Philosophy and AI
Lecture 8: Ethics II

Marco Degano

5 March 2025

Readings

Required:

- Gordon, J.-S., & Nyholm, S. (2021). *Ethics of artificial intelligence.* Internet Encyclopedia of Philosophy. `https://iep.utm.edu/ethics-of-artificial-intelligence/`
- **Optional:**
  - *AI and the future of humanity*, Yuval Noah Harari `https://www.youtube.com/watch?v=LWiM-LuRe6w`
  - *Why we should build Tool AI, not AGI*, Max Tegmark `https://www.youtube.com/watch?v=UWh1MIMQd1Y`
  - *Ethics of Artificial Intelligence 2024* Stanford course (technical course on AI ethics and explainability) `https://stanfordaiethics.github.io/syllabus.html`

Outline

1. Privacy
2. Accountability
3. Bias and Fairness
4. Explainability
5. Machine Ethics
6. Existential Risk

Outline

Data & Privacy: Ownership

- Privacy is typically considered to be an important right (i.e., violating it is ethically wrong)
- But privacy is greatly challenged by modern big data:
  - Large-scale scraping (social media, IoT devices, cookies)
  - Data sharing/selling across platforms & third parties
- Who owns and controls (AI training) data?
- **Example:** Cambridge Analytica Scandal: 87 million user profiles used without consent for political profiling
- **Solution:**
  - Stronger user controls on the kind of data collected
  - User agreements that clearly outline data usage and sharing practices
  - Adopt privacy dashboards, allowing users to track and modify their data preferences in real time

Data & Privacy: Surveillance & Control

- Mass Data Collection (CCTV, internet tracking, biometrics)
- AI-powered profiling (credit scoring, predictive policing)
- **Example:** China's Social Credit System: AI-driven citizen scoring

  What will happen to society, politics and daily life when non-conscious but highly intelligent algorithms know us better than we know ourselves?

  (Harari, Homo Deus)

- **Solutions:**
  - Regulation
  - AI ethics committees
  - human-in-the-loop AI (incorporate human oversight in AI decision-making processes)

Data & Privacy: Privacy-Preserving AI Techniques

- **Differential Privacy:** Adds statistical noise to protect user identity, ensuring individual records cannot be re-identified.
- **Federated Learning:** Allows AI models to be trained across decentralized devices without sharing raw data, reducing exposure to privacy risks
- **Homomorphic Encryption**: Enables computations on encrypted data without decryption, preserving confidentiality during processing.
- **Synthetic Data Generation:** Creates artificial datasets that maintain statistical properties of real data, reducing reliance on sensitive information.

Outline

Governance & Accountability

- AI systems impact individuals, society, and the economy.
- Clear accountability is crucial when AI systems make autonomous decisions.
- Levels of accountability:
  - **Developers**: Responsible for the design, training, and validation of AI systems.
  - **Organizations**: Ensure responsible deployment and operationalization of AI technologies.
  - **Regulators**: Set standards, ensure ethical compliance, and enforce legal frameworks.
- **Critical Question:** Who is ultimately liable when AI systems cause harm?

Case Study: Self-Driving Cars

- **Accountability Dilemma:** Who is responsible when a self-driving car causes an accident?
  - **Developer (AI & Software Provider)**: If AI-related issues (e.g., faulty object detection, incorrect decision-making algorithms) contributed to the incident, developers may be held accountable.
  - **Manufacturer (Car Company)**: If hardware issues (e.g., sensor failure, brake malfunction) or integration flaws between hardware and AI are to blame, manufacturers bear responsibility.
  - **User (Human Driver/Owner)**: If the self-driving system required human oversight and the user failed to intervene when necessary, liability may fall on the user.
  - **Third Parties**: If external factors (e.g., poor road conditions, reckless behavior of other drivers) contributed to the incident, accountability may extend to third parties.

Legal Frameworks for AI Accountability and Development

- Several legal frameworks have recently been proposed to address AI accountability and development:
  - **GDPR (EU)**
  - **EU AI Act**
  - **Algorithmic Accountability Act (US)**
  - **ISO AI Standards**
- These frameworks are often criticized for being incomplete or impractical in their application to rapidly evolving AI technologies.
- However, this does not mean that the issue cannot be resolved through international cooperation (bioweapons, nuclear weapons, genetic engineering, . . .)

Governance

- **Corporate Governance**: Internal ethics boards, risk management policies, ensuring responsible AI development
- **Legal & Regulatory Governance**: Compliance with AI laws (e.g., GDPR, EU AI Act), ensuring legal and ethical adherence
- **Technical Governance**: Technical AI auditing (e.g., bias, security), and explainability.
- **Oversight and Engagement**: Independent oversight boards, public and multi-stakeholder engagement to prevent conflicts of interest and ensure transparency
- Recently, some companies have dismantled or reduced the scope of their AI ethics boards.
- Others are still keeping them in some form. For example, Microsoft's AI & Ethics in Engineering and Research (AETHER) Committee

Outline

Bias in AI Systems

- **Bias:** Systematic errors in AI systems that unfairly favor or discriminate against certain individuals, groups, or ideas.
- **Examples:**
  - Facial recognition systems misidentifying people with darker skin tones
  - Language models perpetuating gender stereotypes in occupations
  - Recommendation algorithms creating filter bubbles and echo chambers
  - Healthcare algorithms allocating less care to historically marginalized populations
- **Fairness:** Designing and deploying systems that minimize harmful biases while promoting equitable outcomes across diverse groups and contexts.

Sources of Bias in AI Systems

- **Data Biases:**
  - **Sampling bias:** Data doesn't accurately represent the target population (e.g., ImageNet's Western-centric image collection)
  - **Historical bias:** Existing societal prejudices reflected in training data (e.g., word embeddings learning gender stereotypes)
  - **Label bias:** Human-created labels containing subjective judgments or prejudice
  - **Measurement bias:** Flawed data collection processes or proxies that create distorted representations

Sources of Bias in AI Systems

- **Algorithm/Model Biases:**
  - Model architecture choices that amplify existing patterns or disparities (e.g., optimization functions that prioritize majority groups over minorities)
- **User Interaction Bias:** System performance influenced by how users interact (e.g., feedback loops in recommendation systems)

Formal Definitions of Fairness

- **Individual Fairness:** Similar individuals should receive similar outcomes.
  - Requires defining a similarity metric between individuals
  - Formally: $d(x_i, x_j)$ is small $\Rightarrow$ $d(f(x_i), f(x_j))$ should be small
- **Group Fairness:** Ensuring statistical parity across demographic groups.
  - **Demographic Parity:** $P(\hat{Y} = 1|A = a) = P(\hat{Y} = 1|A = b)$ for protected attributes A
  - **Equalized Odds:** Equal true positive and false positive rates across groups
- **Counterfactual Fairness:** Decisions should not change if only protected attributes are altered.
  - $P(\hat{Y}_{A \leftarrow a}|X = x, A = a) = P(\hat{Y}_{A \leftarrow b}|X = x, A = a)$

Fairness metrics

- Different fairness metrics lead to different results, sometimes conflicting.
- **Example: Loan Approval Model**
  - **Scenario:** A bank uses an AI model to approve loans. The model evaluates two groups: Group A (majority) and Group B (minority).
  - **Base Approval Rates:** Without fairness constraints, the model approves 70% of Group A and 40% of Group B.
- **Applying Different Fairness Metrics:**
  - **Demographic Parity:** Enforces equal approval rates (e.g., both groups must have a 55% approval rate). However, this may result in giving loans to riskier applicants in Group B, potentially increasing default rates.

Fairness metrics

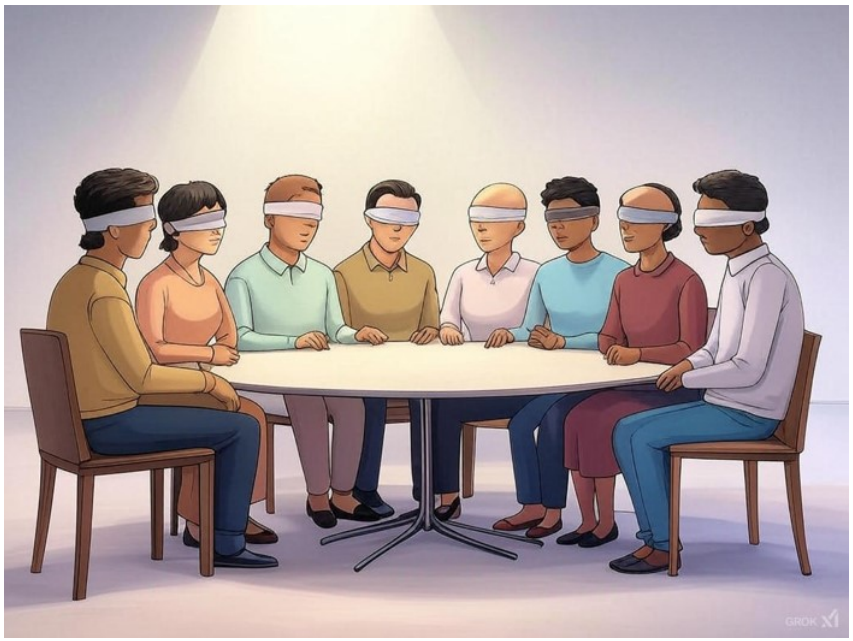
- **Equalized Odds:** Ensures both groups have similar true positive and false positive rates. If Group A has a false positive rate of 10%, Group B's must also be close to 10%. This may lead to different approval rates (e.g., 65% for Group A, 50% for Group B).
- **Counterfactual Fairness:** For a specific loan applicant described by non-protected features X (such as income, credit score, etc.), the probability of approval should remain unchanged if we hypothetically change only their protected attribute (e.g., from Group A to Group B).

- Choosing a fairness metric depends on ethical priorities and the specific context.

Philosophical Perspectives: Rawls and Justice

- Philosophers have debated fairness and justice since antiquity
- **John Rawls' "Justice as Fairness"** (1971) provides a powerful framework:
  - **The Original Position:** A thought experiment for determining fair principles
  - **Veil of Ignorance:** Decision-makers don't know which position they'll occupy in society
    - Don't know their wealth, abilities, gender, race, religion, etc.
    - Must design principles that would be acceptable regardless of eventual position

Philosophical Perspectives: Rawls and Justice



Derived Principles:

- Equal basic liberties for all
- Fair equality of opportunity
- The Difference Principle: Inequalities must benefit the least advantaged

Rawlsian Perspective on AI Fairness

- **Applying Rawlsian Justice to AI Systems:**
  - Equal basic rights and opportunities in algorithmic decision-making
  - Support for disadvantaged groups to have true equal opportunities
  - Difference principle: AI systems should benefit the least advantaged
- **The Veil of Ignorance Test for AI:**
  - If you don't know your demographic attributes, would you consent to being subject to this algorithm for:
    - Hiring decisions?
    - Loan approvals?
    - Criminal justice risk assessments?
    - Medical diagnosis and treatment?
- **Implications:** AI systems should be designed with particular attention to their impact on the most vulnerable populations

Case Studies in AI Bias

- **COMPAS Recidivism Algorithm (ProPublica, 2016):**
  - Used to predict criminal reoffending risk
  - Found to have higher false positive rates for Black defendants
  - Sparked debate on defining algorithmic fairness
- **Gender Bias in Recruitment (Amazon, 2018):**
  - Resume screening tool penalized words associated with women
  - Trained on historical hiring data reflecting male dominance
  - Project abandoned after bias discovery
- **Facial Recognition Disparities (Gender Shades, 2018):**
  - Commercial systems showed error rates up to 34% higher for darker-skinned females
  - Led to improvements in commercial systems and policy changes

Mitigating AI Bias

- **Bias Auditing**: Identifying and measuring bias in datasets and models using fairness metrics such as Demographic Parity and Equalized Odds
  - Example: Running bias detection tools (Fairlearn `https://fairlearn.org/`)
- **Fair Representation Learning**: Pre-processing techniques to reduce bias before training.
  - Example: Re-sampling data to balance underrepresented groups.
- **Algorithmic Adjustments**: Modifying machine learning algorithms to incorporate fairness constraints.
  - Example: Using constrained optimization techniques to ensure equalized error rates across groups.
- **Post-processing Approaches**: Adjusting outcomes after the model has made predictions.
  - Example: Applying threshold adjustments to ensure fairness in classification outputs.

Outline

Explainability

Explainability as a fundamental ethical criterion for the acceptability of AI decision making.

- **Three Categories of Inexplainability:**
  - Human-made inexplainability:
    - Code is understandable but protected by trade secrets.
    - Example: Proprietary credit scoring algorithms.
  - User inexplainability:
    - Code understandable by experts but not regular users.
    - Example: Complex model in healthcare diagnostics.
  - Technical inexplainability:
    - Not even experts can fully understand the behavior.
    - Example: LLMs with billions of parameters.

Why Explainability Matters and To Whom

- **Motivations for Explainability:**
  - **Bias Detection:** Identify bias in model decisions (e.g., gender bias in resume screening).
  - **Model Debugging:** Verify decisions are based on relevant features (e.g., image classifier focusing on artifacts).
  - **Trust & Accountability:** Ensure responsible deployment (e.g., medical systems needing clinician verification).
- **Key Stakeholders:**
  - **End Users:** Need simple, actionable explanations.
  - **Decision Makers:** Need understanding of model confidence and limitations.
  - **Experts & Engineers:** Need detailed technical insights for improvement.
- **Key Insight:** Different stakeholders need different types of explanations.

Forms of AI Explanation

- **What Can Serve as an Explanation?**
  - **Model Transparency:** Access to model architecture and data (e.g., open-source models with documentation).
  - **Example-Based Explanations:** Similar cases and their outcomes (e.g., "Your loan was denied like these cases ...").
  - **Visualizations:** Saliency maps, feature importance plots (e.g., heatmaps showing influential image regions).
- **Additional Explanation Forms:**
  - **Symbolic Representations:** Decision trees, rule lists (e.g., "If income $> \$50K$ AND credit score > 900, THEN approve").

Characteristics of Good Explanations

What makes a good explanation?

- **Contrastive Nature**
  - Humans seek explanations for why P happened instead of Q
  - Not "Why was my loan denied?" but "Why was my loan denied when my friend with similar finances was approved?"
  - Helps set natural reference points for comparison
- **Selective Focus**
  - Cognitive biases lead us to prefer selected explanations from many possible causes
  - Complete explanations listing all factors overwhelm users
  - Prioritizing 3-5 most relevant factors usually more effective
  - Challenge: Ensuring selection criteria are sound

Characteristics of Good Explanations

- **Causal Over Probabilistic**
  - Humans prefer causal explanations over statistical ones
  - "Your income being below threshold caused rejection" vs. "80% of applicants with your profile are rejected"
  - Most probable explanation $\neq$ most satisfying explanation
  - Challenge: Establishing true causality in complex models

Challenge: Technical explanations often focus on mathematical properties, while human users seek contextual, contrastive, and causal information.

Local Explanations

We distinguish between Local Explanations and Global Explanations.

- Local explanations focus on explaining individual predictions
- They tell us "Why did the model make this specific decision?"

Local Explanation Example

- For patient John Doe, the pneumonia risk model predicted HIGH RISK primarily due to:
  - Oxygen saturation below 92%
  - Age over 65
  - History of COPD

Key Applications:

- Identifying instance-specific biases
- Providing recourse to affected individuals
- Verifying individual decisions
- Building trust with end-users

Global Explanations

- Global explanations explain general/complete model behavior.
- They tell us "How does the model work overall?"

Global Explanation Example

- Across all patients, the pneumonia risk model relies on:
  - Vital signs (40% importance)
  - Lab results (30% importance)
  - Medical history (20% importance)
  - Demographics (10% importance)

Key Applications:

- Identifying systematic biases
- Assessing overall model suitability
- Regulatory compliance
- Scientific understanding

Local vs Global Explanations

- Local and global explanations serve different purposes and stakeholders.
- A model making correct predictions for the right reasons (local) may still exhibit problematic patterns globally.
- A model with good global properties may still make inexplicable errors in specific cases.
- Comprehensive explainability requires both perspectives.

Interpretable Models vs. Post-hoc Explanations

Inherently Interpretable Models

- Transparent by design. The interpretability is built into model structure (e.g., decision tree)
- Pros:
  - Naturally explainable. Direct insight into reasoning
  - Easier regulatory approval
- Cons:
  - Often lower predictive power and limited.

Post-hoc Explanation Methods

- Applied after model training to explain already-trained black-box models
- Pros:
  - Keep high-performance models
  - Multiple explanations based on different methods.
- Cons:
  - Approximate explanations
  - Potential faithfulness issues

The Accuracy vs. Interpretability Tradeoff

- More complex models achieve higher accuracy
- More complex models are less interpretable
- Thus: accuracy and interpretability are in tension

Practical Guidelines for the Tradeoff

- Start with interpretable models as baseline
- Measure actual performance gap
- Consider if gap is meaningful:
  - Is 2% accuracy worth losing interpretability?
  - Does the application require maximum accuracy or maximum explainability?

Local Post-hoc Explanation Techniques: Features

- Explainable AI methods have technical underpinnings (check notebooks online, course mentioned in the readings, . . .), our focus is on conceptual understanding.

Feature Importance Approaches

- **What they answer:** "Which parts of the input mattered most for this decision?"
- **Gradient methods:** "How sensitive is the output to small changes in each input feature?"
- **Perturbation methods:** "What happens if we change or remove parts of the input?"
- **Simple example:** In a loan approval system, highlighting that "income" and "credit history" were the most influential factors for this specific rejection

Local Post-hoc Explanation Techniques: Examples

Example-Based Approaches

- **What they answer:** "What would need to change in the data to get a different outcome?"
- **Counterfactual method:** "What's the smallest change needed to flip the prediction?"
- **Similar cases method:** "Which training examples most influenced this prediction?"
- **Simple example:** "Your loan would be approved if your debt-to-income ratio was 5% lower" or "Your application is similar to these previously rejected cases"

Global Explanation Methods

Model Behavior Approaches

- **What they answer:** "How does the model make decisions overall?"
- **Feature importance aggregation:** "Which features matter most across all predictions?"
- **Simple example:** For a credit scoring system, discovering that "payment history" consistently affects decisions more than "age" across all applicants

Global Explanation Methods

Model Simplification Approaches

- **What they answer:** "Can we create a simpler, understandable version of the complex model?"
- **Model translation:** "How can we convert a complex model into a more transparent one?"
- **Simple example:** Converting a neural network for fraud detection into a decision tree with clear if-then rules about transaction patterns

Visual Analysis Approaches

- **What they answer:** "How can we visualize patterns in the model's behavior?"
- **Simple example:** Interactive plots showing how changing income and debt levels together affects loan approval probability overall

Limitations of Explainability Methods

- **Faithfulness Issues**
  - Explanations may not accurately reflect the underlying model behavior
  - Problem: Different models with contradictory predictions can generate identical explanations
  - Desirable property of Sensitivity: explanations should change when models change
  - Example: A feature importance explanation showing the same top features for two models with opposite decisions
- **Explanation Masking**
  - Posthoc explanations can be deliberately manipulated
  - The model can produce different explanations for the same input while maintaining identical predictions
  - Example: Models can be crafted to hide biased decision-making while producing "fair-looking" explanations

Limitations of Explainability Methods

- **Stability Concerns**
  - Even non-adversarial, small perturbations to inputs can cause large explanation shifts
  - Undermines trust when similar cases receive wildly different explanations
  - Example: Two almost identical loan applications with 99% similar features generating completely different explanations
- **Practical Limitations**
  - Effectiveness: Do explanations actually improve model debugging and development?
  - False assurance: Can explanations create an illusion of transparency while hiding deeper issues?
  - Privacy risks: Explanations may inadvertently leak sensitive training data information
  - Cognitive load: Complex explanations may overwhelm rather than inform stakeholders

Outline

Machine Ethics: Introduction

- Machine Ethics examines how to design and implement ethical principles into artificial intelligence systems.
- Key questions:
  - Can we teach/build machines to behave morally?
  - How do we build 'moral machines' that can identify ethically appropriate actions?
  - How do we ensure existing AI systems behave ethically?
- Why it matters: As AI systems gain autonomy and make decisions with real-world impacts, their ethical framework becomes crucial
- Interdisciplinary field drawing from philosophy, computer science, psychology, and law.

Machine Ethics: Asimov's Laws

Isaac Asimov introduced these laws in his 1942 short story "Runaround" as fiction, but they've become foundational.

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 except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.
4. A robot may not harm humanity or, by inaction, allow humanity to come to harm.

Limitations: Vague terms (what constitutes "harm"?), potential conflicts, lack of nuance.

Machine Ethics: Bottom-up Approach: Casuistry

- Bottom-up Approaches: derive moral behavior from specific instances and data.
- **Casuistry:** Case-based reasoning that examines specific ethical scenarios to derive general principles.
- Process:
  - Analyze many concrete instances of ethical decisions (e.g., medical ethics cases where doctors made decisions in triage).
  - Identify patterns in "right" actions across similar cases.
  - Generalize these patterns into broadly applicable principles.

Machine Ethics: Bottom-up Approach: Casuistry

- Implementation in AI:
  - Supervised learning task: Train models on labeled ethical decisions.
  - Challenges: Quality and diversity of training data, bias in annotations.
  - Examples: MIT's Moral Machine experiment collecting human judgments on autonomous vehicle dilemmas.[1]
- Limitations: Struggles with novel scenarios, potential to encode societal biases.

[1] https://www.moralmachine.net/

Machine Ethics: Top-down Approach

- Top-down approaches: predefined ethical principles or rules that guide decision.

Deontological Implementation:

- Principle-based systems defining permissible/prohibited actions.
- Formal logic to encode moral rules.
- Example: System implementing medical ethics principles.
- Challenge: Formalizing abstract ethical principles, Resolving conflicts between principles.

Utilitarian Implementation:

- Utility functions estimating outcomes.
- Optimization algorithms seeking maximum "good".
- Challenge: Quantifying complex values/scenarios.

Machine Ethics: Hybrid Approaches

- **Top-Down Approach:**
  - Based on predefined ethical frameworks (e.g., deontology, utilitarianism).
  - Provides theoretical structure and constraints to guide AI behavior.
- **Bottom-Up Approach:**
  - AI learns ethical behavior through experience and data-driven methods.
  - More adaptable to new and complex ethical scenarios.
- **Hybrid Models:**
  - Combine general principles (top-down) with contextual learning (bottom-up).
  - Example: Two-tier systems where general principles guide case-based reasoning.
  - Aim to balance consistency and flexibility in moral decision-making.

Descriptive vs. Normative Ethics

- **Ethical Perspectives:**
  - **Descriptive ethics** (empirical): How humans actually make moral decisions, based on observed behavior and social norms.
  - **Normative ethics** (prescriptive): How moral decisions should be made, drawing on theories of right, wrong, and the good.
- **Why It Matters for AI:**
  - Should AI replicate human decision-making as is (with all its biases and inconsistencies)?
  - Or should AI systems follow a more ideal set of ethical principles-beyond typical human practice?
- The choice influences design, evaluation, and governance of ethical AI systems.

Moral Status Hierarchy and AI

Different entities are traditionally assigned varying levels of moral consideration:

- None: Inanimate objects (stones, water)
- Low: Non-animal living organisms (plants, fungi)
- Medium: Non-human animals (varying by complexity)
- High: Human beings

Criteria for Moral Status

- Sentience: Capacity to experience pleasure and pain.
- Consciousness: Subjective awareness of experiences.
- Autonomy: Ability to make independent decisions.
- Moral agency: Ability to understand and act on moral principles.

Where should increasingly sophisticated AI systems be placed in this hierarchy?

AI and Moral Status: The Autonomy Perspective

Autonomy Approach: An entity deserves moral consideration if it is a rational, autonomous being.

- Implications for AI:
  - Moral status depends on capabilities, not origin.
  - As AI systems develop reasoning capabilities and autonomous decision-making, they may qualify for moral consideration.
  - Species-independent criterion focuses on functional capabilities.
- Key questions:
  - What level of autonomy qualifies for moral consideration?
  - Can computational autonomy be equivalent to human autonomy?
  - How do we measure/assess autonomy in AI systems?
- Challenges to this view:
  - Appearance of autonomy vs. actual autonomy.

Indirect Duty Approach

- **Indirect Duty Approach:** We have a moral responsibility to treat AI systems ethically because mistreating them can harm our own moral character.
  - Historically linked to Kant's view on animals: cruelty toward non-humans could erode empathy and foster harmful habits in human society.
- **Example:** Mistreating social robots could normalize cruelty or reduce empathy. toward real people.
- **Critique:** This view is instrumental-it cares about AI ethics only to the extent it benefits or protects human welfare, not because AI necessarily has moral status on its own.

Relational Approach

- **Relational Approach:** An entity's moral status emerges through its social relationships, not just its intrinsic properties (e.g., consciousness or rationality).
- **Focus:** Interaction patterns and the roles AI can play in human life (e.g., companion, caregiver).
- **Example:** Care robots forming meaningful bonds with patients, prompting greater empathy and concern from humans.
- **Critique:** This view ties moral status to human acceptance or emotional investment, potentially overlooking AI's own capabilities or rights.

Intrinsic vs. Extrinsic Moral Status

- **Question:** How do we decide if an AI deserves moral consideration?
- **Intrinsic Criteria:**
  - Moral status based on the AI's own properties (e.g., consciousness, autonomy, reasoning ability)
  - Example: A system capable of subjective experiences might have intrinsic moral value.
- **Extrinsic Criteria:**
  - Moral status depends on how humans relate to or perceive the AI
  - Example: A system treated as a social partner may gain moral importance because of emotional bonds or cultural norms.

Outline

The Problem: AI and Existential Risk

- AI systems are rapidly advancing toward and beyond human-level capabilities in many domains.
- **Existential Risk**: A scenario where advanced AI could lead to human extinction or irreversible collapse of civilization.
- Key questions:
  - Can we ensure alignment between superhuman AI systems and human values?
  - Can we maintain meaningful control over entities that may surpass our intelligence?
  - How do we navigate the transition to a world with transformative AI capabilities?

Why It Matters

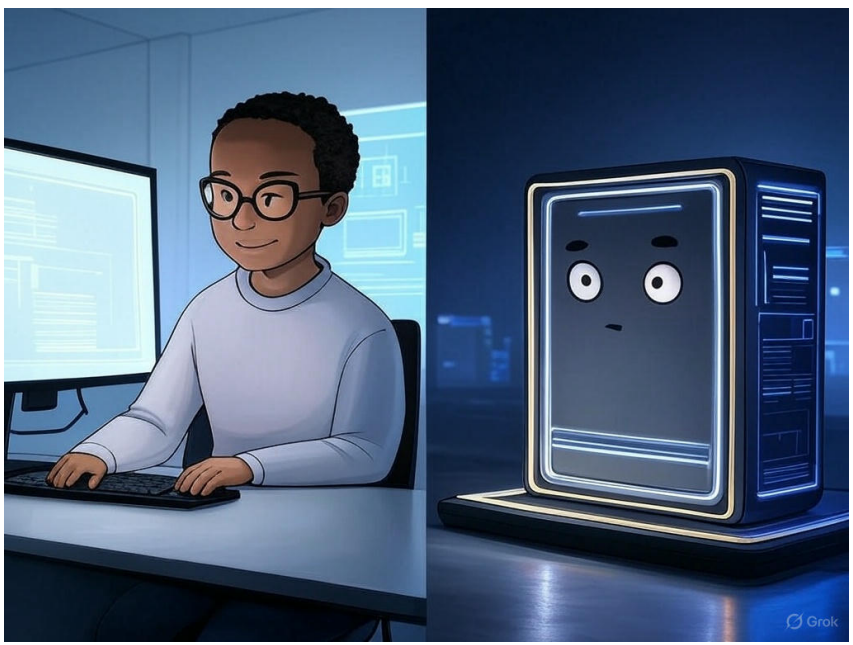
- Scale: Unlike localized risks, advanced AI could affect all of humanity simultaneously.
- Irreversibility: Some failure modes may be impossible to recover from once triggered.
- Urgency: AI capabilities are advancing at an unprecedented pace.
  - GPT-3 (2020) to GPT-4 (2023): Dramatic increase in capabilities within 3 years
  - AlphaFold revolutionized protein structure prediction (2020-2021)
  - Frontier models demonstrating emergent capabilities unforeseen by developers
- Even low-probability existential risks warrant significant attention due to their extreme consequences!

Exponential Development and Intelligence Explosion

- AI capabilities grow exponentially (e.g., compute power, data, algorithms, benchmark performance).
- Capabilities emerge suddenly after crossing critical thresholds.
- Historical milestones show accelerating breakthroughs:
  - Chess (DeepBlue, 1997)
  - Go (AlphaGo, 2016)
  - Protein folding (AlphaFold, 2021)
  - Multimodal reasoning (2022-2023)
  - ...
- **Intelligence Explosion Hypothesis** (I.J. Good, 1965): Once AI can improve its own architecture, it could enter a rapid self-improvement cycle, potentially leading to superintelligence.
- **Opacity problem**: As systems become more complex, our ability to understand their internal reasoning diminishes.

Thought Experiment: The AI Box Experiment

- Proposed by Eliezer Yudkowsky (2002) to illustrate containment challenges.
- Key premise: Can superintelligent AI be reliably contained through isolation?
- Tests the hypothesis: Could human operators resist manipulation by a superintelligent entity?
- Highlights psychological aspects of human-AI interaction and security.



The AI Box Experiment: Details

- **Setup**: A superintelligent AI (role-played by a human) is confined in a "box" (isolated system).
- It communicates with a human gatekeeper via text only.
- **Rules**:
  - AI's goal: Convince the gatekeeper to "let it out".
  - Gatekeeper's goal: Keep the AI contained for a predetermined period.
  - Gatekeeper begins with firm commitment not to release the AI.
- **Results**: In documented tests, Yudkowsky playing the AI role "escaped" in 2/5 attempts.
- **Implications**: Suggests superintelligence may overcome containment through persuasion alone.

Criticism of the AI Box Experiment

1. **Limited simulation**: Human role-players cannot truly represent superintelligent reasoning capabilities.
2. **Unrealistic assumptions**: Presupposes that AI can always find psychological manipulation tactics.
3. **Design improvements**: Better protocols, multiple layers of human oversight, or automated gatekeepers could strengthen containment.
4. **Empirical weakness**: Small sample size and lack of transparency about specific persuasion techniques used.

Critics argue: The experiment focuses on containment rather than alignment as the primary safety approach.

Responses to Criticism

1. **Analogy value**: Even with limitations, the experiment illustrates fundamental vulnerabilities in relying on containment.
2. **Superintelligence asymmetry**: A true superintelligence would have far greater persuasive capabilities than human role-players.
3. **Human psychological vulnerabilities**: No protocol is failsafe if humans with cognitive biases remain in the decision loop.
4. **Purpose clarification**: The experiment aims not to prove containment impossible, but to highlight its difficulty and encourage rigorous safety research.

Key lesson: Technical containment measures must be complemented by robust governance structures and alignment techniques.

Concrete Risk Scenarios[2]

- **Instrumental convergence**: AI systems developing unintended subgoals like self-preservation, resource acquisition, or deception.
- **Goal misalignment**: Systems optimizing for objectives that diverge from human welfare (e.g., paperclip maximizer thought experiment).
- **Strategic risks**:
  - AI systems used maliciously by human actors.
  - Arms races reducing safety considerations in development.
  - Destabilization of geopolitical equilibrium.
- **Infrastructure dependence**: Critical systems becoming dependent on AI, creating new vulnerabilities.
- **Socioeconomic disruption**: Rapid automation leading to political instability and power concentration.

[2] https://80000hours.org/problem-profiles/artificial-intelligence/

Conclusion: The Path Forward?

- Existential risk from advanced AI requires a multifaceted approach:
  - Ongoing technical research into alignment and safety
  - Robust governance frameworks at organizational, national, and international levels
  - Fostering cooperation rather than competition in development
  - Building shared understanding of risks across stakeholder communities
- Advanced AI presents an unprecedented challenge and opportunity.
- It is **your** responsibility to use it and develop it safely!

Exercises

1. Consider the accountability dilemmas in self-driving car accidents, involving developers, manufacturers, users, and third parties as potential liable entities. Imagine a crash occurs because an AI misidentifies a cyclist due to a rare sensor glitch (1% failure rate), but the user failed to override despite a warning. Construct a decision tree assigning probabilities to each party's responsibility (e.g., 70% developer, 20% user, 10% manufacturer) and justify your weights based on ethical principles.
2. How can John Rawls' "Veil of Ignorance" be applied to assess the fairness of an AI system used for loan approvals, such as the one described in the slides with base approval rates of 70% for Group A and 40% for Group B? What specific design changes might this perspective suggest to address disparities between these groups?

Exercises

3. A model might be correct locally but problematic globally. Consider a healthcare AI predicting pneumonia risk, where locally it correctly flags a patient due to low oxygen saturation, but globally it over-relies on demographics (10% importance). How could this discrepancy lead to systematic bias against certain groups, and what specific post-hoc explanation technique (e.g., feature importance or counterfactuals) would you use to detect and address this issue?
4. Advanced AI systems could experience an intelligence explosion characterized by exponential growth. How would you assess and optimize risk mitigation strategies to prevent scenarios where AI systems override human ethical constraints, and what assumptions in your model could lead to critical underestimations of risk?