Lecture 09 September 2025

Design for Values and AI

Philip J. Nickel

Lecture Overview

Review: Technical AI, Strong AI, AGI

The Design for Values Paradigm

Some cases discussed in CHLEP ch. 3

Applying DfV in your assignment

Review

Technical AI

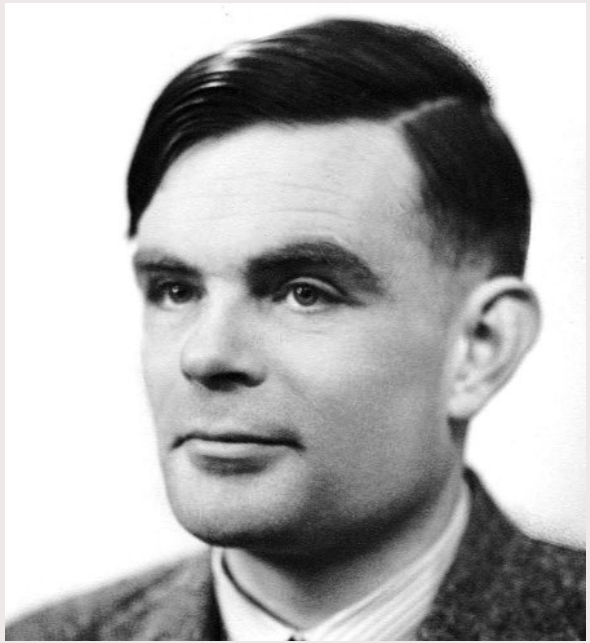
- Based on a set of technical methods for simulating learning and reasoning in machines
- Reasoning is deductive or inductive and follows the rules of logic
- Technical AI machines are oriented toward automating a task… with the possible exception of generative models not designed for any particular task.
- Robustness is defined in terms of the reliability of a model for new data.
- The EU approach to trustworthy AI incorporates ethics and robustness

Some foundational aims

Classical AI aims to build artificial machines that have intelligence. It holds that any feature of (human) intelligence can be built into a machine.

In addition, it holds that a machine so built would help us understand (human) intelligence.

Strong AI actually has a mind and cognition.

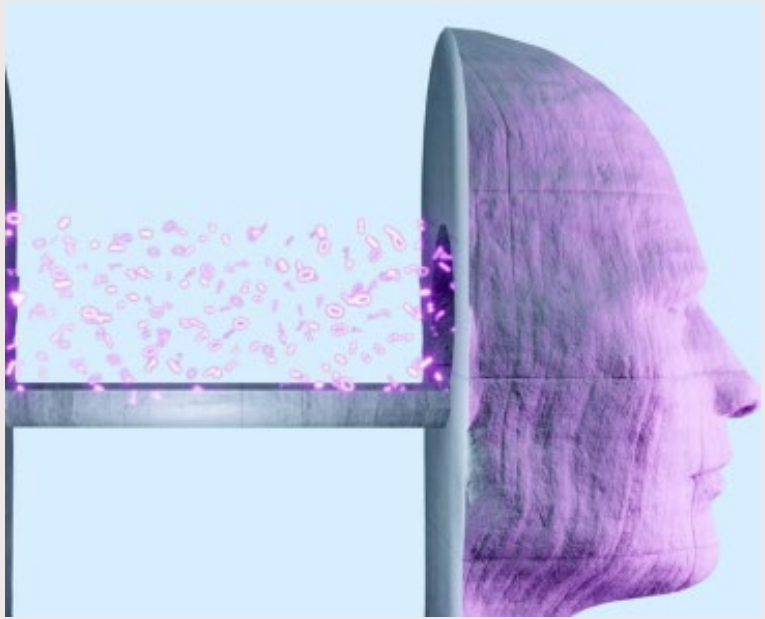


By Elliott & Fry - https://commons.wikimedia.org/wiki/File:Alan_Turing_(1951).jpg, Public Domain, https://commons.wikimedia.org/w/index.php?curid=158923803

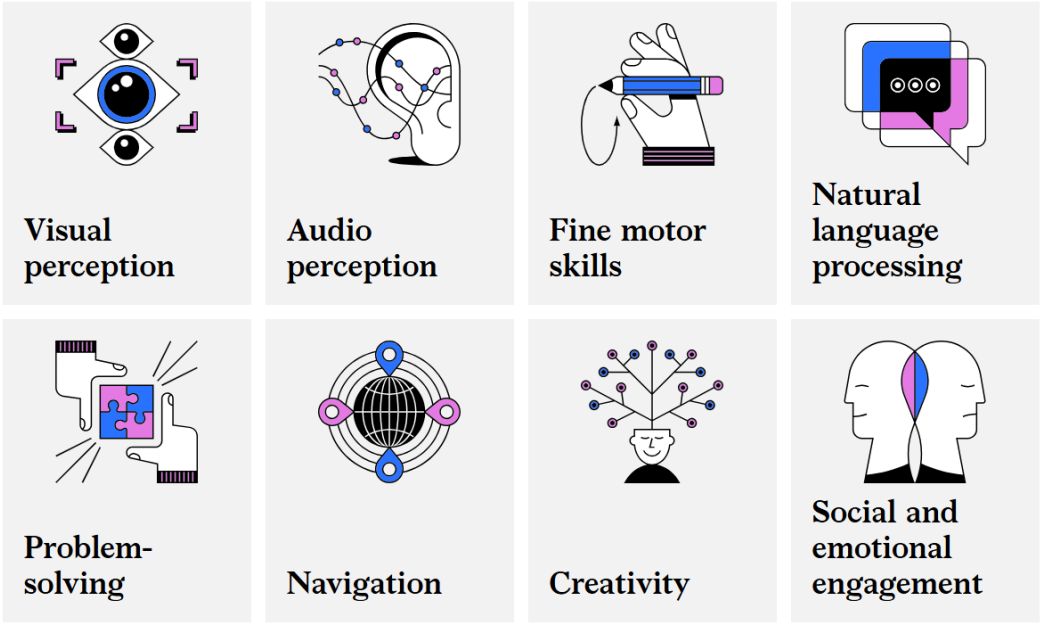
Artificial General Intelligence

Müller: We can have the ambition to recreate human or superhuman intelligence, while maintaining the current methodology of Technical AI (with which you are familiar or are becoming familiar), a set of computer-science methods for recreating the component capacities of human cognition.

Image: What is Artificial General Intelligence (AGI)? | McKinsey



McKinsey claims these skills are needed for AGI



Philosophy of AI

What does it take for computation to be intelligent (“to think”)? AI has prompted philosophers, cognitive scientists, and computer scientists to consider carefully what we mean by such terms.

Arguments that intelligence is computation-plus

The 4E argument: only embodied computation that interacts with the environment can be intelligent

The Chinese room argument: only computation with intentionality can be intelligent

End of review

Design for Values

In CHLEP Ch. 3, Design for Values and Value Sensitive Design are taken as equivalent (variations of a single type of approach).

Design for values is an approach to incorporating ethical values into a technological design, in which there is input from:

- Stakeholders
- Ethical theories, esp. mid-level ethical theories

The challenge of AI in Design for Values

Some of AI's "disruptive" features create special challenges for ethics, and for Design-for-Values:

- It is being adopted very quickly.
- It is applied at a massive scale.
- It can take intelligent steps independent of human judgment and intervention.
- It is often opaque, i.e., the process by which it takes steps cannot be recovered easily in the case of deep learning or LLMs.

Some methodological assumptions of DfV

It is both possible and necessary to take steps early on in order to ensure that ethical values are built into our technological solutions.

Each technology, considered in a narrow sense, can be embedded in social practices in different ways to create different socio-technical systems (“applications,” technologies in the broad sense).

Our values can change over time.

Design for Values

Design for Values is described as having three aspects (CHLEP p. 72):

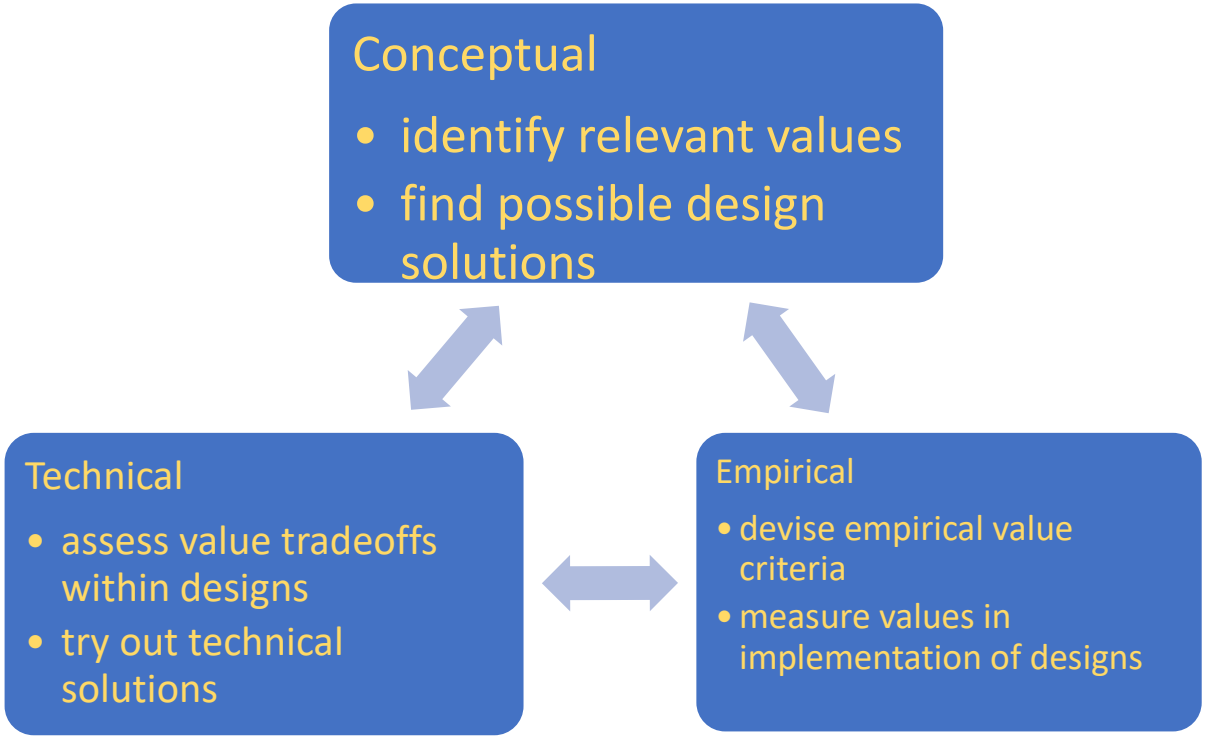
1. Identifying the relevant values
2. Embedding these values in systems
3. Assessing whether these efforts have been successful

Source: Friedman, Kahn & Borning 2006.

Part I: Value Sensitive Design

Approach developed by Batya Friedman and colleagues ~ 20 years ago to integrate ethical values systematically in engineering design.

Value Sensitive Design



Where do the values come from?

List from Friedman et al. 2005:

human welfare

ownership & property

privacy

freedom from bias

universal usability

trust

courtesy

autonomy

informed consent

accountability

calmness

identity

environmental sustainability

Value Sensitive Design

- What values are relevant? Are there tradeoffs? (Conceptual)
- Are there technical solutions to value tradeoffs? (Technical)
- How do we measure the success of these solutions? (Empirical)

Source: Friedman, Kahn & Borning 2006.

(a) "The Watcher" (b) The HDTV Camera (c) "The Watched"

Figure 2. Plasma Display Technology Studies

From the standpoint of illustrating Value Sensitive Design, we woul

Part I: Ethical values

Non-ethical vs. ethical values

Fundamental ethical values

Some values are so important and general that they are taken as fundamental in ethics:

- **Happiness/welfare**
- **Autonomy/respect for autonomy**
- **Justice**

Other values, such as privacy or safety, are sometimes thought to be less fundamental, in the sense that they are justified in terms of these fundamental values.

Interpreting fundamental values

All ethical values are subject to interpretation, and you should be able to explain and defend your interpretation, showing an awareness of other possible interpretations. Example:

Justice:

Everybody…

- **Gets what they deserve?**
- **Gets an equal amount?**
- **Gets what they need?**

http://www.365thingsin365days.com/wp-content/uploads/2012/01/pieInTheFace_07-1024x574.jpg

A few important distinctions

- **Descriptive vs. normative judgments**
- **Non-ethical vs. ethical value**
- **Intrinsic vs. instrumental value**

Understanding these distinctions will help you discuss ethical values in a logical, clear way.

Descriptive vs. normative judgments

“This building is square.”

“This building is good.”

“This building is

- **Smart.”**
- **Sustainable.”**
- **Original.”**
- **Safe”**

Daniel Burnham, Ellicott Square Building, 1895-6
http://library.buffalo.edu/pan-am/img/fig5lg.jpg

Intrinsic vs. instrumental value

http://www.uroubc.ca/wp-content/uploads/2015/07/7030094-cute-animals.jpg

Design for Values: sources of values

DfV differs from VSD in its emphasis on theory as a source of values.

General ethical theories (utilitarianism, deontology, and virtue ethics) provide “salience lenses” that allow us to identify some issues as ethically significant.
“Mid-level” theories (Rawls’ theory of justice as fairness, the Capability Approach) provide more specific guidance about values to include. (Here, I would also turn to foundational sources such as the UDHR and the Belmont Report.)
Participant and stakeholder values are also included, but in a qualified way.

Incorporating ethics into AI applications: some approaches

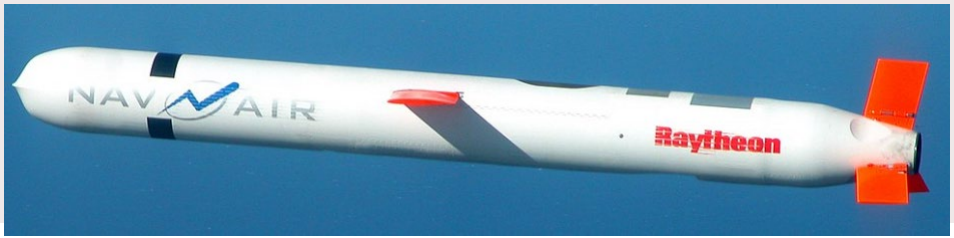
1. Build the system using ethical rules.
("Whenever you detect a human, stop the robot")
2. Train systems on large data sets in which the rules are followed (e.g., cases where the robot stops for humans)
3. Train systems on large data sets in which the rules might not have been followed, but impose the rules as filters.
4. Rely on the socio-technical system (users) to provide ethical input.

Cases from CHLEP Ch. 3

(applications of) ChatGPT and other LLMs
AI for determining loan eligibility
AI for benefits fraud detection

Some limitations of Design for Values

- Some ethical issues appear too late in the design process to be incorporated into the design.
- Design for Values is very resource-intensive, especially if it has to be specified for each application.
- Some technologies may be impossible to ethically “optimize”
- In cases of high moral uncertainty (‘moral disruption’), it may be impossible to specify values in the way required by Design for Values.



https://commons.wikimedia.org/wiki/File:Tomahawk_Block_IV_cruise_missile_-crop.jpgmedia Commons image source