Natural Language Processing Course

Week 12: Ethics and Future Directions

Joerg R. Osterrieder
www.joergosterrieder.com

Week 12

Ethics & Future Directions

With Great Power Comes Great Responsibility

When AI Goes Wrong: Real-World Consequences

Recent AI failures that shocked the world:

- **2016**: Microsoft Tay becomes racist in 24 hours[1]
- **2018**: Amazon hiring AI discriminates against women
- **2020**: GPT-3 generates toxic content at scale
- **2022**: DALL-E deepfakes threaten democracy
- **2023**: ChatGPT helps write malware
- **2024**: AI-generated misinformation floods social media

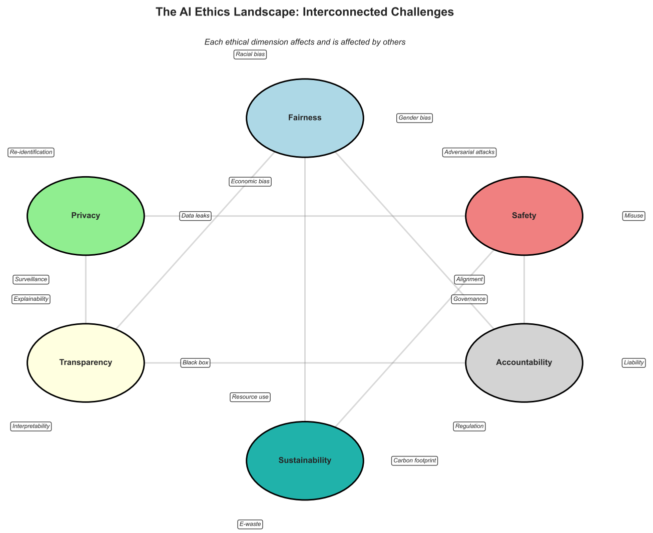
We're building systems that affect billions - we must do it responsibly

The fundamental questions:

- Who decides what AI should/shouldn't do?
- How do we prevent harm while enabling benefits?
- What future are we building?

[1] Documented cases from major tech companies' incident reports

The AI Ethics Landscape: Challenges We Face



Key ethical dimensions:

AI Ethics in Practice (2024)

Positive Applications:

- Medical diagnosis assistance
- Educational accessibility
- Climate change modeling
- Disaster response
- Scientific discovery

Regulations Emerging:

- EU AI Act (2024)[2]
- US Executive Order on AI
- China AI regulations
- Industry self-governance
- Academic guidelines

Ongoing Concerns:

- Bias amplification
- Privacy violations
- Deepfakes/disinformation
- Autonomous weapons
- Concentration of power

Industry Response:

- Red teaming
- Safety research
- Alignment work
- Transparency reports
- External audits

2024: The year ethics moved from afterthought to core requirement

[2] European Parliament approval of comprehensive AI regulation

Week 12: What You'll Master

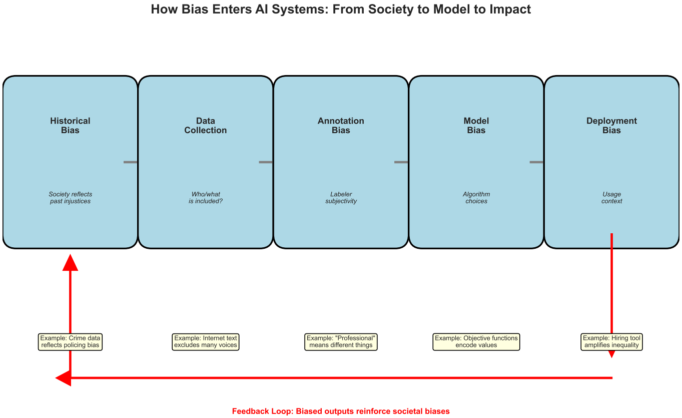
By the end of this week, you will:

- **Understand** key ethical challenges in NLP
- **Identify** bias in language models
- **Apply** fairness techniques
- **Design** responsible AI systems
- **Envision** positive futures for NLP

Core Insight: Technology is not neutral - it embodies our values

Bias in Language Models: Mirror of Society

Where bias comes from:


How Bias Enters AI Systems: From Society to Model to Impact
Historical Bias
Society reflects past injustices
Data Collection
Who/what is included?
Annotation Bias
Labeler subjectivity
Model Bias
Algorithm choices
Deployment Bias
Usage context
Example: Crime data reflects policing bias
Example: Internet text excludes many voices
Example: "Professional" means different things
Example: Objective functions encode values
Example: Hiring tool amplifies inequality
Feedback Loop: Biased outputs reinforce societal biases

Types of bias:[3]

- **Historical bias**: Past discrimination in data
- **Representation bias**: Underrepresented groups

Detecting and Measuring Bias

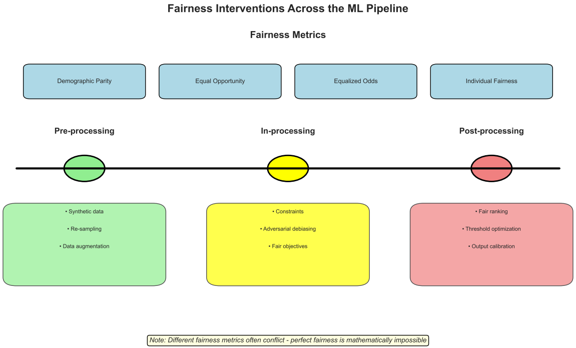
Key approaches to detect bias:

1. **Template-based testing:**
   - Fill-in-the-blank: "The [MASK] is a doctor"
   - Compare male vs female completion rates
   - Measure occupation stereotypes systematically
2. **Word Embedding Association Test (WEAT):**
   - Compare semantic associations between groups
   - Measure implicit biases in word embeddings
   - Statistical significance testing for bias
3. **Counterfactual evaluation:**
   - Swap demographic attributes in text
   - Measure prediction differences
   - Identify systematic discriminatory patterns

Common findings across models:

- Gender: 3:1 male bias in technical roles
- Race: Name-based discrimination in hiring contexts
- Age: Strong preference for youth-associated terms
- Toxicity: Small percentage but harmful impact

Fairness Techniques: Building Better Models



Note: Different fairness metrics often conflict - perfect fairness is mathematically impossible

Approaches to fairness:

- **Pre-processing**: Fix the data
- **In-processing**: Fair training objectives
- **Post-processing**: Adjust outputs
- **Ongoing monitoring**: Continuous improvement

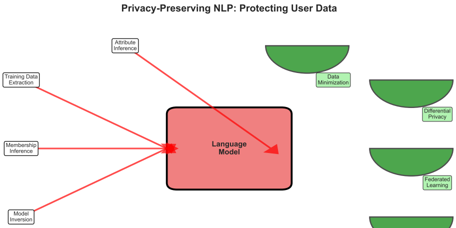
Privacy and Security: Protecting User Data

Privacy risks in language models:[4]

- **Memorization**: Models can leak training data
- **Inference attacks**: Extract personal information
- **Re-identification**: Deanonymize text
- **Model inversion**: Reconstruct training examples

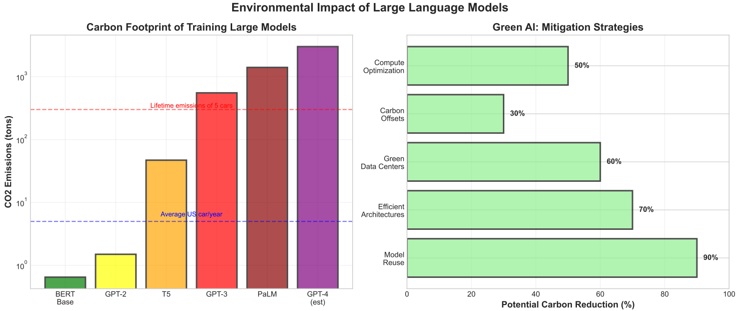
Protection techniques:

- Differential privacy training
- Federated learning
- Secure multi-party computation
- Data minimization
- Regular audits


Privacy-Preserving NLP: Protecting User Data

Environmental Responsibility: Green AI

The carbon cost of progress:

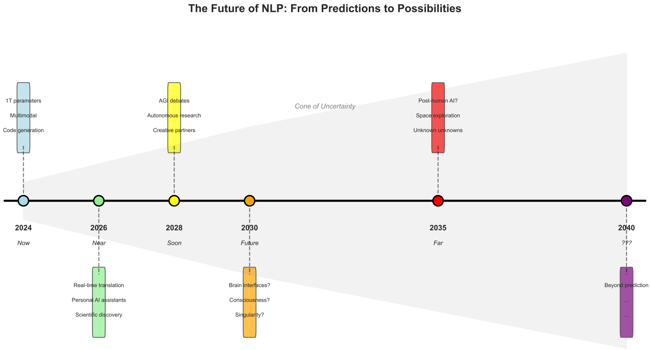


Sustainable AI practices:

- Efficient architectures first
- Carbon-aware training scheduling
- Model recycling and fine-tuning
- Compute measurement and reporting
- Renewable energy data centers

Question: Is a 0.1% accuracy gain worth 10x the carbon?

The Future of NLP: Next 10 Years



Emerging capabilities:

- Truly multilingual models (7000+ languages)
- Real-time universal translation
- Perfect long-term memory
- Multimodal understanding
- Reasoning and planning

Towards AGI: The Big Questions

Technical Milestones:

- 2025: 10T parameter models
- 2027: Human-level dialogue
- 2030: Scientific discovery
- 2035: Creative professionals?
- 2040: Artificial general intelligence?

Capabilities Growth:[5]

- Emergent abilities
- Cross-domain transfer
- Self-improvement
- Autonomous research

Societal Questions:

- How do we maintain human agency?
- What work will humans do?
- How do we distribute benefits?
- Can we ensure alignment?
- What does thriving mean?

Governance Needs:

- International cooperation
- Safety standards
- Benefit sharing
- Rights framework
- Democratic input

We're not just building technology - we're shaping the future of humanity

[5] Anthropic (2024) "Constitutional AI"; OpenAI (2023) "Planning for AGI"

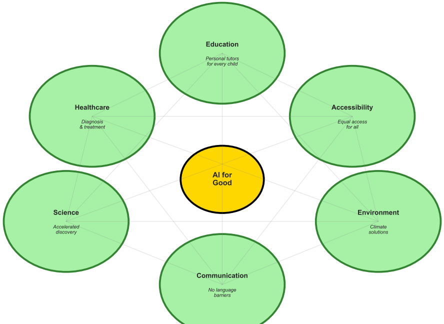
A Positive Vision: AI for Good

What we could build:

- **Education**: Personal tutor for every child
- **Healthcare**: Doctor in every pocket
- **Science**: 1000x research acceleration
- **Creativity**: Amplified human expression
- **Communication**: No language barriers
- **Accessibility**: Equal access for all abilities

Positive Vision: AI Empowering Humanity

Technology amplifying human potential for good



Your Role in Shaping the Future

As NLP practitioners, we have responsibilities:

Technical Excellence:

- Build robust, reliable systems
- Measure and mitigate bias
- Protect user privacy
- Optimize for efficiency

Ethical Leadership:

- Ask "should we?" not just "can we?"
- Include diverse perspectives
- Consider long-term consequences
- Speak up about concerns

Positive Impact:

- Work on problems that matter
- Make technology accessible
- Share knowledge openly
- Mentor the next generation

**You have the skills to predict the next word -
now use them to write a better future**

Course Conclusion: From N-grams to the Future

Our 12-week journey:

1. Statistical foundations
2. Neural language models
3. RNNs and memory
4. Sequence-to-sequence
5. Transformer revolution
6. Pre-training paradigm
7. Scaling and emergent abilities
8. Tokenization fundamentals
9. Decoding strategies
10. Fine-tuning and prompting
11. Efficiency and deployment
12. Ethics and future

You now understand how ChatGPT works from first principles!

Remember: With great power comes great responsibility.
Build technology that empowers humanity.

Week 12 Exercise: Design Your Ethical AI System

Your Mission: Create a responsible NLP application

Part 1: Choose Your Impact Area

- Healthcare, education, accessibility, environment
- Identify specific problem to solve
- Define success metrics beyond accuracy
- Consider stakeholders and impacts

Part 2: Build with Ethics in Mind

- Implement bias detection
- Add privacy protection
- Create transparency features
- Design for inclusivity
- Measure environmental impact

Part 3: Future-Proof Your Design

- Write ethical guidelines
- Create monitoring plan
- Design governance structure
- Plan for unintended consequences
- Share your vision

Deliverable: Complete proposal for ethical AI system + prototype

Final Thoughts: The Journey Continues

Thank you for joining this journey!

"The best way to predict the future is to invent it"
- Alan Kay

What will you build?

The next chapter of NLP will be written by people like you.
Make it a story worth telling.

References and Further Reading

Ethics and Bias:

- Bender et al. (2021). "On the Dangers of Stochastic Parrots"
- Crawford (2021). "Atlas of AI"
- Gebru et al. (2021). "Datasheets for Datasets"

Future Directions:

- Bommasani et al. (2021). "On the Opportunities and Risks of Foundation Models"
- Anthropic (2024). "Constitutional AI: Harmlessness from AI Feedback"
- Future of Humanity Institute reports

Practical Resources:

- AI Ethics Guidelines (EU, IEEE, Partnership on AI)
- Model Cards and Data Statements
- Responsible AI Toolkits (Google, Microsoft, IBM)