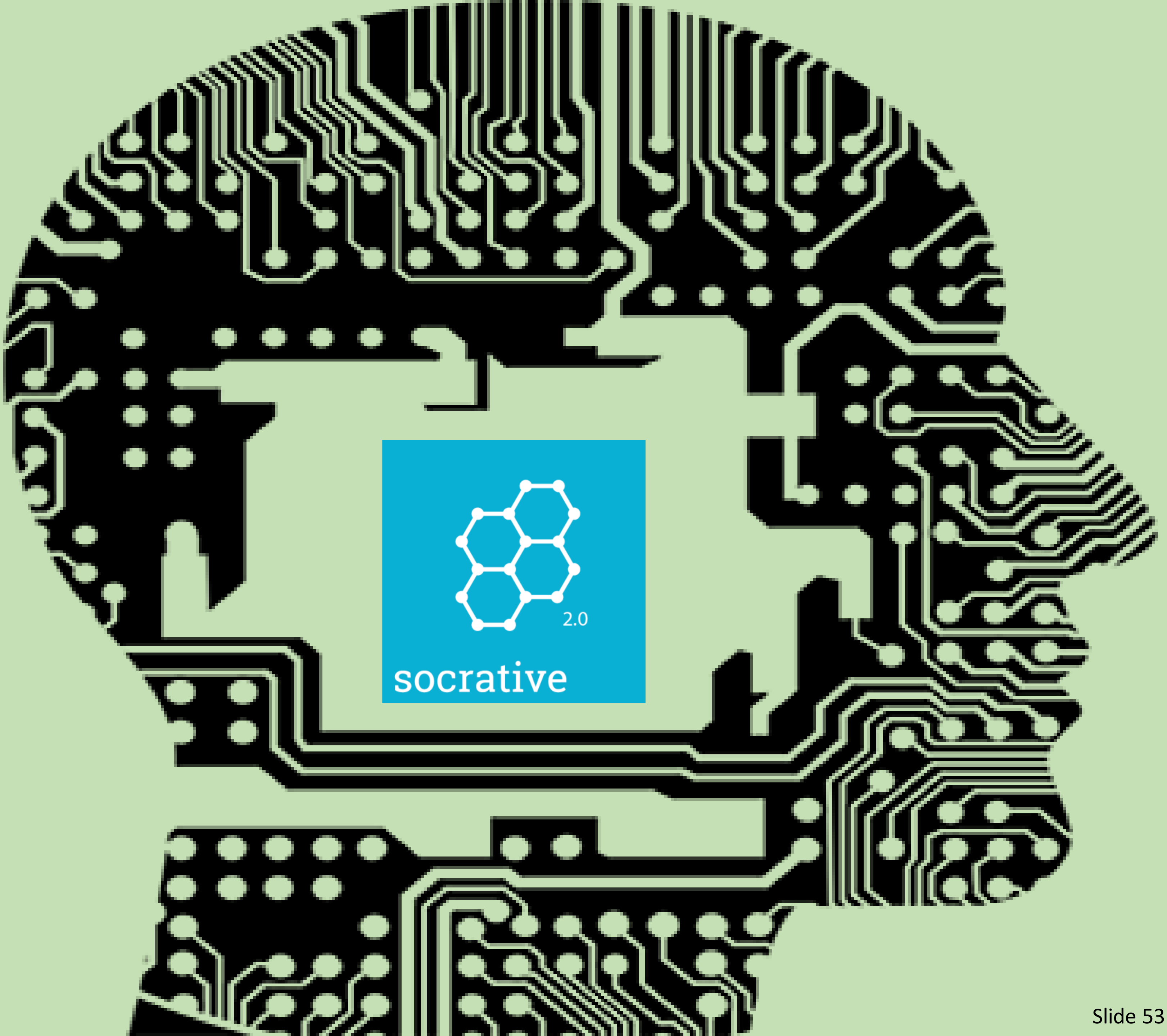
As we see more autonomous technology, with access to more data and able to make decisions that impact on human lives and on society, it’s important to consider the ethics of that technology.

Many issues, recognised globally as important by governments, organisations and experts.

"As yet, no consensus or clear solutions." <- last year summary

Today: There is some consensus, but there's still much to be done.

Research is being done so that we have better AI systems and a better understanding of these AI systems and how they impact us and themselves.


2.0
socrative

DISCUSSION: How worried should we be?

- Room number: 6CCS3AIN

socrative

by MasteryConnect