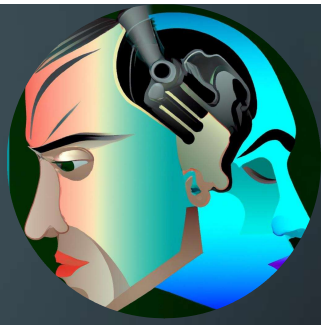


[AI-PHI] INAUGURAL SESSION
INTRODUCTION TO THE PHILOSOPHY OF AI

Aïda Elamrani
email: aidaelamrani@outlook.fr
twitter: @AidaElam

OUTLINE

1. Glossary
2. Digital Transformation
3. AI Ethics
4. Artificial Minds

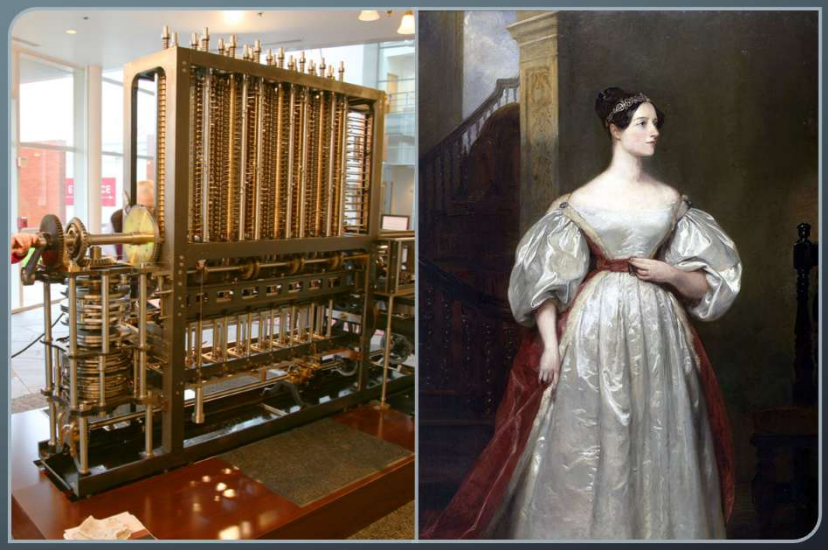
BASIC AI-PHI GLOSSARY

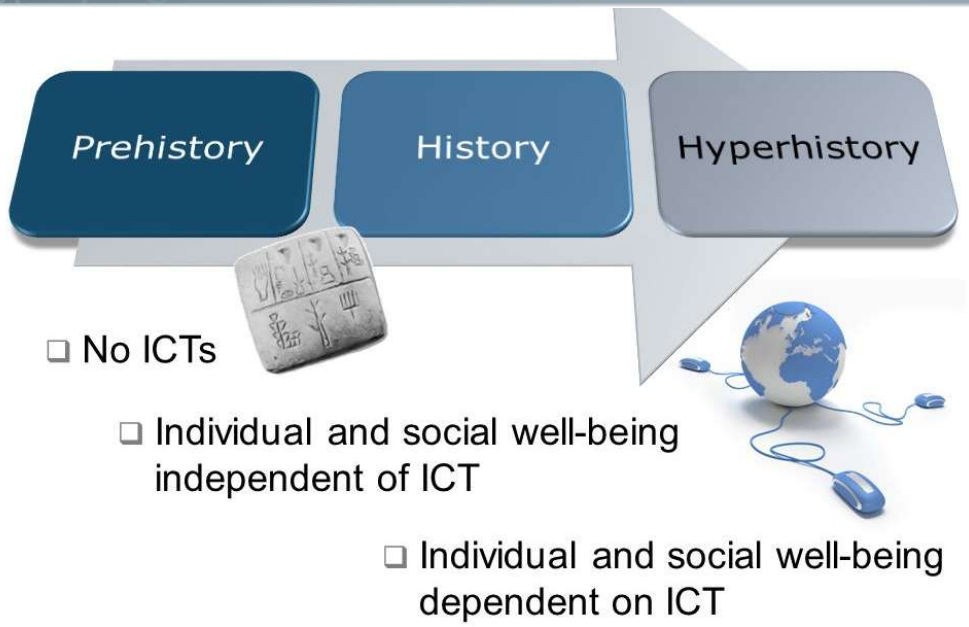
- Artificial Intelligence:
  - Technology imitating human intellectual abilities
  - Discipline studying and designing agents capable of performing intellectual tasks
- Philosophy: The queen of all sciences.
- Epistemology:
  - The branch of philosophy concerned with knowledge (*episteme* as opposed to *doxa*, opinion)
  - The study of the history of science
- Metaphysics: The branch of philosophy concerned with the study of reality.
- Ontology: The branch of philosophy concerned with the study of *what there is*.
- Morality: a set of personal or social standards for good or bad behaviour and character [Cambridge dictionary]
- Ethics / Moral philosophy: the study of morality

ADA BYRON COUNTESS OF LOVELACE

Its province is to assist us in making available what we are already acquainted with. This it is calculated to effect primarily and chiefly of course, through its executive faculties; but **it is likely to exert an indirect and reciprocal influence on science itself in another manner**. For, in so distributing and combining the truths and the formulæ of analysis, that they may become most easily and rapidly amenable to the mechanical combinations of the engine, **the relations and the nature of many subjects in that science are necessarily thrown into new lights, and more profoundly investigated.** This is a decidedly indirect, and a somewhat speculative, consequence of such an invention. It is however pretty evident, on general principles, that in devising for mathematical truths a new form in which to record and throw themselves out for actual use, views are likely to be induced, which should again react on the more theoretical phase of the subject**. There are in all extensions of human power, or additions to human knowledge, various collateral influences, besides the main and primary object attained.** (…)

Lovelace, A. (1843). 'Notes on L. Menabrea's 'Sketch of the Analytical Engine Invented by Charles Babbage, Esq.'''. Taylor's Scientific Memoirs, 3(1843), 1843.





WELCOME TO HYPERHISTORY!

THE 4TH REVOLUTION – NEW AGENTS IN THE INFOSPHERE

1. 1543 Nicolaus Copernicus
2. 1859 Charles Darwin
3. 1856-1939 Sigmund Freud
4. 1950 Alan Turing

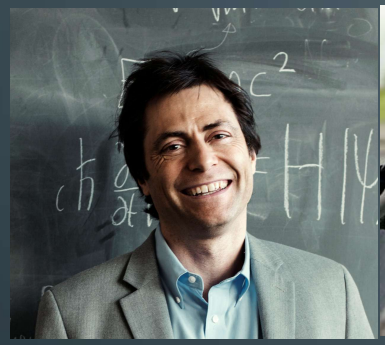
"We are slowly accepting the post-Turing idea that we are not Newtonian, stand-alone, and unique agents, some Robinson Crusoe on an island. Rather, we are informational organisms (inforgs), mutually connected and embedded in an informational environment (the infosphere), which we share with other informational agents, both natural and ***artificial, that also process information logically and autonomously****."*

L. Floridi, The fourth revolution: How the infosphere is reshaping human reality. OUP Oxford, 2014.

« EXPERTS » DISAGREE

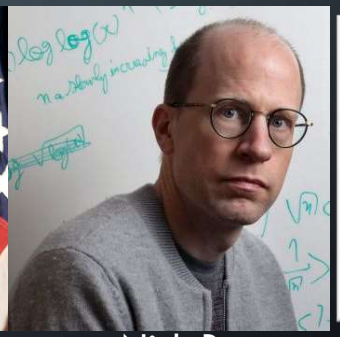
Techno-pessimism

Max Tegmark



Stephen Hawking

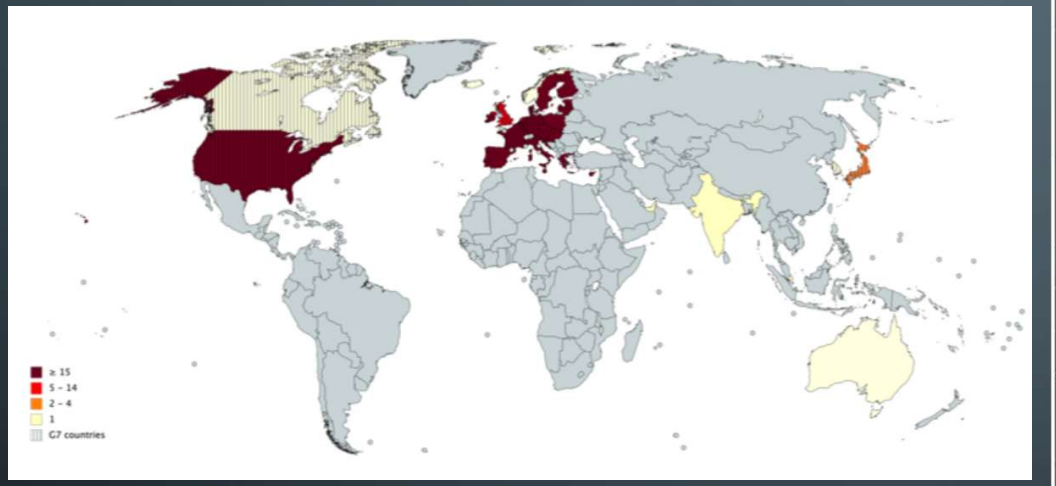
Ray Kurzweil



Nick Bostrom

Techno-optimism

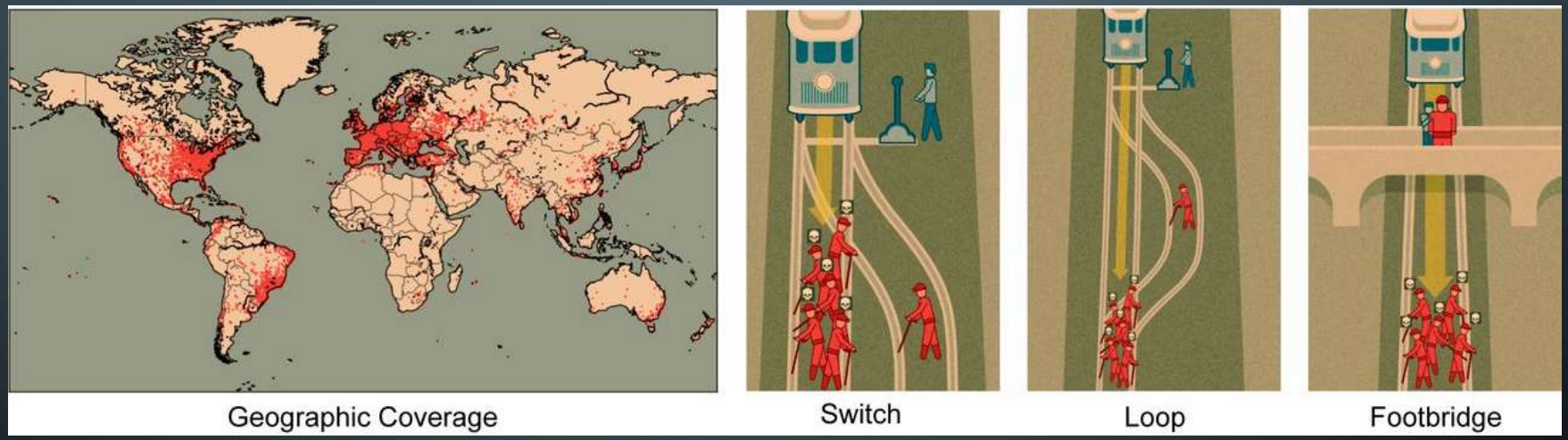
GUIDELINES ON ETHICS OF AI (2016-2019)



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

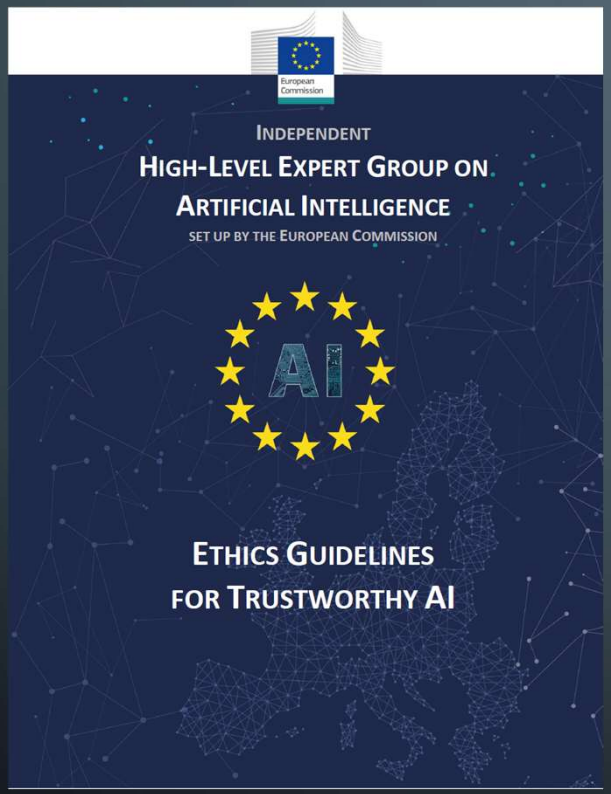
Jobin, A., Ienca, M., & Vayena, E. (2019). Artificial Intelligence: The global landscape of ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399.

ETHICS IS **HARD**



Awad et al. "The Moral Machine Experiment." *Nature* 563, no. 7729 (November 2018): 59–64. https://doi.org/10.1038/s41586-018-0637-6.

THE EUROPEAN WAY



Four fundamental guiding principles:

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. Explicability

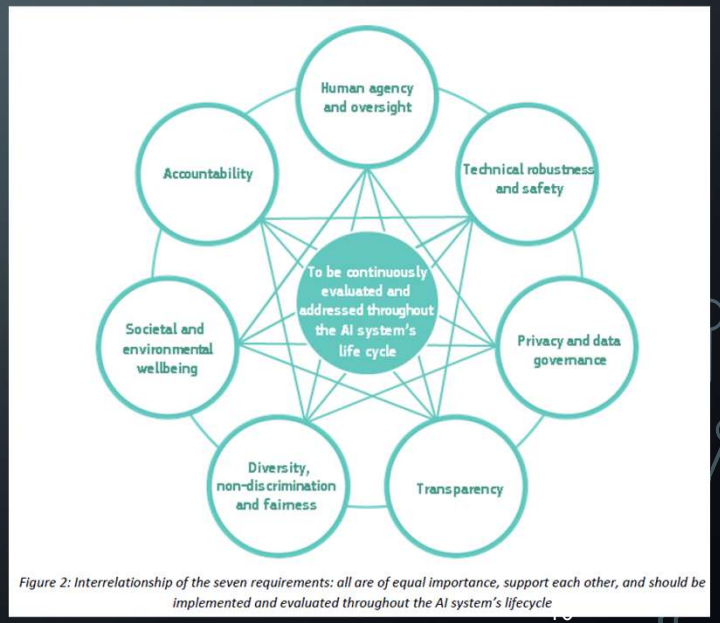
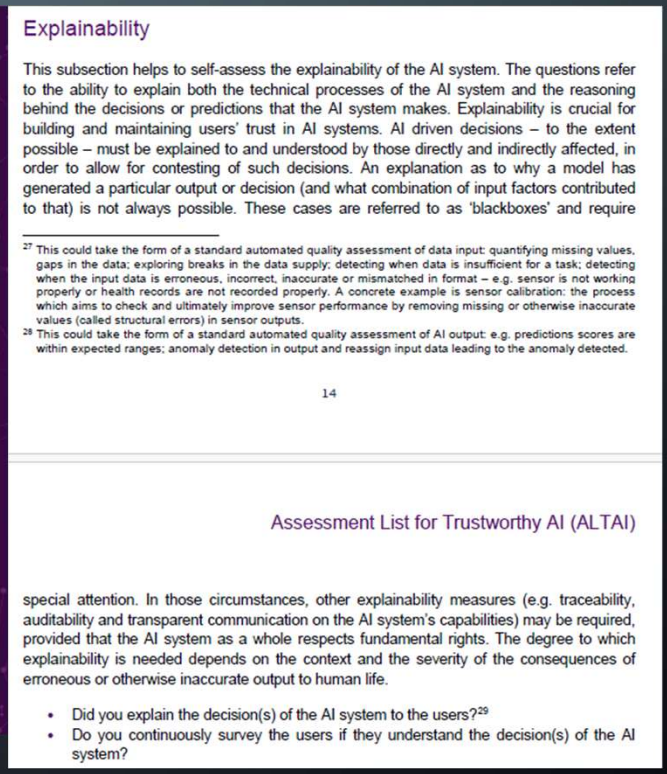


Figure 2: Interrelationship of the seven requirements: all are of equal importance, support each other, and should be implemented and evaluated throughout the AI system's lifecycle

FROM PRINCIPLES TO ACTION?

- ALTAI
- Ethics by Design
- Auditability
- Legal frameworks (GDPR, AI Act,…)



Explainability

This subsection helps to self-assess the explainability of the AI system. The questions refer to the ability to explain both the technical processes of the AI system and the reasoning behind the decisions or predictions that the AI system makes. Explainability is crucial for building and maintaining users' trust in AI systems. AI driven decisions – to the extent possible – must be explained to and understood by those directly and indirectly affected, in order to allow for contesting of such decisions. An explanation as to why a model has generated a particular output or decision (and what combination of input factors contributed to that) is not always possible. These cases are referred to as 'blackboxes' and require

[27] This could take the form of a standard automated quality assessment of data input: quantifying missing values, gaps in the data; exploring breaks in the data supply; detecting when data is insufficient for a task; detecting when the input data is erroneous, incorrect, inaccurate or mismatched in format – e.g. sensor is not working properly or health records are not recorded properly. A concrete example is sensor calibration: the process which aims to check and ultimately improve sensor performance by removing missing or otherwise inaccurate values (called structural errors) in sensor outputs.

[28] This could take the form of a standard automated quality assessment of AI output: e.g. predictions scores are within expected ranges; anomaly detection in output and reassign input data leading to the anomaly detected.

14

Assessment List for Trustworthy AI (ALTAI)

special attention. In those circumstances, other explainability measures (e.g. traceability, auditability and transparent communication on the AI system's capabilities) may be required, provided that the AI system as a whole respects fundamental rights. The degree to which explainability is needed depends on the context and the severity of the consequences of erroneous or otherwise inaccurate output to human life.

- Did you explain the decision(s) of the AI system to the users?[29]
- Do you continuously survey the users if they understand the decision(s) of the AI system?

NEW AGENTS IN THE INFOSPHERE

- **Ethical-impact agents:** not their purpose, but the impact they have nonetheless as a technology
- **Implicit ethical agents:** designed to satisfy ethical requirements (such as safety)
- **Explicit ethical agents:** designed to solve ethical problems (ex: medical resource allocation)
- **Full ethical agents:** "*can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent. We typically regard humans as having consciousness, intentionality, and free will. Can a machine be a full ethical agent?*"

Machine Ethics

The Nature, Importance, and Difficulty of Machine Ethics

James H. Moor, *Dartmouth College*

Implementations of machine ethics might be possible in situations ranging from maintaining hospital records to overseeing disaster relief. But what is machine ethics, and how good can it be?

The question of whether machine ethics exists or might exist in the future is difficult to answer if we can't agree on what counts as machine ethics. Some might argue that machine ethics obviously exists because humans are machines and humans have ethics. Others could argue that machine ethics obviously doesn't exist because ethics is simply emotional expression and machines can't have emotions.

A wide range of positions on machine ethics are possible, and a discussion of the issue could rapidly propel us into deep and unsettled philosophical issues. Perhaps, understandably, few in the scientific arena pursue the issue of machine ethics. You're unlikely to find easily testable hypotheses in the murky waters of philosophy. But we can't—and shouldn't—avoid consideration of machine ethics in today's technological world.

As we expand computers' decision-making roles in practical matters, such as computers driving cars, ethical considerations are inevitable. Computer scientists and engineers must examine the possibilities for machine ethics because, knowingly or not, they've already engaged—or will soon engage—in some form of it. Before we can discuss possible implementations of machine ethics, however, we need to be clear about what we're asserting or denying.

Varieties of machine ethics

When people speak of technology and values, they're often thinking of ethical values. But not all values are ethical. For example, practical, economic, and aesthetic values don't necessarily draw on ethical considerations. A product of technology, such as a new sailboat, might be practically durable, economically expensive, and aesthetically pleasing, absent consideration of any ethical values. We routinely evaluate technology from these nonethical normative viewpoints. Tool makers and users regularly evaluate how well tools accomplish the purposes for which they were designed. With technology, all of us—ethicists and engineers included—are involved in evaluation processes requiring the selection and application of standards. In none of our professional activities can we retreat to a world of pure facts, devoid of subjective normative assessment.

By its nature, computing technology is normative. We expect programs, when executed, to proceed toward some objective—for example, to correctly compute our income taxes or keep an airplane on course. Their intended purpose serves as a norm for evaluation—that is, we assess how well the computer program calculates the tax or guides the airplane. Viewing computers as technological agents is reasonable because they do jobs on our behalf. They're normative agents in the limited sense that we can assess their performance in terms of how well they do their assigned jobs.

After we've worked with a technology for a while, the norms become second nature. But even after they've become widely accepted as the way of doing the activity properly, we can have moments of realization and see a need to establish different kinds of norms. For instance, in the early days of computing, using double digits to designate years was the standard and worked well. But, when the year 2000 approached, programmers realized that this norm needed reassessment. Or consider a distinction involving AI. In a November 1999 correspondence between Herbert Simon and Jacques Berleur,[1] Berleur was asking Simon for his reflections on the 1956 Dartmouth Summer Research Project on Artificial Intelligence, which Simon attended. Simon expressed

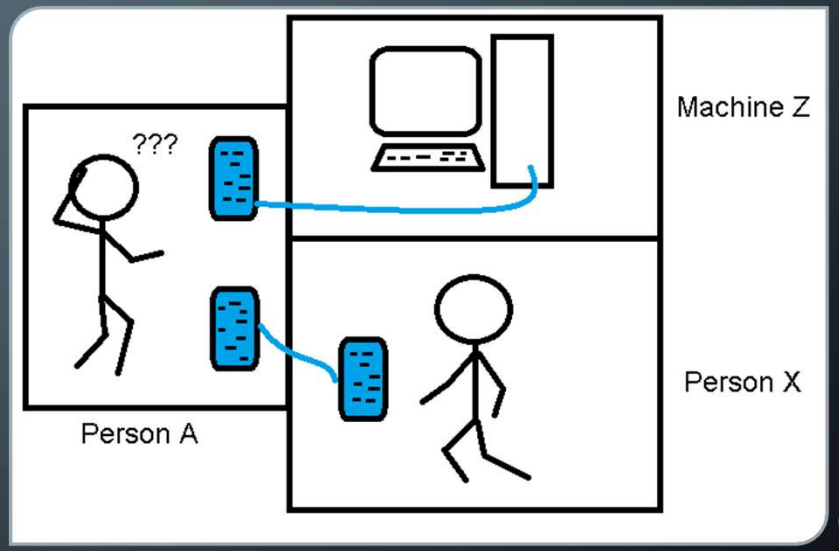
1541-1672/06/$20.00 © 2006 IEEE
Published by the IEEE Computer Society

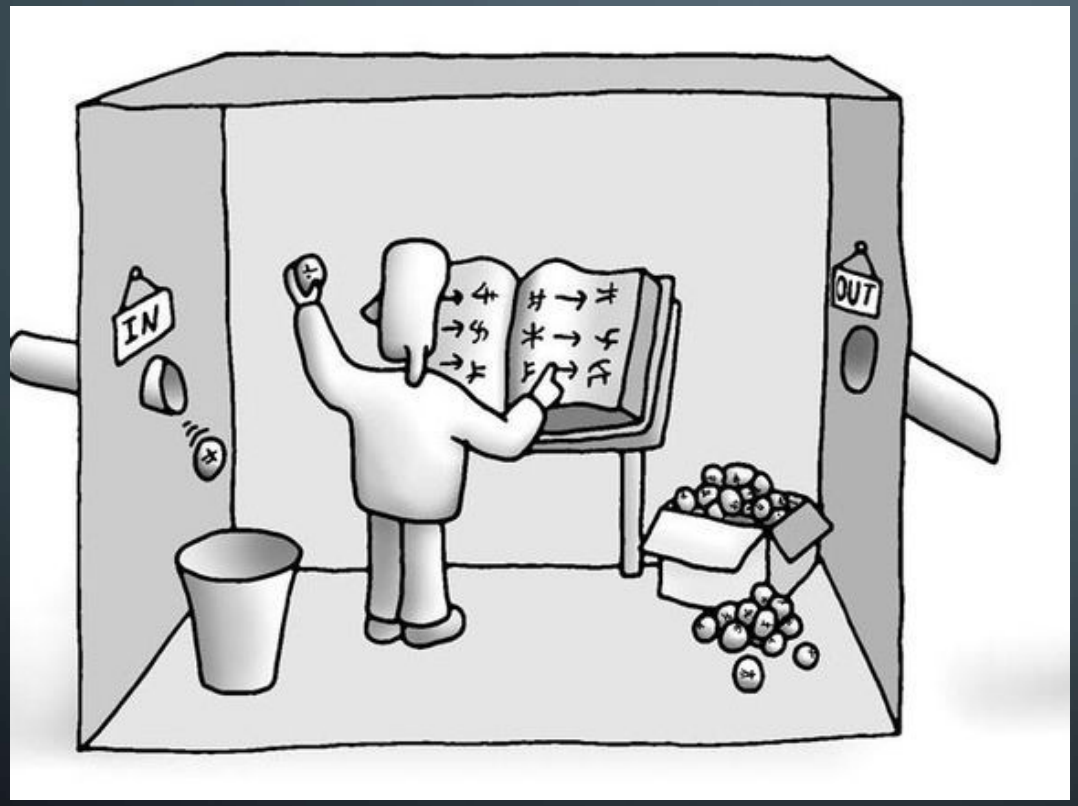
- **Ethical-impact agents:** not their purpose, but the impact they have nonetheless as a technology
- **Implicit ethical agents:** designed to satisfy ethical requirements (such as safety)
- **Explicit ethical agents:** designed to solve ethical problems (ex: medical resource allocation)
- **Full ethical agents:** can make explicit ethical judgments and generally is competent to reasonably justify them. An average adult human is a full ethical agent. We typically regard humans as having consciousness, intentionality, and free will. ***Can a machine be a full ethical agent?***

IMITATION GAME – "TURING TEST"

I do not wish to give the impression that I think there is no mystery about consciousness. There is, for instance, something of a paradox connected with any attempt to localise it. But I do not think these mysteries necessarily need to be solved before we can answer the question with which we are concerned in this paper.

Turing, A. M. "Computing Machinery and Intelligence." *Mind* LIX, no. 236 (October 1, 1950): 433–60.





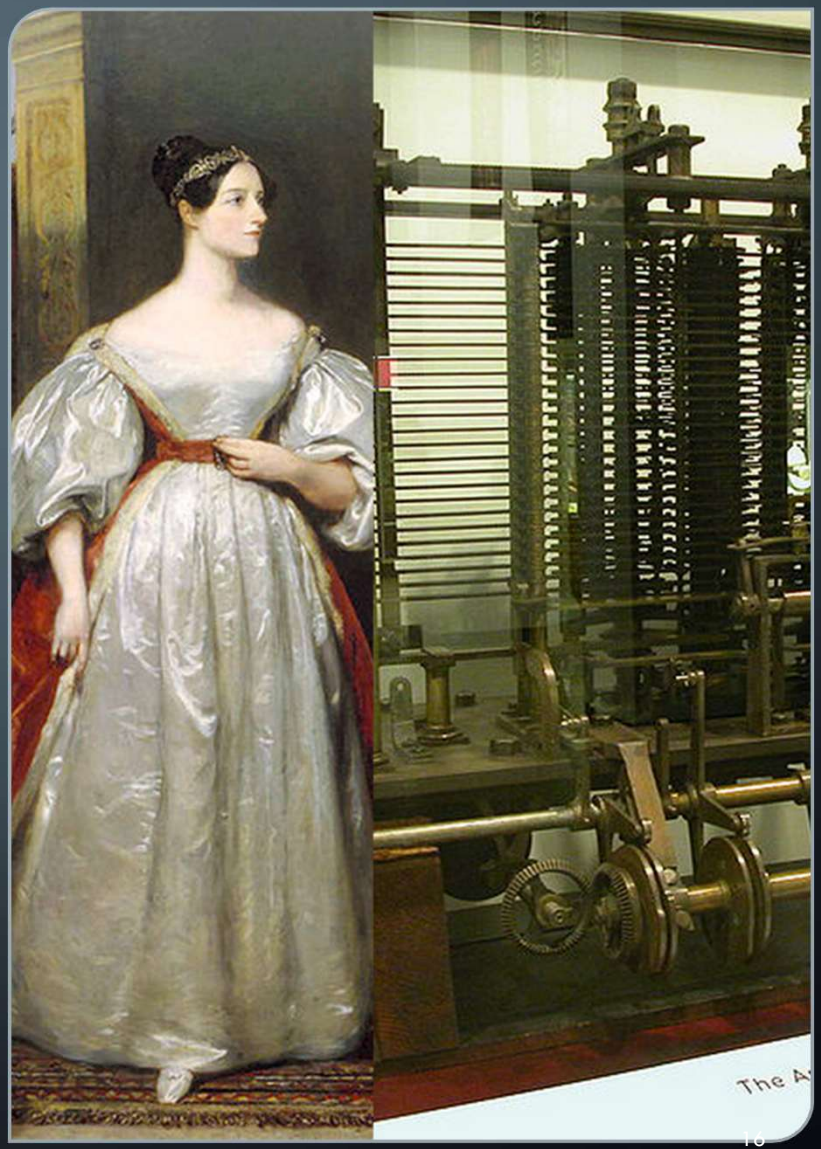
THE CHINESE ROOM

Searle, J. R. (1980). Minds, Brains and Programs. *Behavioral and Brain Sciences*, 3, 417–457.

LOVELACE TEST

Bringsjord, S., Bello, P., & Ferrucci, D. (2003). Creativity, the Turing Test, and the (Better) Lovelace Test. In J. H. Moor (Ed.), *The Turing Test: The Elusive Standard of Artificial Intelligence* (pp. 215–239). Springer Netherlands.

"The Analytical Engine has no pretensions whatever to originate anything. It can do whatever we know how to order it to perform. It can follow analysis; but it has no power of anticipating any analytical relations or truths."



GOING DEEPER DOWN THE HOLE

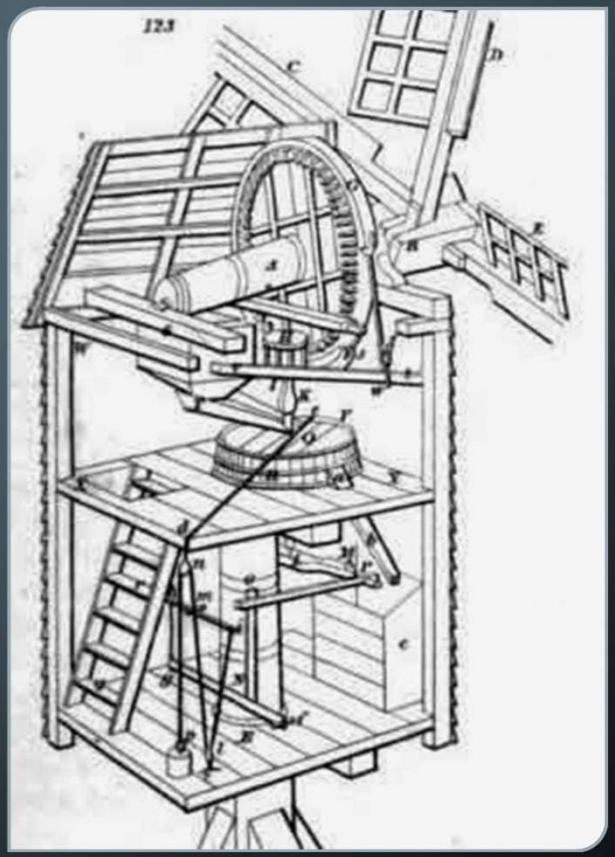
Can a computer be conscious?
Understand? Desire?

Four thought experiments to the rescue:

1. Leibniz's Mill
2. P-Zombies
3. Chinese Nation
4. Mary's Room

LEIBNIZ'S MILL

Leibniz. (1714). Monadologie.



P-ZOMBIES

Kirk, Robert, and Roger Squires. "Zombies v. Materialists." *Proceedings of the Aristotelian Society, Supplementary Volumes 4* (1974): 135–63.

CHINESE NATION

Block, Ned. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9 (1978): 261–325.

MARY'S ROOM

Jackson, Frank. "Epiphenomenal qualia." *The Philosophical Quarterly (1950-)* 32, no. 127 (1982): 127-136.

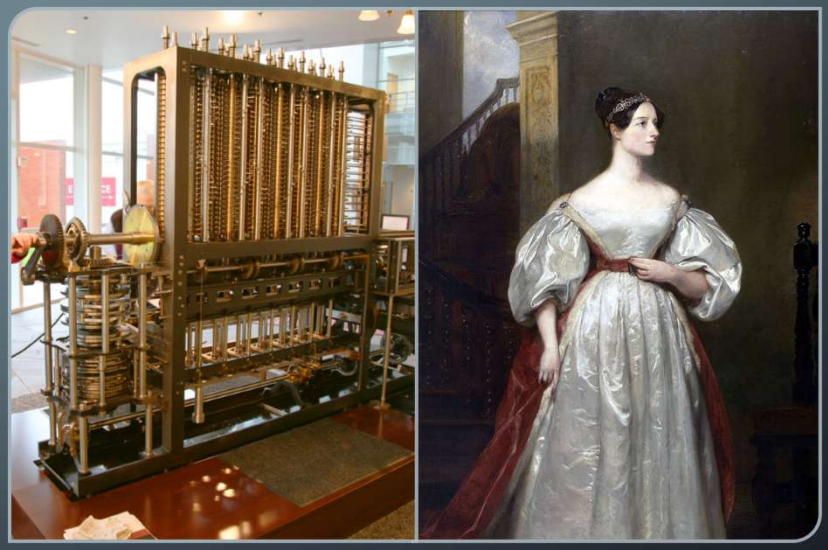
THE EXPLANATORY GAP

- Hard problem of consciousness = qualia, phenomenal consciousness
- "Easy" problem of consciousness = identifying the underlying physical attributes of consciousness

ADA BYRON COUNTESS OF LOVELACE

To return to the executive faculties of this engine: the question must arise in every mind, are they really even able to follow analysis in its whole extent? ***No reply, entirely satisfactory to all minds, can be given to this query, excepting the actual existence of the engine, and actual experience of its practical results.***

Lovelace, A. (1843). 'Notes on L. Menabrea's 'Sketch of the Analytical Engine Invented by Charles Babbage, Esq.'''. Taylor's Scientific Memoirs, 3(1843), 1843.



THANK YOU!