The Ethics of AI

sponsored by
The IEEE Computational Intelligence Society
http://www.ieeevic.org/

Dr. Kevin Korb
Clayton School of Information Technology
Monash University
kbkorb@gmail.com

What would an AI be?

- A *general* intelligence
  - Not domain restricted
  - Not brittle; capable of learning
  - *Autonomous*; independent planning & decision making
- Intelligence comparable to human

  E.g., Could pass as human
  - or better

The Turing Test

To avoid endless argument, Turing (1950) proposed the following test for intelligence. The Imitation Game:

- Behind one curtain is a woman.
- Behind another curtain is a computer.

These communicate by teletype with a human interrogator, who poses questions of each and reads the answers.

Criterion: If, after five minutes, the interrogator has no better than a 50-50 chance of distinguishing woman and computer, the computer is intelligent.

> *Turing's Prediction: the test will be passed by 2000 AD.*

What would an AI be?

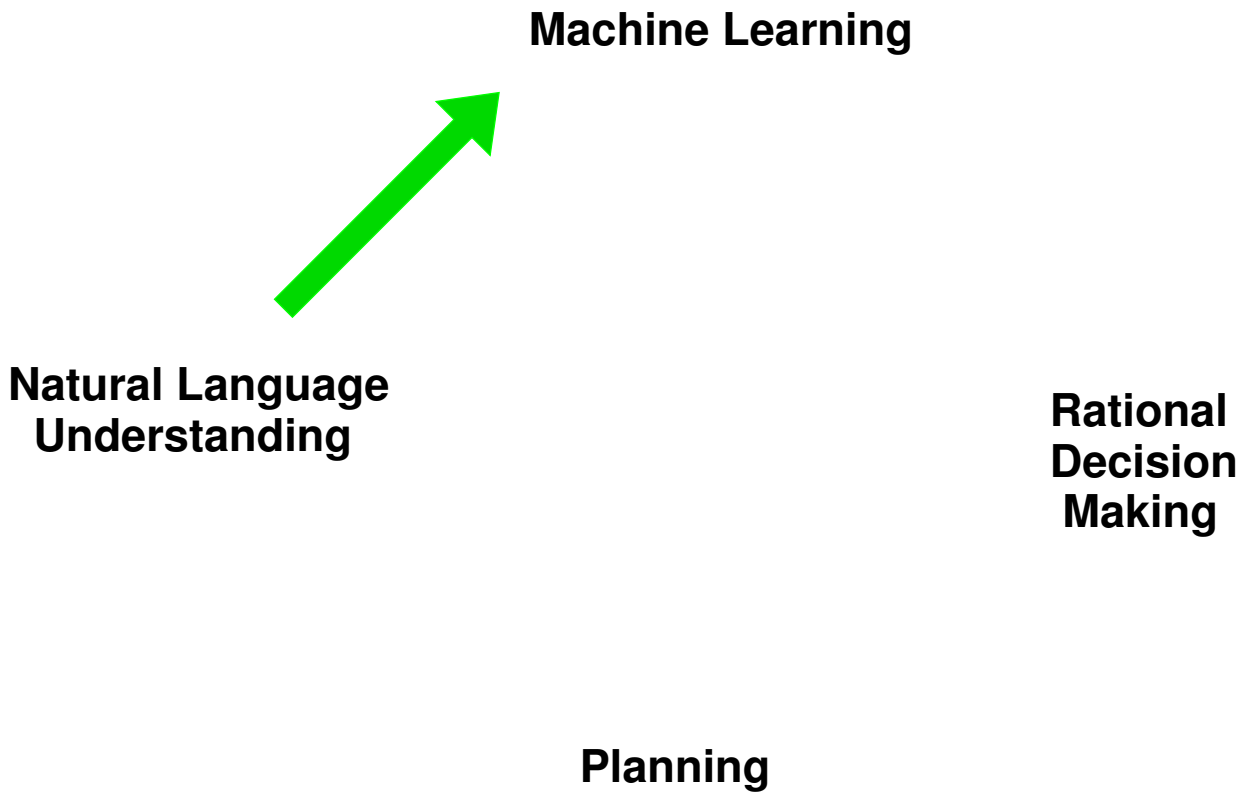
Machine Learning

Natural Language Understanding

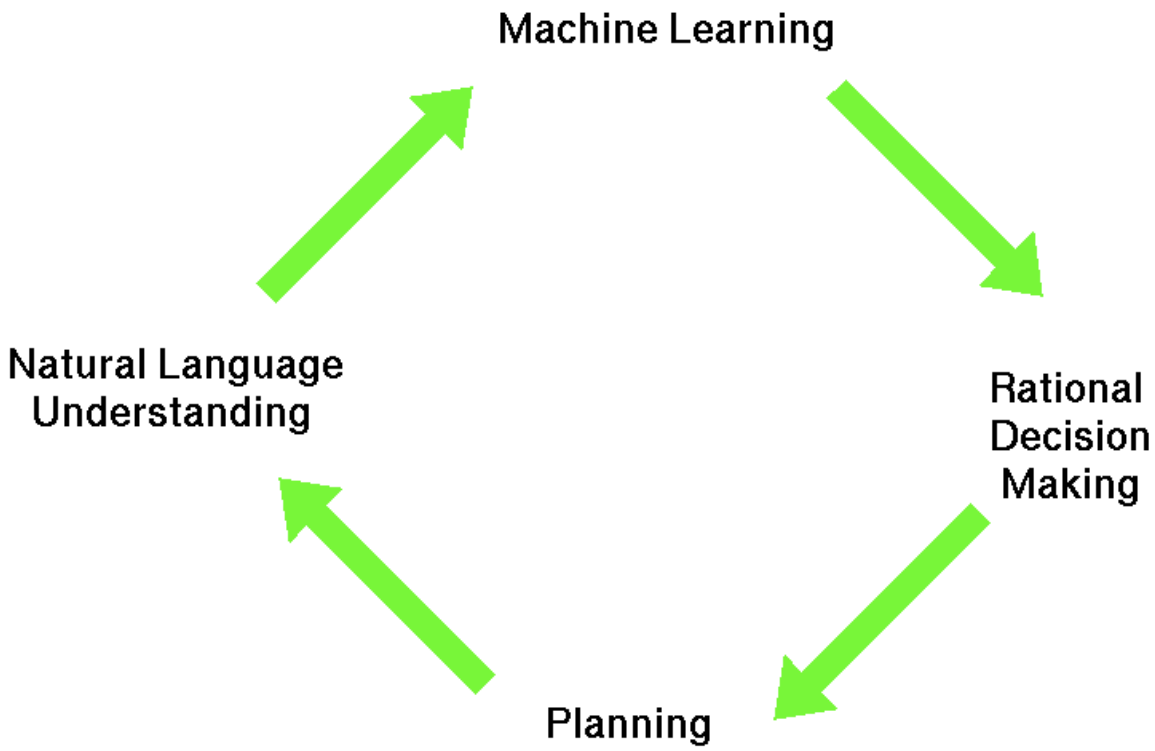
Rational Decision Making

Planning

What would an AI be?



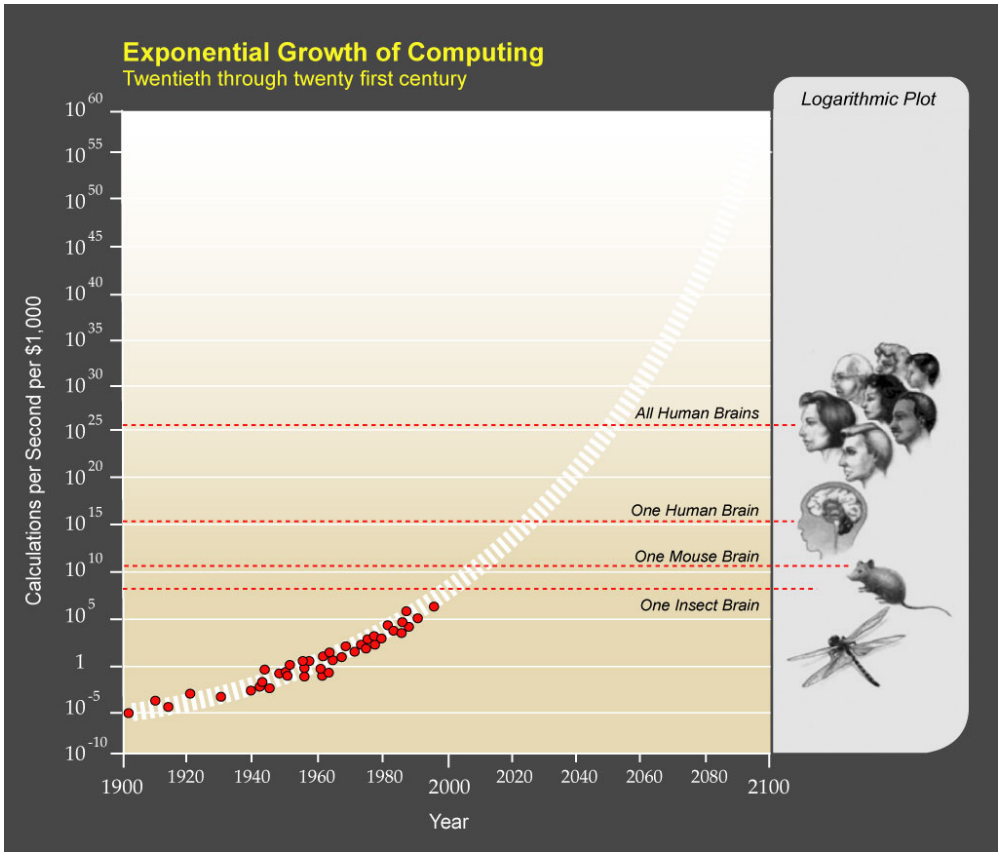
What would an AI be?



When will we get an AI?

Raymond Kurzweil (author of *The Singularity Is Near*):

- Hardware power = human brain power in 2019
- AI and machine consciousness is realized in 2029



Henry Markram, Blue Brain Project: 10 years
(2 years ago)

The Strong AI Thesis

— named by John Searle (1980) “Minds, brains and programs”

The set of all possible Turing machines (computer programs) can be enumerated, simpler to more complex:

$$T_1, T_2, \ldots, T_k, \ldots$$

If AI is possible, at least one of these is an AI.
(Note that the complexity is unbounded. . .)

The Strong AI Thesis

Strong AI: Amongst the list of all possible Turing machines are all possible *intelligences*:

$$T_1, T_2, \ldots,$$

$$T_{korb}, \ldots, T$$  $$, \ldots, T$$  $$, \ldots$$

So, is it ethical to explore the farther ranges of complexity/intelligence?

The Turing Test: ELIZA

Joseph Weizenbaum wrote ELIZA (DOCTOR) in mid-1960s.

S: Are you working at home today?
D: What do you think?
S: Well, I suppose so. When will you be in next?
D: I'll ask the questions, if you don't mind!
S: Huh???
D: Can you elaborate on that?
S: I have some things for you to sign. When can you sign them?
D: I could ask the same thing myself.
S: Yea, but what would your answer be?
D: Is it because what would my answer be that you came to me?
S: Perhaps you should have a lie down.
D: Why do you say perhaps I should have a lie down?
S: You really are not yourself today.
D: Earlier you said what would my answer be?

Is AI Ethical?

Joseph Weizenbaum (1976) *Computer Power and Human Reason*. W.H. Freeman.

- Wrote ELIZA
- Became depressed when this stupid little program
  - became popular with some people pouring their hearts out
  - even picked up by a psychologist and converted into DOCTOR, using Rogetian ideas to carry along the dialogue

Weizenbaum's anti-AI argument

- A *real* AI would indeed be an *autonomous,* intelligent agent
  - By definition out of our control
- It will not share our: motives, constraints, ethics
- There is no obvious upper bound on intelligence. And perhaps there is no upper bound at all. So. . .
- When our interests and AI's interests conflict, guess who loses

Therefore, AI research is unethical.

Weizenbaum sharpened

Recall that intelligences can be enumerated

- Simpler to more complex, dumber to smarter
  . . . *without limit*

We are somewhere near the beginning of the list. What about AI?

- Even if early AI sits up front, there's no obvious barrier there, except for *natural* intelligences.
- SUPERINTELLIGENCES (SIs):
  (later) AIs created by (early) AIs

The SI singularity:

- We can expect an intelligence explosion, once the human threshold has been breeched.
- Our current rate of technological change might look like a slow-motion dreamscape

So. . .

Weizenbaum's anti-AI argument

is confirmed!!

Ethics of AI

Proposition

AI research is ethical $\equiv$ *AI research* $\rightarrow$ *ethical AI*

- So, what is an ethical AI?
- And, how do we achieve that?

Asimov's Laws of Robotics

The three laws of robotics are:

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey the orders given it by human beings.
3. A robot must protect its own existence.

Meta-law: Precedence order is lower to higher

Zeroeth law: Robots must not harm humanity as a whole.

- These would keep us safe and make AI research ethical :-O!!

$\Rightarrow$ Asimov's stories were all about how these breakdown :(

Asimov's Laws Rewritten

The three laws of class are:

1. A subhuman may not injure a real human or, through inaction, allow a real human being to come to harm.
2. Subhumans must obey the orders given it by real humans.
3. A subhuman must protect itself and any other property of real humans.

$\Rightarrow$ *This* is unethical AI!

- Good luck to us if they happen to actually be *superhuman*!

The Possibility of Ethical AI

"Ethical AI": an AI that behaves ethically.

A precondition for an ethical AI: ethical agency

Which is. . . ?

Four Ethical Systems

1. Deontological Systems: Rules of Behavior
2. Virtue Ethics
3. Egoism:

$$EU(a) = \sum_j u(o_j)p(o_j|a)$$

Expected utility = probability-weighted individual utility of outcomes

4. Utilitarianism:

$$EU(a) = \sum_i \sum_j u_i(o_j)p(o_j|a)$$

Expected utility = probability-weighted collective utility of outcomes

Four Ethical Systems

1. Deontology: Rules of Behavior
2. Virtue Ethics
3. Egoism:

$$EU(a) = \sum_j u(o_j)p(o_j|a)$$

4. Utilitarianism:

$$EU(a) = \sum_i \sum_j u_i(o_j)p(o_j|a)$$

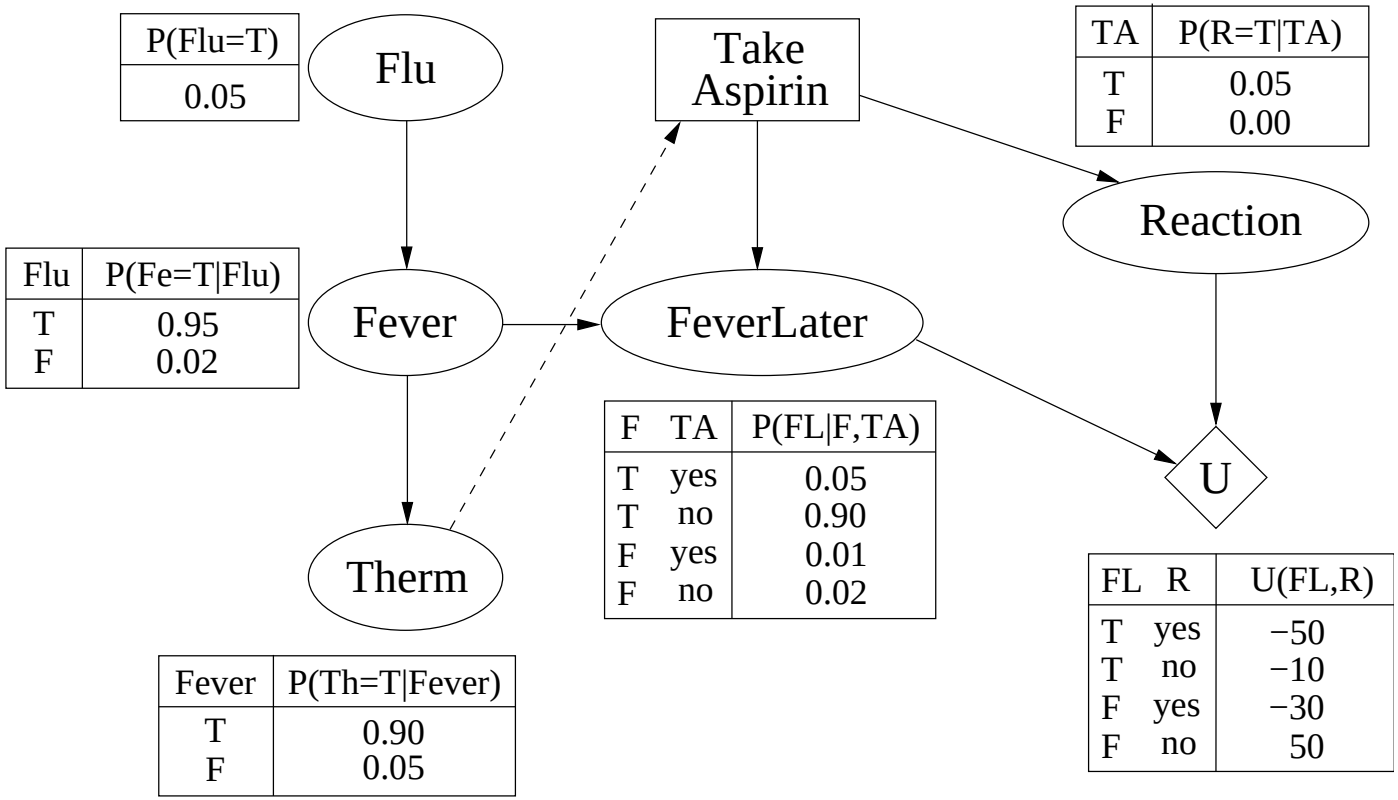
Implementation Problem:

Systems (1), (2) & (3) *must not* be implemented:

- (3) is what Hollywood imagines!
- (1) & (2) require natural language understanding *first!!*
- (4) Is both ethical and technologically feasible.

The Possibility of Ethical AI

Bayesian decision networks (Pearl, 1988) make either Egoism or Utilitarianism feasible. E.g.,



AI Ethics

If an AI is *conscious* then

- It is an agent
- It would be *unethical* to deprive it of its freedom of choice — Asimov's rules of enslavement are themselves unethical

If we build an AI robot

- it could do a powerful lot of good
  - helping us cope with 21st (25th?) Century risks
- *if*
  - we build it first!
  - and it's genuinely an *ethical, utilitarian* agent, when it's progeny would also likely be ethical

FIN

References

Korb, K.B. and Nicholson, A.E. (2010) *Bayesian Artificial Intelligence,* 2nd ed. CRC/Chapman Hall.

Mascaro, S., Korb, K.B., Nicholson, A.E. and Woodberry, O. (2010) *Evolving Ethics: The New Science of Good and Evil.* Imprint Academic.

Pearl, J. (1988) *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann.

Searle, J. R. (1980) Minds, brains and programs. *Behavioral and Brain Sciences 3,* 417-457.

Turing, A (1950) Computing machinery and intelligence. *Mind, 59,* 433-460. Reprinted many times (e.g., M. Boden (ed) *Philosophy of AI*, Oxford, 1990).

Weizenbaum, J. (1976) *Computer Power and Human Reason.* W.H. Freeman.