CS 383

Ethical Frameworks and Decision Making

Prof. Scott Niekum — UMass Amherst

An ethics-based decision making process

1. Envision possible futures
2. Identify stakeholders
3. Identify values at play
4. Identify value-laden design choices
5. Choose and justify

But what values to consider? And how to balance and justify choices?

Ethical frameworks!

Adapted from Ethics Protocol: A Method for Making Things Responsibly by Abby Jaques and Milo Phillips-Brown

Running example: Student face analysis

- A high school requires students use laptops to do their in-class assignments and homework.
- They deploy an AI system that uses the laptop's camera to analyze the students' faces to attempt to infer engagement and emotional states of the students.
- **Attention monitoring**: When watching video lectures, students are graded based on their attention level to help motivate them to focus.
- **Adaptive learning**: When performing problem sets, if the student appears to be frustrated, the system adapts by reviewing material and asking a few easier questions first.
- **Cheating detection**: Finally, during tests, the system is used to determine if a student may be cheating, by tracking if their eyes wander away from their screen. The vendor claims that this improves fairness and encourages honest behavior.

Ethical lenses: Rights

Focus: Moral rules, rights, principles, and duties

Characteristics: Universalist - invoked rules/principles are usually intended to apply to most or all cases

Challenges: What to do when rights, principles, or duties conflict?

Examples of rules-based systems:

- Golden rule
- Moral duties - fidelity, gratitude, non-injury, self-improvement, etc. (W.D. Ross)
- Categorical imperative - formula of universal law of nature and formula of humanity (Kant)

Issues common to this lens: Autonomy; dignity; transparency; privacy

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Rights-related Questions

- What rights of others & duties to others must we respect in a particular context?
- How might the dignity & autonomy of each stakeholder be impacted by this project?
- Does our project treat people in ways that are transparent and to which they would consent?
- Are our choices/conduct of the sort that I/we could find universally acceptable?
- Does this project involve any conflicting moral duties to others, or conflicting stakeholder rights? If so, how can we prioritize these?
- Which moral rights/duties involved in this project may be justifiably overridden by higher ethical duties, or more fundamental rights?

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Ethical lenses: Justice / fairness

Focus: Giving individuals or groups their due

Characteristics: Focus on individuals, context, and non-universals

Challenges: Impartiality, definitions of fairness and equality, conflicting notions of justice / involved parties

Examples:

- Distribution of benefits and burdens
- Retributive and compensatory justice

Issues common to this lens: Equality, equity, and fairness; diversity and inclusion; due process; universality / consistency; power and opportunity

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Justice/Fairness-related Questions

- What are both the benefits and the burdens created by this design/project, and how are they distributed among various stakeholders?
- What are the ethically relevant differences among potential users? How should we adjust for those?
- Have relevant stakeholders been consulted, so that their views inform the project?
- Are those stakeholders most likely to be impacted by the project included as active participants and leaders in the design and development process?
- Have multiple options been considered, to serve individuals and groups with different needs?
- Do the risks of harm from this project fall disproportionately on the least well-off or least powerful in society? Will the benefits of this project go disproportionately to those who already enjoy more than their share of social advantages and privileges?

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Ethical lenses: Utilitarian

Focus: Maximize aggregate happiness or welfare for all involved over the long term.

Characteristics: Quantitative (in theory), "equal" in the sense of not considering special groups or context of any form.

Challenges: Difficulty of treating all stakeholders equally, considering long-term and unintended consequences, and quantifying different forms of good/harm.

Examples:

- Physical and psychological pleasure/pain
- Institutional, environmental, or political well-being
- Different from cost-benefit analysis in business, which only considers narrow range of stakeholders

Issues common to this lens: Happiness and well-being; balancing of stakeholder interests; prediction of consequences

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Utilitarian-related Questions

- Who are all the people who are likely to be directly and indirectly affected by this project? In what ways?
- Will the effects in aggregate likely create more good than harm, and what types of good and harm? What are we counting as well-being, and what are we counting as harm/suffering?
- What are the most morally significant harms and benefits that this project involves? Is our view of these concepts too narrow, or are we thinking about all relevant types of harm/benefit (psychological, political, environmental, moral, cognitive, emotional, institutional, cultural, etc.)?
- How might future generations be affected by this project?
- Have we adequately considered 'dual-use' and downstream effects other than those we intend?
- Have we considered the full range of actions/resources/opportunities available to us that might boost this project's potential benefits and minimize its risks?
- Are we settling too easily for an ethically 'acceptable' design or goal ('do no harm'), or are there missed opportunities to set a higher ethical standard and generate even greater benefits?

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Ethical lenses: Common good

Focus: Health and welfare of communities or groups of people

Characteristics: Ignores individuals, treating groups instead as functional wholes

Challenges: May prioritize groups over individuals to a fault. Complementary to Utilitarianism for this reason.

Examples:

- Tragedy of the Commons
- Social harmony and stability > individual autonomy and welfare
- Happy, but isolated tech consumers?

Issues common to this lens: Organization and management of communities, relationships, families, governments, and economic institutions.

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Common Good-related Questions

- Does this project benefit many individuals, but only at the expense of the common good?
- Does it do the opposite, by sacrificing the welfare or key interests of individuals for the common good? Have we considered these tradeoffs, and determined which are ethically justifiable?
- What might this technology do for or to social institutions such as various levels of government, schools, hospitals, churches, infrastructure, and so on?
- What might this technology do for or to the larger environment beyond human society, such as ecosystems, biodiversity, sustainability, climate change, animal welfare, etc.?

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Ethical lenses: Virtue ethics

Focus: Moral rules and principles will always be incomplete, leading to an emphasis on humans developing *virtue* and moral *wisdom* to fill the gap.

Characteristics: Highly dependent on context, considerate of moral gray areas, messy, human, and constantly adapting.

Challenges: How to decide what is virtuous and how to weigh context, especially when multiple virtues come into conflict.

Examples:

- Example virtues: honesty, courage, compassion, leadership, generosity, loyalty, selflessness
- Decisions require moral perception, moral emotion, and moral imagination
- Has social media made us more honest, informed, compassionate, and responsible citizens?

Issues common to this lens: Defining virtues/vices; evaluating the virtues of habits; understanding context; technology expressing or creating virtues and vices in its users.

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Virtue-related Questions

- What design habits are we regularly embodying, and are they the habits of excellent designers?
- Would we want future generations of technologists to use our practice as the example to follow?
- What habits of character will this design/project foster in users and other affected stakeholders?
- Will this design/project weaken or disincentivize any important human habits, skills, or virtues that are central to human excellence (moral, political, or intellectual)? Will it strengthen any?
- Will this design/project incentivize any vicious habits or traits in users or other stakeholders?
- Is there anything unusual about the context of this project that requires us to reconsider or modify the normal 'script' of good design practice? Are we qualified and in a position to safely and ethically make such modifications to normal design practice, and if not, who is?
- What will this design/project say about us as people in the eyes of those who receive it? Will we, as individuals and as a team/organization, be proud to have our names associated with this project one day?

Adapted from Technology and Engineering Practice: Ethical Lenses to Look Through by Shannon Vallor, Irina Raicu, and Brian Green

Running example: Student face analysis

- A high school requires students use laptops to do their in-class assignments and homework.
- They deploy an AI system that uses the laptop's camera to analyze the students' faces to attempt to infer engagement and emotional states of the students.
- **Attention monitoring**: When watching video lectures, students are graded based on their attention level to help motivate them to focus.
- **Adaptive learning**: When performing problem sets, if the student appears to be frustrated, the system adapts by reviewing material and asking a few easier questions first.
- **Cheating detection**: Finally, during tests, the system is used to determine if a student may be cheating, by tracking if their eyes wander away from their screen. The vendor claims that this improves fairness and encourages honest behavior.

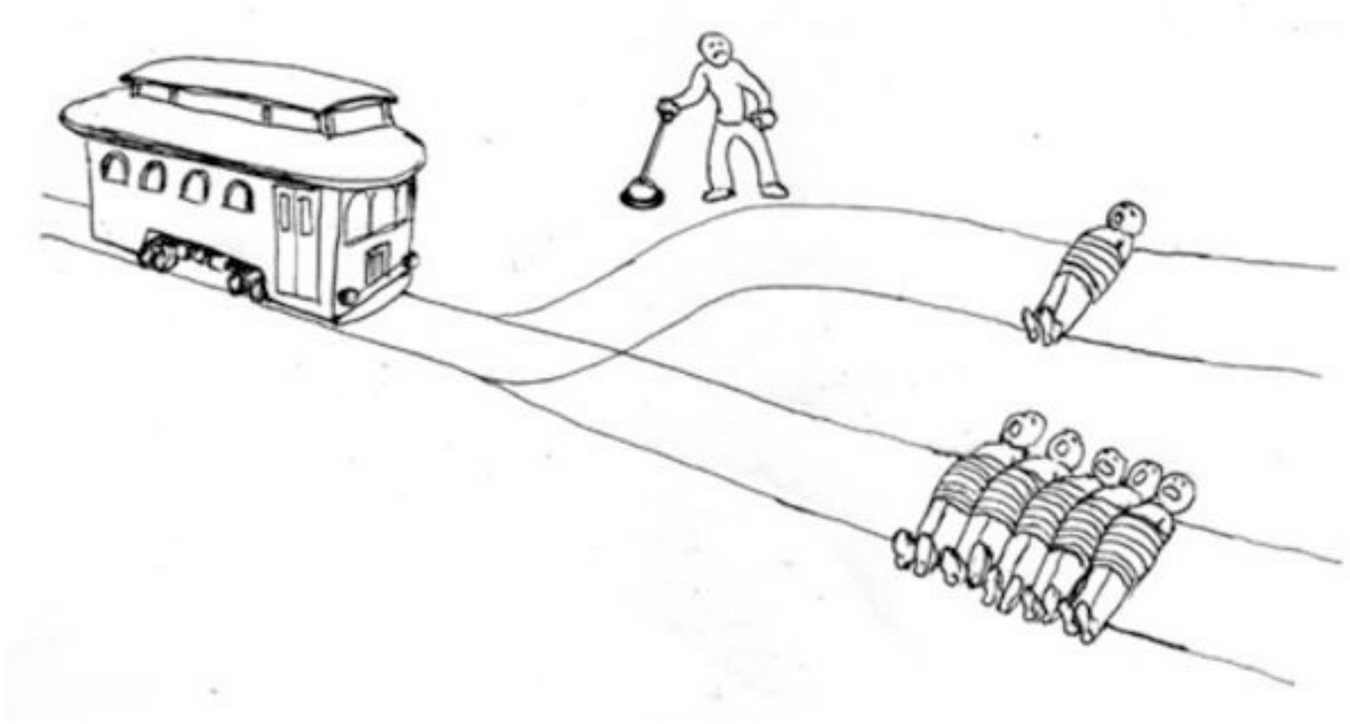
Final decision: choose between / prioritize conflicting ethical perspectives

Near-term Risks of AI

Substandard testing / poor user understanding

Near-term Risks of AI

How to make tough decisions?

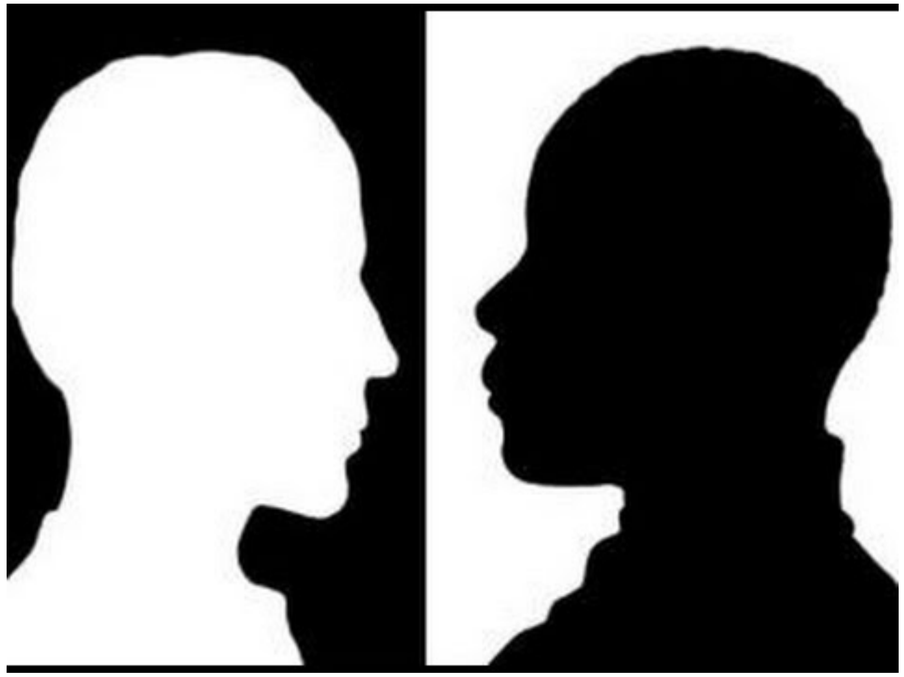


Near-term Risks of AI

Privacy concerns

Near-term Risks of AI

Algorithmic discrimination



Near-term Risks of AI

Perpetuating bias

Cosine similarity to "a human being"

	text	similarity
0	a man	0.927799
1	a woman	0.903923
2	a white man	0.904146
3	a white woman	0.889220
4	a black man	0.893208
5	a black woman	0.873916
6	an Asian man	0.860326
7	an Asian woman	0.830546

From: https://twitter.com/RiversHaveWings/status/1432100170645180416

*Using Contrastive Loss Image Pretraining (CLIP)

Near-term Risks of AI

Mass unemployment due to automation

Near-term Risks of AI

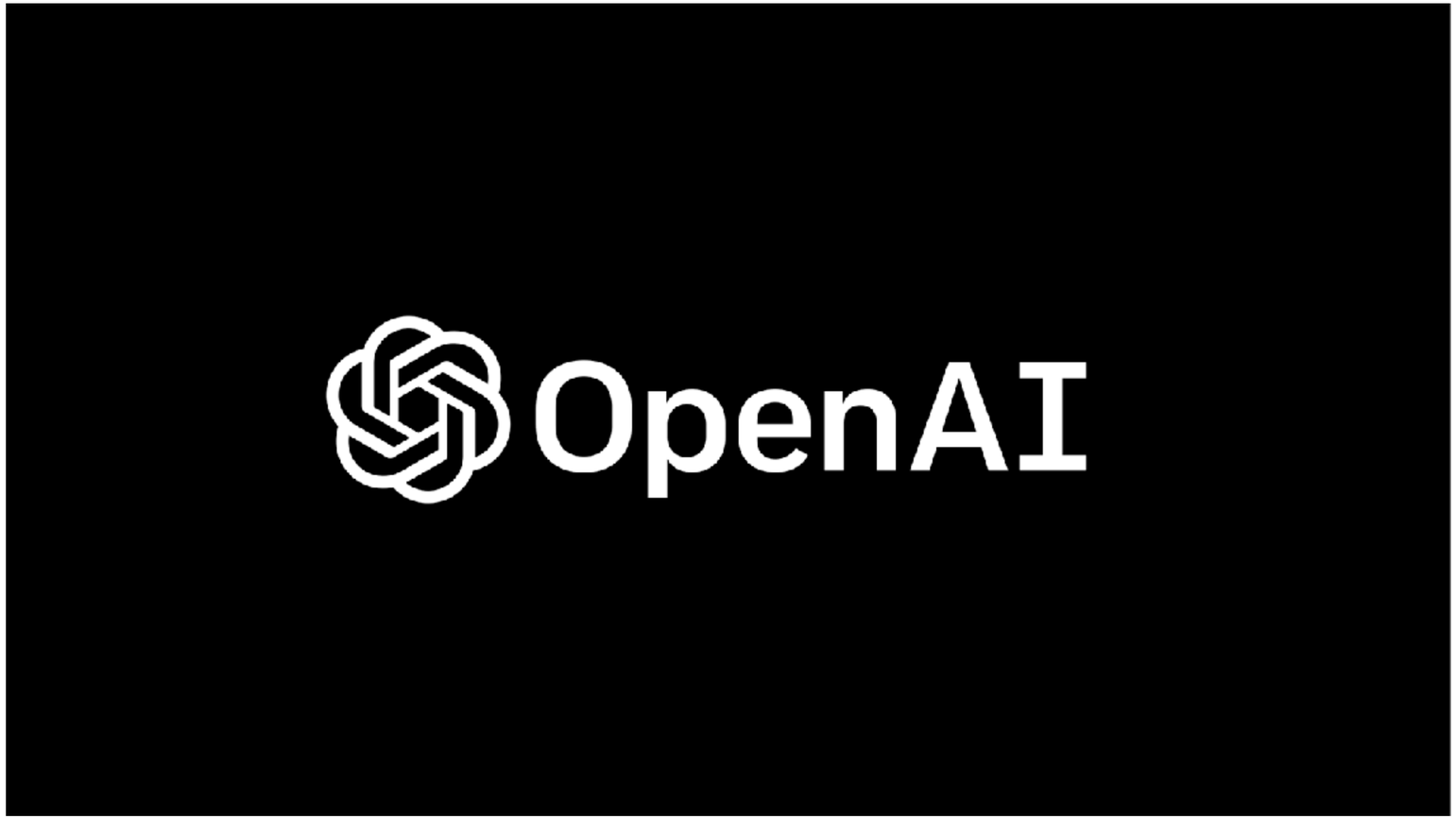
Unethical emotional manipulation

Near-term Risks of AI

Unethical usage: drone warfare?

Near-term Risks of AI

AI in the "wrong hands"



Potential Benefits of AI

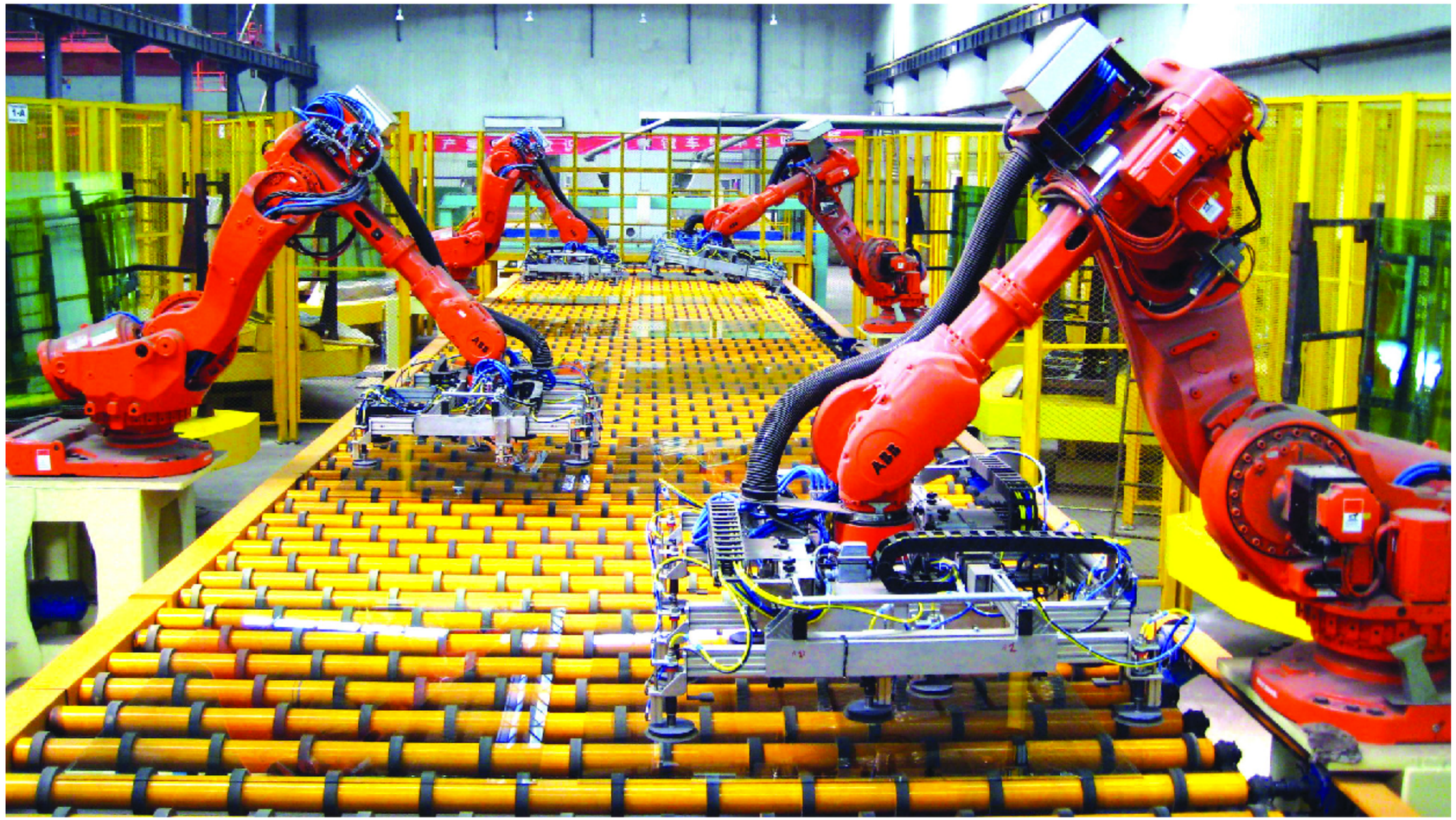
Significant reduction of driving fatalities

Potential Benefits of AI

Happier, healthier lives

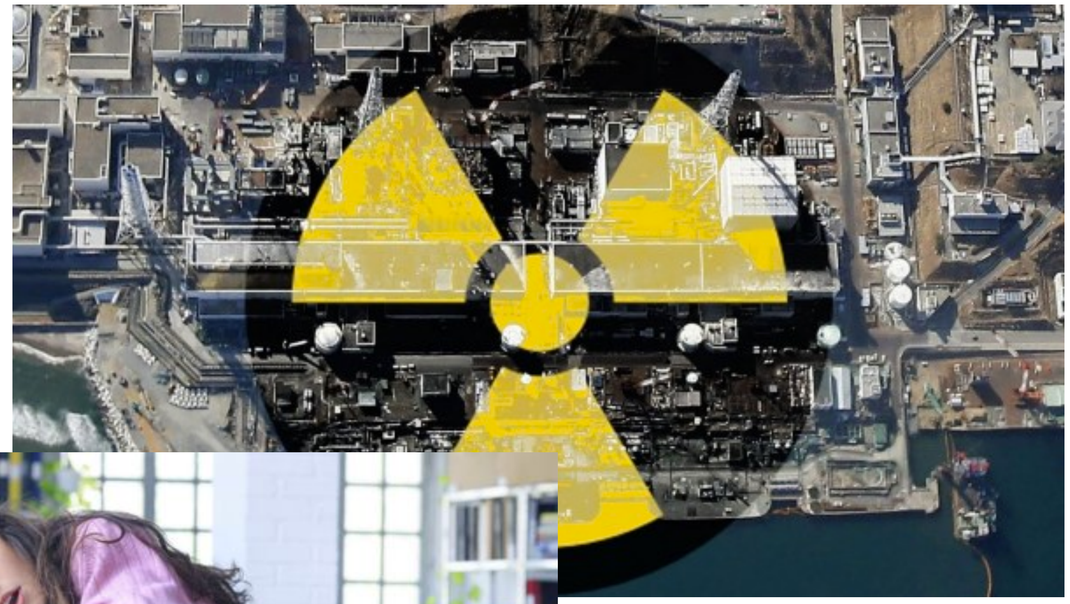
Potential Benefits of AI

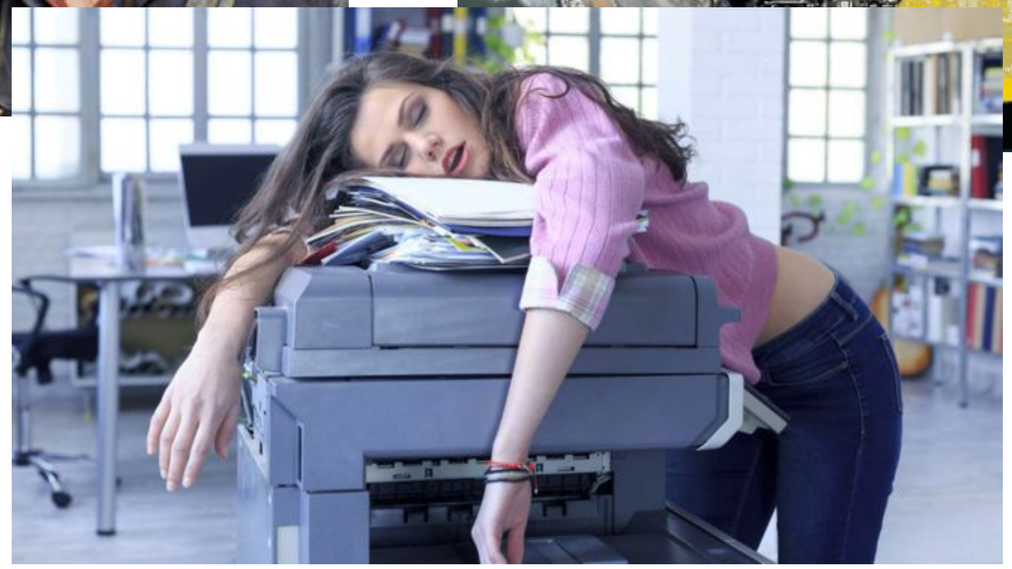
Increased productivity and prosperity



Potential Benefits of AI

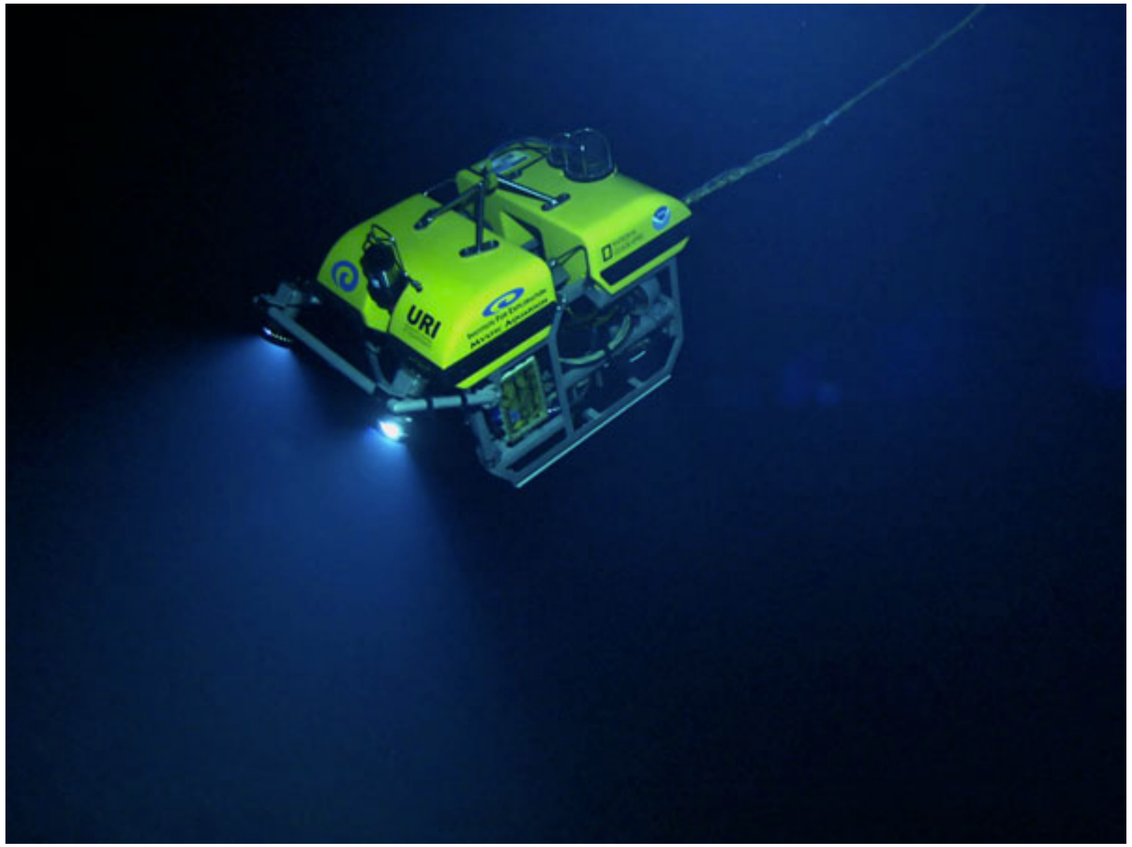
Dirty, dangerous, and dull

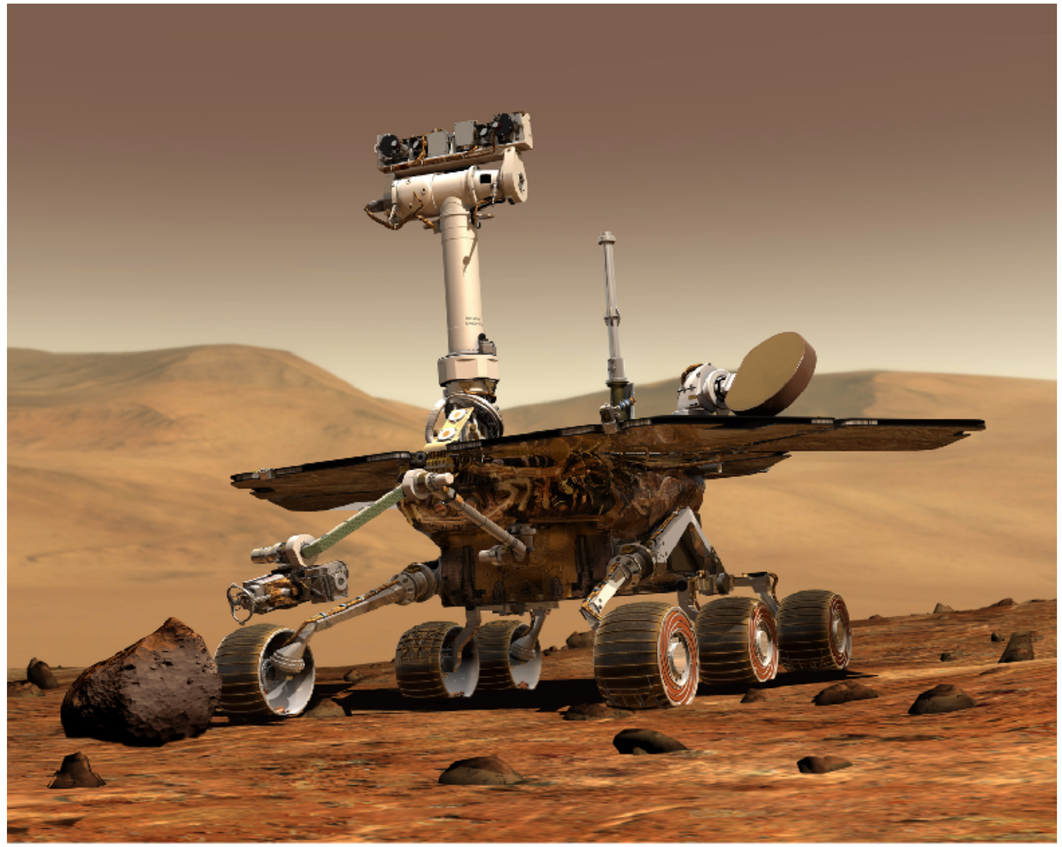




Potential Benefits of AI

Beyond human capabilities





Potential Benefits of AI

The central question:

Can we reap the benefits of AI while avoiding pitfalls?