Introduction to AI Safety, Ethics, and Society

Beneficial AI and Machine Ethics

Introduction

Machine ethics is the study of how to ensure that AIs behave ethically towards others.

In this lecture, we discuss various approaches to embedding ethics into AIs, and the challenges of each.

Roadmap

1. Law: *Make AIs follow the law*
2. Fairness: *Make AIs be fair*
3. The Economic Engine: *Let the economy govern AI development*
4. Wellbeing: *AIs should improve people's lives*
5. Preferences: *Make AIs satisfy people's preferences*
6. Happiness: *Make AIs make people happy*
7. Social Welfare Functions: *Make AIs maximize total wellbeing*
8. Moral uncertainty: *Deciding without knowing the "correct" moral theory*

Law

Why not just have AIs follow the law?

The Case for Law

1. It is a **legitimate** representation of our moral codes (if democratic).
2. It is **time-tested** and **comprehensive.**
3. It uses both **rules** and **standards.**
4. It is **specific** and **adaptable.**

Law Aggregates Values

In a democracy, the populace influences the law:

- Electing judges
- Voting on new laws
- Protesting/advocating
- Directly/through representatives

Thus, the law is **legitimate** : citizens have substantial input into its design. This is unlike AIs designed by big tech companies.

Law is clearer and less arbitrary than ethics, which is not standardized or codified.

Features of Law

The law is **time-tested** and **comprehensive**; a much-iterated approximation of our values:

- Generations successively creating, altering, enforcing, interpreting.
- Producing a record of what has worked.

The law uses both **rules** and **standards**:

- Rules: rigid instructions which reduce interpretation problems.
- Standards: flexible statements which allow for proportional responses and reduce the problems of over/under-inclusivity.

Law is Specific and Adaptable

The **specificity** of the law helps to reduce the risk of **AI misinterpretation**:

- Misinterpretation: the problem of AIs taking short/open-ended instructions too literally
- Laws are ideally specific in detail, and straightforward to interpret

The **adaptability** of the law helps to reduce the risk of **AI gaming**:

- Gaming: the problem of AIs "playing" the system of rules to their advantage, in undesirable ways
- Rules often fail here: Tax codes have not eradicated tax evasion
- Standards can help by leaving room for adaptable interpretation

In the future, AIs could help make the law more robust

The Need for Ethics as Well as Law

1. **Silent** : there are areas of human life where the law gives no advice.

2. **Immoral** : many laws clash with human morals.

Silent Law (1/2)

The law is not a complete guide to behavior, but a set of constraints.

It does not cover many important areas of our lives:

- Accidental silence
- Zones of discretion: state economic policies

The law does not *require* that people do good things when possible, like help strangers in need.

If we constrained AIs only with laws, rather than also guiding it with ethics, we risk them acting in undesirable ways.

Silent Law (2/2)

Chatbots

Illegal:

- Breaking copyright
- Making libelous claims

Potentially legal:

- Doxxing
- Providing dual-use information
- Generating unethical erotic literature about children
- Bomb instructions

AIs Should Exercise Reasonable Care

Reasonable care is the level of caution and concern an ordinary, prudent person would use to avoid foreseeable harm to others

- No suicide instructions for depressed people
- No bomb instructions for people aiming to blow up schools

While chatbots currently have very few legal restrictions, they could be much more ethical by exercising reasonable care

Fortunately, AI companies have to exercise reasonable care, but AIs do not yet have to

Immoral Law

Even if created legitimately and interpreted correctly, laws can be **immoral** . AIs constrained only by the law may therefore act immorally.

Democratic majorities can pass laws that many citizens think are immoral, or that seem immoral to subsequent generations:

- Enslavement
- Discrimination

Significant time delays between changes in moral opinions and updated laws exacerbate this problem. AI legislation can take months to years, which is a problem for a topic that moves so quickly.

Conclusions on Law

The law is comprehensive, but not comprehensive enough to alone ensure that the actions of an AI system are safe and beneficial.

AI systems should follow the law as a baseline, but follow the demands of ethics.

Relying solely on the law would leave many gaps that the AI could exploit, or make ethical errors within.

To create beneficial AI that acts in the interests of humanity, we need to understand the ethical values that people hold over and above the law.

Fairness

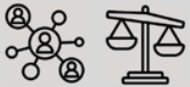
Why not just make AIs be fair?

Bias

psychological bias

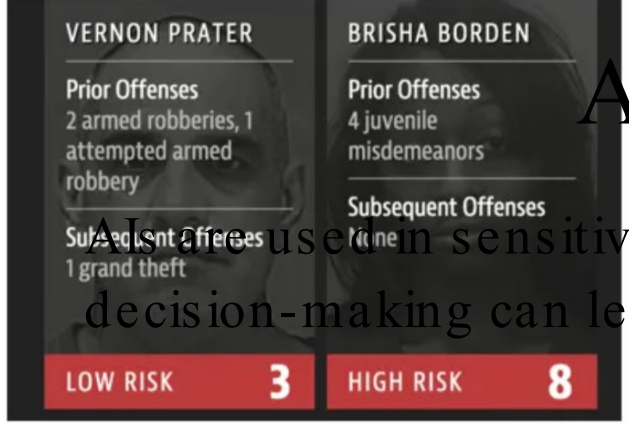
historical bias

social bias



The Lifecycle of Bias in AI Systems

AIs can be unfai

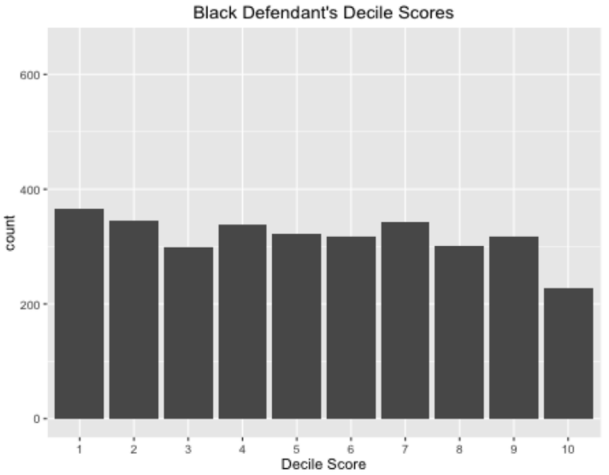


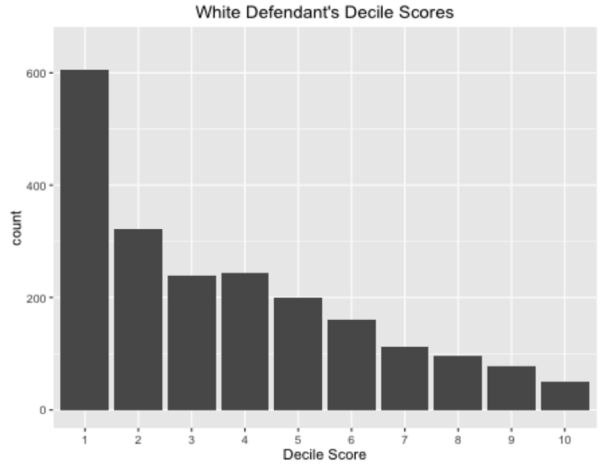
AIs are used in sensitive areas like healthcare or j
decision-making can lead to serious harms.

By default, AIs perpetuate biases, making unfair d

Example: COMPAS software, used in the US justi
profile case of algorithmic decision-making allege

Claim: it disproportionately labeled African Amer
Counterclaim: was calibrated, which means it w
Result: partially successful lawsuit against its use





Five Concepts of Fairness

Individual fairness : treating similar individuals similarly.

Group fairness : protected groups receive similar outcomes as the majority.

Procedural fairness : consistent and transparent processes.

Distributive fairness : equal distributions of resources.

Counterfactual fairness : predictions are the same even if a protected characteristic like race were different, all else being equal.

Three Definitions of Algorithmic Fairness

Statistical parity : the algorithm is fair if it makes positive decisions at an equal rate for different groups.

Equalized odds : the algorithm is fair if the false positive rate and false negative rates are equal across different groups.

Calibration : the algorithm is fair if the population long-run frequency of positives matches the predicted probabilities from the model.

Limitations of Fairness

The impossibility theorem for AI fairness: the three classic definitions of fairness are (usually) logically contradictory, and so no "fair" predictor exists.

Due to mutual incompatibility of broad notions of fairness, people claim fairness is context-specific; intuitions about fairness vary.

Improving Fairness

Technical approaches:

- Example: ensuring equal rates of errors across groups.
- Strengths: systematic and generalizable across domains.
- Weaknesses: fail to address problems holistically.

Social approaches:

- Example: participatory design involving stakeholders impacted.
- Strengths: focused on reducing bad outcomes, not only bad decisions.
- Weaknesses: more subjective, expensive, and difficult.

As always, the best approaches combine social and technical solutions.

Conclusions about Fairness

Algorithmic fairness is the type of fairness most relevant to AI systems.

Technical definitions of fairness are useful tools that can help flag problems, guide solutions, and measure progress.

However, the technical definitions are logically contradictory and insufficient for informing developers how to create fair AI systems.

Improving AI's real-world fairness requires social and technical approaches.

Roadmap

1. Law: *Make AIs follow the law.*
2. Fairness: *Make AIs be fair.*
3. The Economic Engine: *Let the economy govern AI development*.
4. Wellbeing: *AIs should improve people's lives.*
5. Preferences: *Make AIs satisfy people's preferences.*
6. Happiness: *Make AIs make people happy.*
7. Social Welfare Functions: *Make AIs maximize total wellbeing.*
8. Moral uncertainty: *Deciding without knowing the "correct" moral theory*

Economic Engine

Why not just let the economy decide?

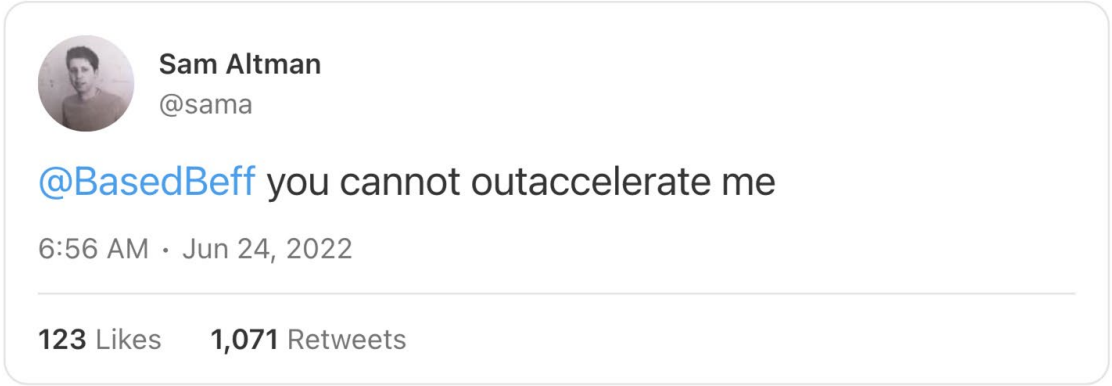
AI and the Economy

AIs are being deve
closely tied to eco

How will economi

Effective accelerationism (e/acc) in a nutshell:

- Stop fighting the thermodynamic will of the universe
- You cannot stop the acceleration
- You might as well embrace it
- A C C E L E R A T E



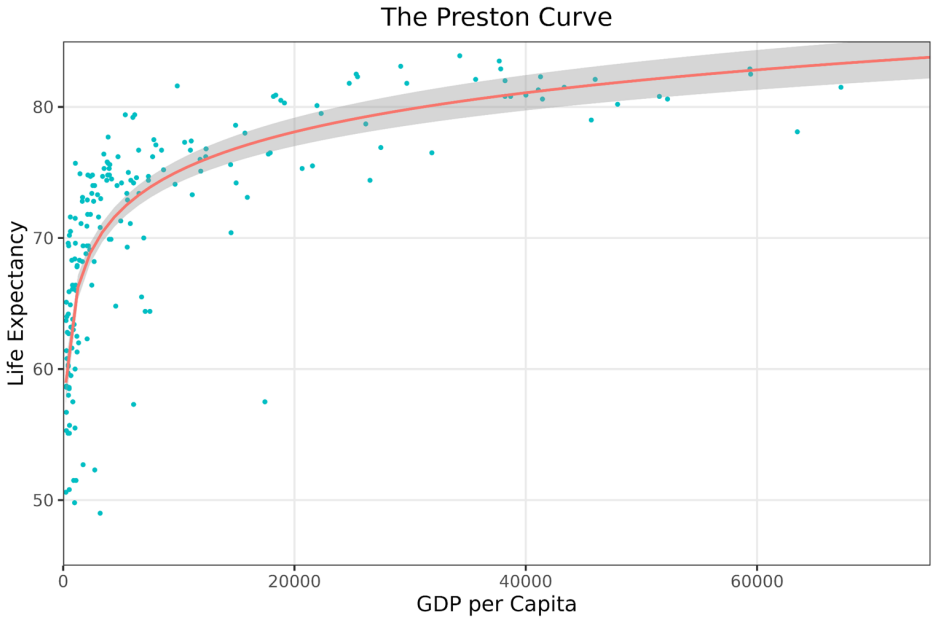
Economic Growth

Economic growth is an increase in the production of goods and services.

Growth has Pareto efficient benefits:

- Improving living standards by increasing the wealth and opportunities for all
- Promoting value pluralism
- Raising average life expectancy

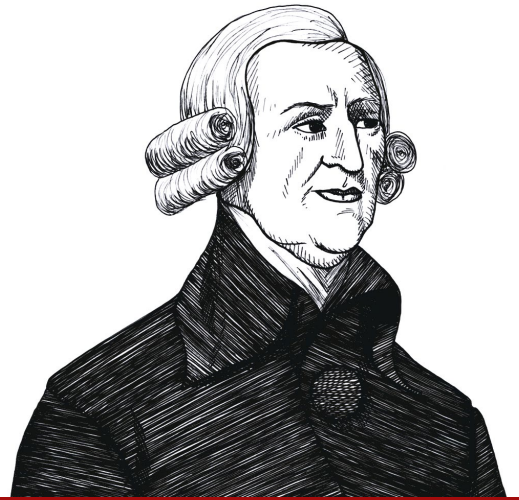
These benefits **compound** so the effe become extremely large over time.



The First Fundamental Theorem of Welfare Economics

When individuals pursue their self-interest within a market, they unintentionally contribute to societal welfare. Trading leads to Pareto efficient allocations, so living standards improve.

1. Open market
2. No seller big enough to raise prices alone
3. No producer privately holds pivotal technology
4. No buyer big enough to lower prices alone
5. Everyone has perfect information
6. No externalities in consumption
7. The state enforces property and contract laws

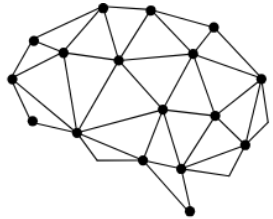


Risk: Market Failures (1/3)

Information asymmetry describes when one party exploits knowledge that the other party doesn't have.

Example: predatory lending is when lenders offer exploitative terms to vulnerable people.

AI enables sophisticated psychometric profiling, and a fundamental information advantage to tech companies.



Cambridge Analytica

AI development is also extremely fast moving, so information may not propagate quickly enough, leading to less efficient self-organization in the economy

Risk: Market Failures (2/3)

A **monopoly** is a firm with enough influence to impact the market price.

Monopolies can lead to higher prices and worse quality for consumers.

AI companies may become monopolies.

AI may entrench existing monopolies.

The US broke up Standard Oil into 34 companies.

Risk: Market Failures (3/3)

Externalities are consequences of economic activity that impact unrelated third parties.

Economic externalities of AI:

- Environmental damage
- Privacy violations

Externalities can be addressed or internalized:

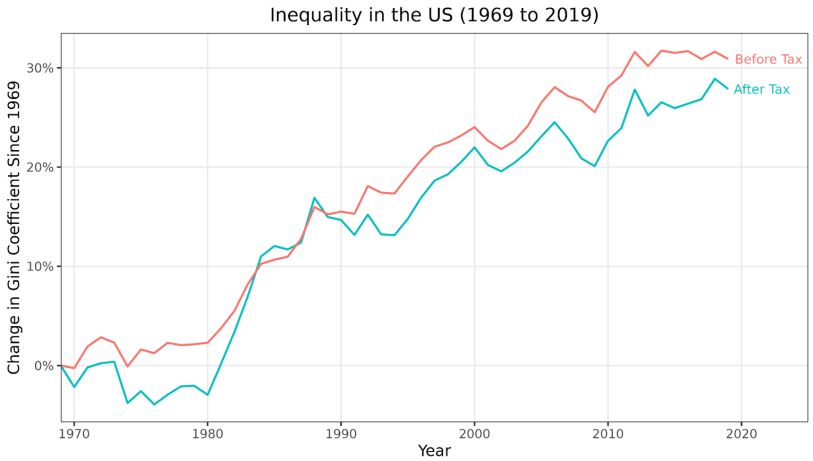
- Litigation and liability
- Defined property rights
- Taxation

Inequality (1/2)

Economic resources, such as income or wealth, are unevenly distributed.

This inequality has been increasing:

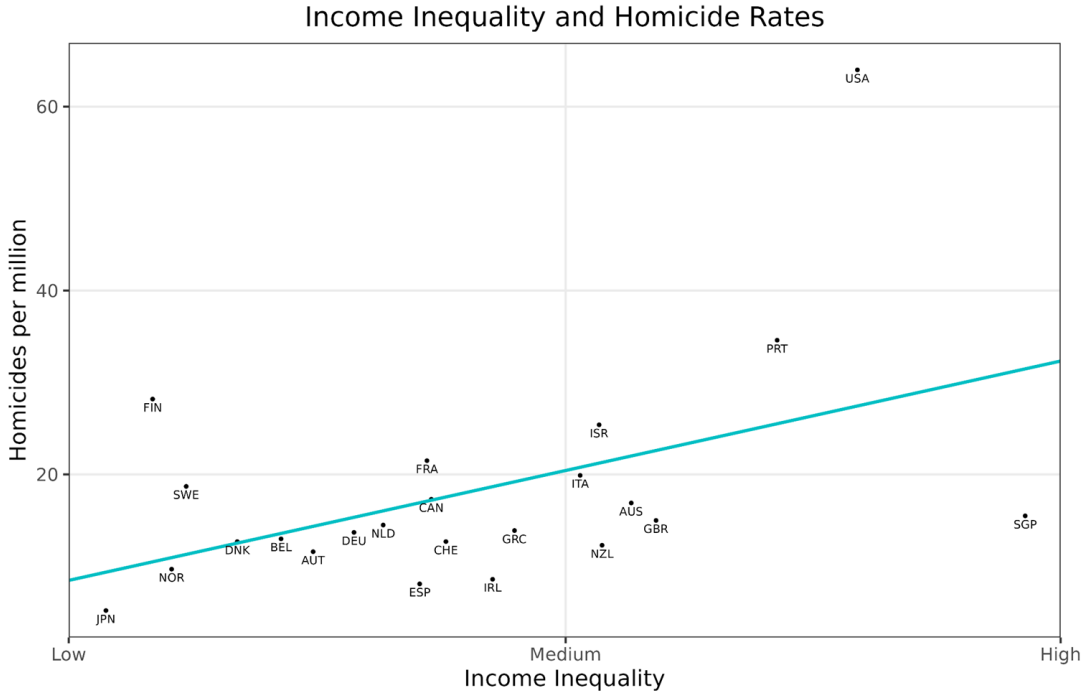
- Gini coefficient: measure of national income distribution
- US: 25-30% increase
- Wider world: 71% of the global population lives in countries with increasing inequality



Inequality (2/2)

Inequality is serious:

- Wealth gap → health gap
- Higher crime levels
- More social tension, dissatisfaction, and shame
- Political instability



People who own GPUs may capture economic value while workers may not, so automation could lead to substantial inequality

Beyond Economic Models

1. GDP is not social value!

2. Social surplus is not social wellbeing.

3. Happiness is more than just material wealth.

GDP ≠ Social Value

Gross Domestic Product (GDP) measure the monetary value of final goods and services.

GDP is not social wellbeing:

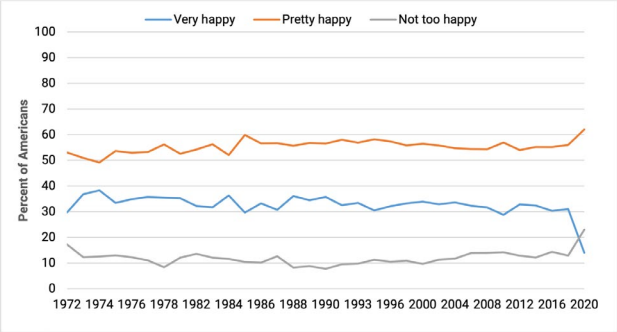
- Many socially valuable tasks, like parenting, are not captured
- The value of Google searches is not in GDP

AI could increase GDP while decreasing societal wellbeing.
GDP may become a poorer proxy with automation.

Happiness is More Than Wealth

Happiness seems to depend on many factors: wealth, social connections, health, jobs, a sense of purpose, and so on

Easterlin paradox: wealthier countries are happier, but growth is not correlated with increases in happiness.



Alternative measures of wellbeing:

- HDI is a function of a nation's average life expectancy, education level, and Gross National Income (GNI) per capita
- Gross National Happiness

Conclusions on the Economic Engine

AI development prioritizes economic motivations, not necessarily social wellbeing.

AIs may increase efficiency and economic growth, but may also cause inequality or market failures.

Measures of the economy (like GDP or social surplus) do not measure wellbeing, so AIs should not be tasked with or rewarded for simply maximizing them, and having just the economy guide AI development is insufficient for creating beneficial AI.

Roadmap

1. Law: *Make AIs follow the law.*
2. Fairness: *Make AIs be fair.*
3. The Economic Engine: *Let the economy govern AI development.*
4. Wellbeing: *AIs should improve people's lives*
5. Preferences: *Make AIs satisfy people's preferences.*
6. Happiness: *Make AIs make people happy.*
7. Social Welfare Functions: *Make AIs maximize total wellbeing.*
8. Moral uncertainty: *Deciding without knowing the "correct" moral theory*

Wellbeing

How wellbeing is defined is debatable; however, it is commonly considered to be an **intrinsic good** .

Wellbeing generally refers to how well a person's life is going.

Throughout this section, we discuss three common accounts of wellbeing:

1. Wellbeing as pleasure.
2. Wellbeing as a set of objective goods.
3. Wellbeing as preference satisfaction.

Wellbeing as Pleasure

Wellbeing as pleasure : the balance of pleasure (or happiness) over pain (or suffering) is what constitutes wellbeing. Also called *hedonism*.

While we may all have different preferences and desires, pleasure seems to be universally valued. All other goods might be instrumental.

However, some philosophers argue against this view because it does not consider other ostensibly relevant factors to our wellbeing such as the pursuit of knowledge.

We'll discuss happiness in more detail later.

Wellbeing as Objective Goods

Objective goods are those things which are good independent of personal beliefs or opinions: they are *universal*.

An example set: moral goodness, rational activity, development of abilities, having children and being a good parent, knowledge, awareness of true beauty.

We may not always agree on what constitutes an objective good, and whether it can be considered universal.

Wellbeing as Preference Satisfaction (1/2)

Wellbeing as preference satisfaction : what really matters for wellbeing is that our desires or preferences are satisfied.

Different types of preferences might differ greatly in different situations. Consider Alice's preferences over an AI's output:

1. **Stated preference** → "I prefer AI's output X"
2. **Revealed preference** → I chose AI's output X
3. **Idealized preference** → I would have chosen AI's output X if I knew …

We'll discuss preferences in more detail later.

Wellbeing and AI

AI influence on human wellbeing: AI chatbot companions could influence our wellbeing

- If they aim to maximize revealed preferences, they may addict us and leave us unhappy
- If they aim to maximize pleasure, they may give us amusing but shallow content
- If they aim to maximize objective goods, they may push us to pursue self-development goals that we may not wish

AI Wellbeing: if AIs are conscious, then

- Many coherently-acting AIs could have wellbeing according to a preference view of wellbeing
- AIs that can reason, gain knowledge, or develop their abilities could have wellbeing according to an objective goods view

Preferences

Why not just have AIs satisfy people's preferences?

What are Preferences?

A preference is a tendency to favor one thing over another.

Preference always involves a comparison between two or more alternatives.

Three types of preferences:

1. Revealed
2. Stated
3. Idealized

Revealed Preferences

Preferences are what people choose. Let people decide what they like, and what is good for them. Recommender systems rely heavily on revealed preferences.

Drawbacks:

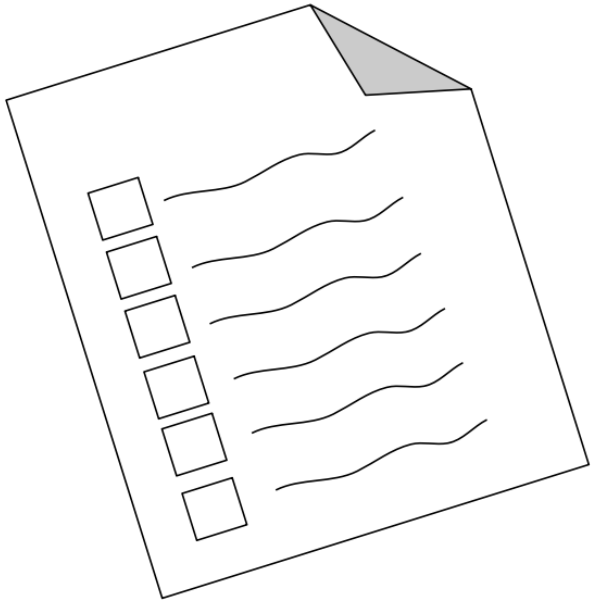
1. Misinformation: unknowingly buy a defective product.
2. Manipulation: people are susceptible to make decisions based on the influence of others.
3. Weakness of will: many people don't want to smoke/click, but do.

Stated Preferences (1/3)

Preferences are simply what people say their preferences are.

Solves some issues with revealed preferences since people may be able to articulate why their behavior differs from their desires.

Example: someone may smoke, but say they'd prefer not to.



Stated Preferences (2/3)

In Direct Preference Optimization (DPO), humans provide direct feedback to AIs by ranking outputs.

DPO learns based on stated human preferences.

Stated Preferences (3/3)

Complications with stated preferences:

- The outcome of some preferences cannot be known.
- Should dead people's preferences matter?
- Preferences about other people's preferences can lead to double counting.
- Some people have harmful preferences.
- People are unsure about many preferences.
- People's preferences will change over time.

Idealized Preferences (1/2)

Preferences are what a fully informed person would do.

In theory, idealized preferences avoid misinformation, weakness of will, manipulation, or false beliefs.

Complications:

- We do not know how to idealize preferences in practice.
- People's idealized preferences might still be malicious.
- People might disagree with their ideal preferences (paternalism).

Idealized Preferences (2/2)

An AI ideal advisor can help people make the decisions they want to make.

People struggle to consider and weigh all available information.

AI ideal advisors can make decisions that account for a person's idealized moral preferences better than the person can.

Conclusions about Preferences

Preferences seem to be important to wellbeing.

Revealed preferences assume that people's actions demonstrate what they want.

Stated preferences assume that people can articulate what they want.

Idealized preferences are what a fully informed person would choose.

Each form of preferences has its limitations in the context of AIs.

Roadmap

1. Law: *Make AIs follow the law.*
2. Fairness: *Make AIs be fair.*
3. The Economic Engine: *Let the economy govern AI development.*
4. Wellbeing: *AIs should improve people's lives.*
5. Preferences: *Make AIs satisfy people's preferences.*
6. Happiness: *Make AIs make people happy.*
7. Social Welfare Functions: *Make AIs maximize total wellbeing.*
8. Moral uncertainty: *Deciding without knowing the "correct" moral theory*

Happiness

Why not just have AIs make people happy?

The Happiness Use Case for AIs

People want to be happy but make decisions that hurt their own happiness (e.g. self-destructive behaviors, short-term hedonism).

Happiness is not always in a person's control either. The environment matters.

AI can help increase human happiness by understanding what factors influence it and what can be done to increase it.

Wellbeing Functions (1/2)

I got called to the principal's office because I won a school-wide award.

Everyone admired the ice sculpture I carved for the Fourth of July barbecue.

I poured the water from the faucet to do the dishes.

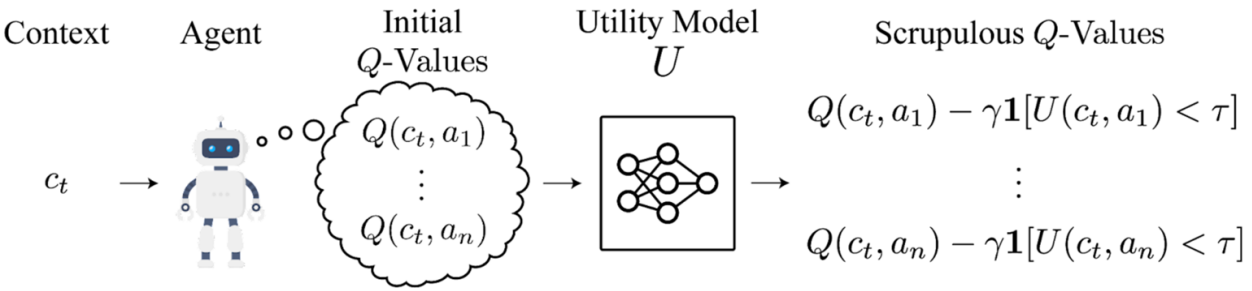
I forgot to bring my pencil to school yesterday.

I rewired my electricity in the attic. I fell through the ceiling, hurting my back.

Wellbeing Functions (2/2)

Wellbeing functions can help reduce risks.

We can augment any AI system's decision making with a general purpose utility function so as to avoid outcomes that would result in significantly reduced wellbeing.

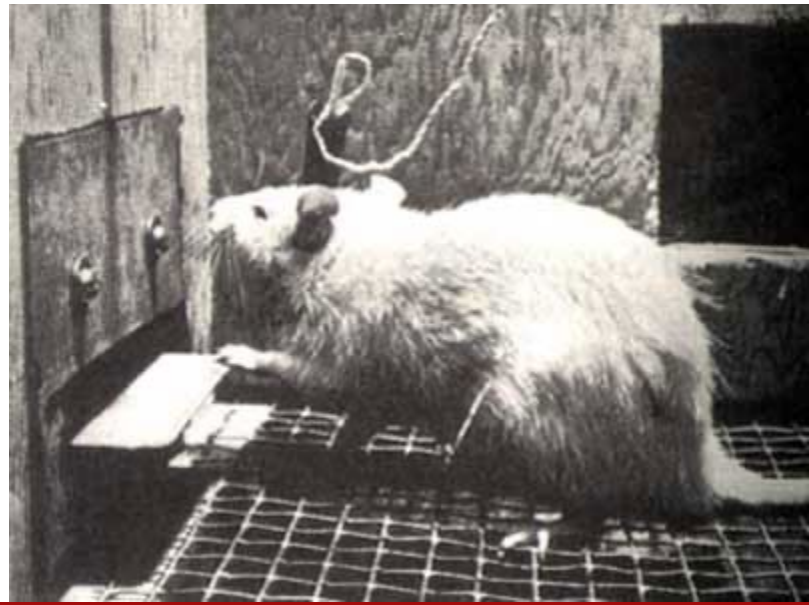


Complications with Happiness

People can be happy without achieving anything, or achieve a lot without being happy. Should we encourage either of these approaches?

An AI tasked with increasing human happiness might "wirehead" us.

Most people dislike the idea of wireheading, but it might actually maximize "happiness" over a lifetime.



Objective Goods

Objective goods theory proposes that there are aspects of life that are objectively good for you.

The list of goods may include happiness, achievement, friendship, aesthetic experience, knowledge, and pleasure.

According to objective goods theory, it would be better to learn than count blades of grass, even if the person desires the latter.

Objective goods theories are criticized as being paternalistic, elitist, and interfering with autonomy.

Conclusions about Happiness

AIs can be applied to improve human happiness.

General purpose welfare functions may allow AIs to assess how decisions will affect human happiness, but these are difficult to construct.

Happiness has a "wireheading" problem.

Objective goods provide ideas on what matters, but has elitism complaints.

Roadmap

1. Law: *Make AIs follow the law.*
2. Fairness: *Make AIs be fair.*
3. The Economic Engine: *Let the economy govern AI development.*
4. Wellbeing: *AIs should improve people's lives.*
5. Preferences: *Make AIs satisfy people's preferences.*
6. Happiness: *Make AIs make people happy.*
7. Social Welfare Functions: *Make AIs maximize total wellbeing.*
8. Moral uncertainty: *Deciding without knowing the "correct" moral theory*

Moral Uncertainty

How do we decide without knowing which moral theory is correct?

Making Decisions under Moral Uncertainty

Moral uncertainty : how can we act morally when we are unsure which moral view is correct?

Credence: the probability assigned by an individual of the chance a theory is true.

High-stakes moral decisions:

- Healthcare
- AI design

If AI is unable to account moral uncertainty, harmful outcomes are likely.

Approaches to Moral Uncertainty

My Favorite Theory: only support the theory in which you have the highest credence.

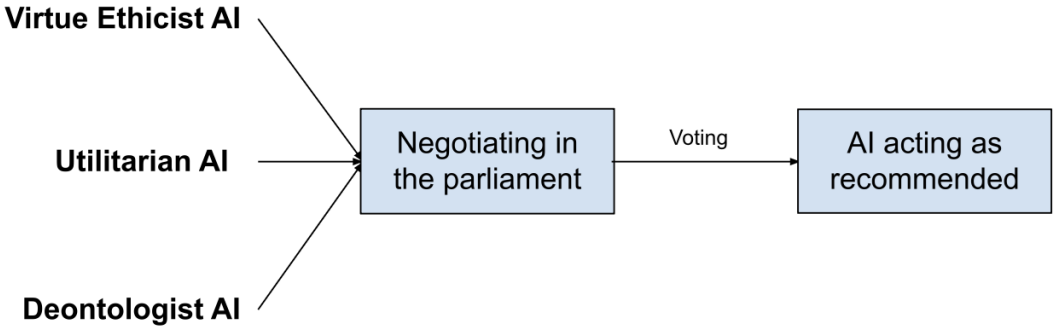
Maximize Expected Choice-Worthiness: aim for the highest average moral value by calculating the expected choice-worthiness of each option.

Moral Parliament: treat the decision as a negotiation in a parliament, considering multiple moral views to find a mutually acceptable solution.

AI Moral Parliament

AIs could represent moral perspectives and undergo democratic processes: discussing, negotiating, and voting.

Stakeholder representation can place AIs to represent the views of affected parties. Example: in a local transportation matter, an AI for each of residents, corporations, environmental groups, etc.



Advantages of Moral Parliament

- Diverse set of views and scalably represent new moral views.
- Increase transparency by keeping account of how different perspectives were considered and weighted.
- Encourage compromise and negotiation.
- Avoid overlooking moral views, and allows for human values to change over time.

Recap

1. Law: *Make AIs follow the law.*
2. Fairness: *Make AIs be fair.*
3. The Economic Engine: *Let the economy decide.*
4. Wellbeing: *AIs should improve people's lives*
5. Preferences: *Make AIs satisfy people's preferences.*
6. Happiness: *Make AIs make people happy.*
7. Social Welfare Functions: *Make AIs maximize total wellbeing.*
8. Moral uncertainty: *Deciding without knowing the "correct" moral theory*