True or False?

- AI will soon replace humans in most jobs
- AI will surpass human intelligence in the next X years
- AI works similarly to the human brain
- AI systems “think”
- AI systems have their own desires and goals
- AI systems can do things that their designers didn’t intend
- AI systems could become conscious
- AI systems can be trusted
- AI algorithms can discriminate or exhibit prejudice
AI Everywhere
“But what about Skynet?”
Utopia or Dystopia?
Not so fast…
DR. SARAH ABRAHAM
CS349

ETHICAL FRAMEWORKS
ETHICS IN COMPUTER SCIENCE

- Machine ethics
  - The moral behavior of artificially intelligent beings
- Roboethics
  - The moral behavior of designing artificially intelligent beings
- Computer ethics
  - The moral behavior of using computers and computing systems
WHAT ARE ETHICAL FRAMEWORKS?

- Systems that guide ethical choices and provide a reason for that choice
- This is an unsolved problem!
  - Numerous approaches that result in vastly different outcomes and behaviors
- Three broad frameworks:
  - Duty-based framework
  - Consequentialist framework
  - Virtue framework
ETHICAL THEORIES

- Non-consequentialist
  - Concerned with agent’s intent rather than consequence
- Consequentialist
  - Concerned with consequence of agent’s actions
- Agent-centered
  - Concerned with ethical makeup of agent
NON-CONSEQUENTIALIST (DUTY-BASED)

- Often associated with Immanuel Kant’s “categorical imperative”
  - “Act only according to that maxim by which you can at the same time will that it should become a universal law.”
- Ethical conduct means choosing actions that are right and good
- Consider duties and obligations when choosing
PROBLEMS?

- Good intents are valued over good outcomes
- Does not answer how to act when two duties conflict
- Does not provide definition of ethical behaviors
CONSEQUENTIALIST

- Based on Utilitarian philosophy
  - Weights good and bad produced by action to determine overall best action
- Ethical conduct means attempting to do the most good and the least harm
- Considers the impact on all individuals involved when choosing
PROBLEMS?

- The needs of the many override the needs of the few
- Any action can be justified if enough good comes out of it
- Does not address how to predict outcomes based on actions
AGENT-CENTERED (VIRTUE)

- Based on ideas of Aristotle and Confucius
  - Agents should act according to their ideal self
- Ethical conduct means determining an agent’s traits and behaviors and building on those that foster good
- Considers entirety of an agent’s life rather than individual actions
PROBLEMS?

- Focuses on personal character rather than a system for determining action
- High level approach requires a depth of understanding and interpretation to implement effectively
- Does not define virtuous traits
SOME OTHER VARIANTS

- **Duty-based**
  - Rights approach: rights of those affected should be protected
  - Fairness approach: fair and equal conduct is ethical conduct
  - Divine command: higher powers dictate what is right

- **Consequentialist**
  - Ethical egotism: self-interest leads to interest and respect of others
  - Common good approach: what is good for the general society is what is good for all

- **Virtue**
  - Eudaimonism: living a good life leads to ethical action
  - Ethics of care: feminist theory that care and nurturing should be valued as much as justice and autonomy
ETHICAL DILEMMAS

- Rushworth Kidder defines ethical dilemmas as choices that are right vs right:
  - Truth vs loyalty
  - Justice vs mercy
  - One vs many
  - Short-term vs long-term
Acting Rationally

Acting rationally simply means maximizing utility

…but can this go wrong?
Unforeseen Consequences of Maximizing Utility?
What went wrong?

• **Is this realistic?**
  • Robots aren’t smart enough to learn from one example. But let’s assume they will be able to in the future.
  • It wouldn’t have a concept of “human” to go seek out. It only knows about colors and cubes.

• **Bad design!**
  • Objectives must be designed carefully: robot should only be rewarded for making cubes red, not any object.
  • Actions should be limited: only actions available should be to pick up a block or paint a block.
  • Plans should be verified for safety before / during execution: cancel any trajectory that will come in contact with a human.
  • Don’t continue learning after deployment.

• **Is this any more dangerous than any factory with non-intelligent machinery that doesn’t automatically stop if someone is in the way?**
  • It is bad design, but we know how to use engineering to avoid these situations!
<table>
<thead>
<tr>
<th>Human</th>
<th>AI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Evolved for survival</td>
<td>Designed by engineers</td>
</tr>
<tr>
<td>Sets own goals</td>
<td>Goals programmed explicitly (usually)</td>
</tr>
<tr>
<td>Complex, general purpose system</td>
<td>Specific, constrained system</td>
</tr>
<tr>
<td>Continually learns</td>
<td>Can turn off learning, or not use learning</td>
</tr>
<tr>
<td>Learns from all observed data</td>
<td>Data access can be controlled</td>
</tr>
<tr>
<td>Learns only from own experiences</td>
<td>Can share data with other robots</td>
</tr>
<tr>
<td>Can make any choice at any time</td>
<td>Available actions can be restricted</td>
</tr>
</tbody>
</table>
Realistic Risks of AI

Mass unemployment due to automation
Realistic Risks of AI

Substandard testing / poor user understanding
Realistic Risks of AI

How to make tough decisions?
Realistic Risks of AI

Privacy concerns
Realistic Risks of AI

Algorithmic discrimination
Realistic Risks of AI

Unethical emotional manipulation
Realistic Risks of AI

Unethical usage: drone warfare?
Realistic Risks of AI

AI in the “wrong hands”
Realistic Benefits of AI

The central question:

Can we ensure that the benefits of AI outweigh the potential risks?
Realistic Benefits of AI

Significant reduction of driving fatalities
Realistic Benefits of AI

Happier, healthier lives
Realistic Benefits of AI

Increased productivity and prosperity
Realistic Benefits of AI

Dirty, dangerous, and dull
Realistic Benefits of AI

Greater social justice
Realistic Benefits of AI

Beyond human capabilities
Realistic Benefits of AI

But what does social good really mean?

Final project will give you a chance to define this for yourself!