

THE UNIVERSITY *of* EDINBURGH

Researching Responsible and Trustworthy Natural Language Processing

Week 8: Research Ethics

Prof Frauke Zeller

What are ethical considerations in NLP research?



Yutong Liu / https://betterimagesofai.org /
https://creativecommons.org/licenses/by/4.0/

Different Viewpoints

Deontology

Virtue Ethics

Consequentialism/Utilitarianism

…



Lone Thomasky & Bits&Bäume / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Deontology

The word deontology derives from the Greek words for duty (deon) and science (or study) of (logos).

A normative theory regarding which choices are morally required, forbidden, or permitted.

Deontology guides and assesses our choices of what we ought to do (deontic theories).



(quoted from https://plato.stanford.edu/entries/consequentialism/)

Leo Lau & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Virtue Ethics

In contrast to deontology, (aretaic [virtue] theories) guide and assess what kind of person we are and should be.

Falls under the three main approaches in normative ethics: deontology, virtue ethics and consequentialism/utilitarianism.

(quoted from https://plato.stanford.edu/entries/consequentialism/)



Lone Thomasky & Bits&Bäume / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Consequentialism/Utilitarianism

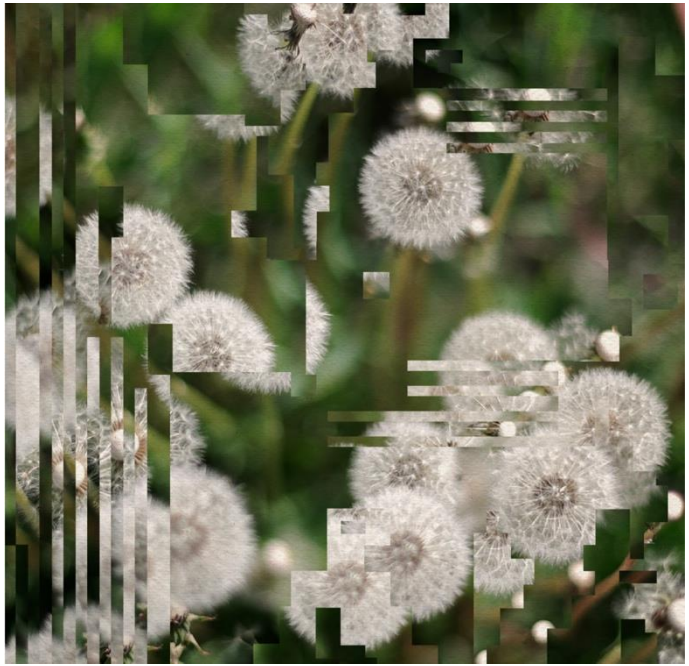
It is the view that normative properties depend only on consequences.

This historically important and still popular theory embodies the basic intuition that what is best or right is whatever makes the world best in the future, because we cannot change the past, so worrying about the past is no more useful than crying over spilled milk.



(quoted from https://plato.stanford.edu/entries/consequentialism/)

Lone Thomasky & Bits&Bäume / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/



Lone Thomasky & Bits&Bäume / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Holistic Approach

Including ethical considerations throughout the entire process:

- Design – ensuring diversity representation in data collection
- Transparency in model development
- Implementing mechanisms for accountability
- Ongoing evaluation of the model's outcomes and user interaction

Interdisciplinarity

Ethical approaches require input from all stakeholders

It is also very useful to discuss these with experts from other fields



Leo Lau & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

UKRI Six Principles of Ethical Research

1. Research should aim to maximise benefit for individuals and society and minimise risk and harm.
2. The rights and dignity of individuals and groups should be respected.
3. Wherever possible, participation should be voluntary and appropriately informed.
4. Research should be conducted with integrity and transparency.
5. Lines of responsibility and accountability should be clearly defined.
6. Independence of research should be maintained and where conflicts of interest cannot be avoided they should be made explicit.

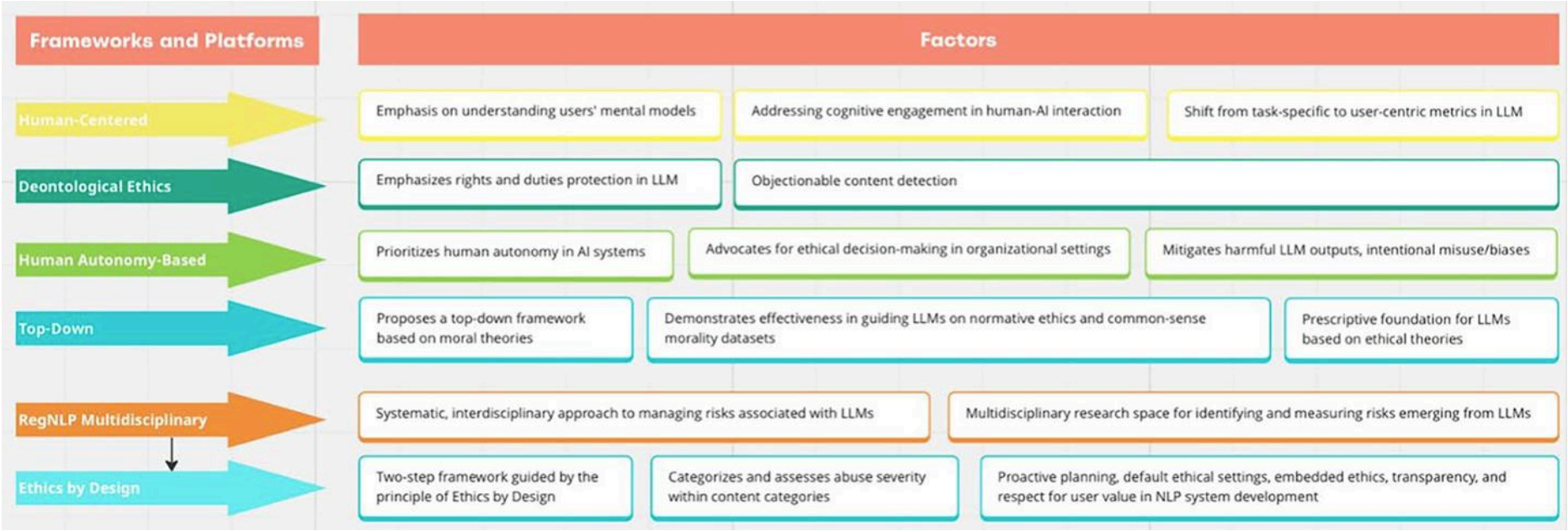
LLM Ethics

Jiao et al. (2025) conducted an inclusive and systematic review of academic papers, reports, case studies, and frameworks regarding LLM ethics, written in English.

The analysis produced 13 relevant topics related to LLM and ethics:

1. Ethical concerns in LLM
2. Ethical frameworks and platforms in LLM
3. Bias and fairness in LLM
4. Privacy and data security in LLM
5. Misinformation and disinformation in LLM
6. Accountability and governance in LLM
7. Case studies in LLM ethics
8. Mitigation strategies in LLM ethics
9. Transparency in LLM
10. Censorhsip in LLM
11. Intellectual property and plagiarism in LLM
12. Abusive LLM, hate speech and cyber-bullying
13. Auditing LLM

Jiao, J., Afroogh, S., Xu, Y. et al. Navigating LLM ethics: advancements, challenges, and future directions. AI Ethics (2025). https://doi.org/10.1007/s43681-025-00814-5



Jiao, J., Afroogh, S., Xu, Y. et al. Navigating LLM ethics: advancements, challenges, and future directions. AI Ethics (2025). https://doi.org/10.1007/s43681-025-00814-5



Jiao, J., Afroogh, S., Xu, Y. et al. Navigating LLM ethics: advancements, challenges, and future directions. AI Ethics (2025). https://doi.org/10.1007/s43681-025-00814-5

THE UNIVERSITY of EDINBURGH

Jiao, J., Afroogh, S., Xu, Y. et al. Navigating LLM ethics: advancements, challenges, and future directions. AI Ethics (2025). https://doi.org/10.1007/s43681-025-00814-5

Major strategies	Key factors	Explanations	References
Bias Mitigation	Dataset Enhancement	Explore methods to improve the diversity and inclusivity of the training dataset	[129, 133, 140 144, 152
	Occupation-Focused Tuning	Implement occupation-focused fine-tuning, as demonstrated by OccuQuest, a specialized dataset encompassing various occupations	[129]
	InfoEntropy Loss Function	Utilize dynamic evaluation of learning difficulty through functions like InfoEntropy Loss to guide the model towards challenging tokens	[126]
	Adversarial Learning	Integrate adversarial learning during pre-training to address bias in natural language generation tasks	[134]
	Semantic Similarity Task	Fine-tune models on tasks emphasizing semantic similarity to reduce gender bias	[133]
	Social-Group-Agnostic	Capture the underlying connection between bias and stereotypes to help reduce bias among social groups	[204]
Privacy Protection and Data Security	Knowledge Unlearning	Knowledge Unlearning: Employ knowledge unlearning techniques to reduce privacy risks for LLMs	[131]
	Embedding Purification	Utilize Embedding Purification methods in conjunction with clean pre-trained weights to mitigate potential backdoors in word embeddings	[130]
	Comparative Testing	Evaluate proposed privacy protection methods against established data preprocessing and decoding methods	[131]
Hallucination Prevention	Logit Output Verification	Implement a method that detects and mitigates hallucinations during content generation by verifying potential hallucinations through the model's logit output values	[139]
	Proactive Detection	Actively mitigate hallucinations before generating content by proactively identifying and addressing potential issues	[139]
	Participatory Design	Involve users in the design process to create features that reduce the impact of hallucinations in LLMs	[151]

Bring these approaches back to our three main normative approaches to ethics:

Deontology
Virtue Ethics
Consequentialism/Utilitarianism

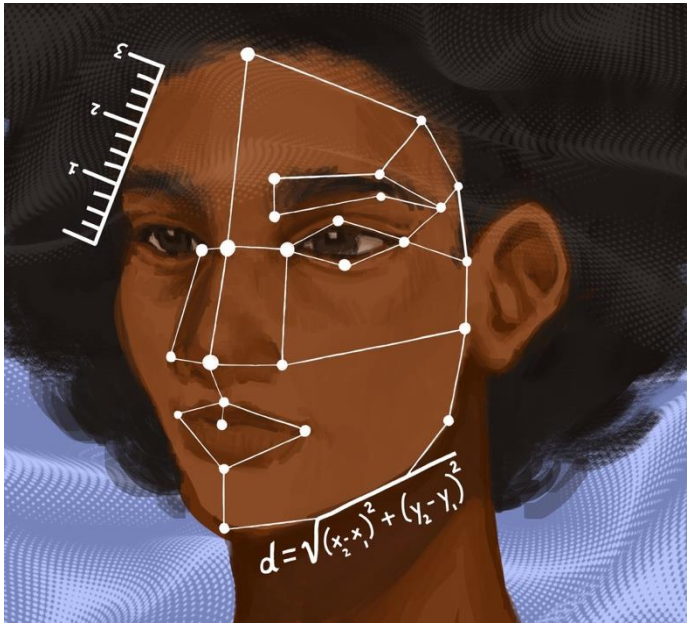
How does this all apply to you/us as researchers?

Case Study: Evaluating HRI

How to measure, evaluate and analyse Human-Robot Interaction?

How to find the perfect scale?
https://hriscaledatabase.psychology.gmu.edu/

How to measure ethics?
How to operationalise ethical norms and guidelines?



Alyssa Chen / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Socially Assistive Robots in Paediatric Healthcare

Objective

- Multi-disciplinary approach to ethical and EDI-based design of socially assistive robots in paediatric care.

Context

- Focus on establishing trustworthiness, acceptability in healthcare robots for patients, their caregivers, and the medical professionals
- Increasing use of AI and robots in healthcare.

Ethical Concerns

- Potential for AI to reinforce societal stereotypes and biases.
- Need for EDI integration in robot design.

Socially assistive robots in paediatric healthcare primarily provide therapeutic benefits through communication and social interaction. Below is more details regarding their assistive roles:

Pain & Anxiety Management

- Application: Stress reduction in needle-based procedures.
- Effectiveness: Statistically significant relief from pain, distress, fear.
- Methods: Cognitive-behavioural strategies for engagement.

Developmental & Physical Needs

- ASD Therapy: Positive impacts on speech, social interaction, emotional expression.
- Physical Therapy: Promising results in pilot studies for cerebral palsy, acquired brain injury.

Ethics and EDI in Robots for Paediatrics:

Addressing bias in design

Trust, privacy, and deception

Gender

Race

Disability

Addressing Bias in Design

Inherent Biases in AI

- Risks of unfair implementation favoring certain accents, skin tones.
- Potential to compromise benefits for marginalized children.

Algorithmic Challenges

- Biases informed by human experts' data.
- Risk of perpetuating stereotypes.

Socio-Economic & Cultural Factors

- Socio-economic status influences comfort and interaction with robots.
- Complexity in addressing identity for culturally sensitive care.

Trust, Privacy and Deception

Acceptability Across Stakeholders

- Design needs to appeal to children, medical staff, and parents.
- Concerns extend beyond functionality to meeting individual needs.

Trust & Reliability

- Trust essential for user comfort and collaboration.
- Debate on evaluating AI for reliability vs trustworthiness.

Privacy Considerations

- Differing views on privacy between adults and children.
- Adult concerns over consent in the context of big data.

Ethical Complexities

Debate on potential deception by emotionally responsive robots.

'Deception' varies in meaning between healthcare and HRI settings.

Gender

Bidirectional Influence

- User's features and robot's features both influence interaction.

Gender Stereotyping

- Research shows robot "gender" invokes human-like biases.
- Metrics used include perceived 'communion' and 'agency'.

Design & Perception

- Facial cues, body shape affect gender perception.
- Gender neutrality insufficient to avoid biases.

Gender

Trust & Suitability

- Gender impacts levels of cognitive and affective trust.
- Female robots more trusted and seen as suitable for caregiving.

Implications for Children

- Vulnerability to stereotyping due to higher tendency to anthropomorphize.

Ethical Concerns

- Gender cues can reinforce harmful stereotypes.
- Need for intentional design to mitigate biases.

Healthcare Context

- Complex trust dynamics due to existing gender issues in heal

Race

Etymological Roots

- Word 'robot' linked to concepts of slavery.
- Relevance for groups with colonial histories, like Indigenous people.

WEIRD Bias in HRI Research

- Research predominantly from Western, Educated, Industrialized, Rich, Democratic countries.
- Lacking diversity in perspectives and representation.

Race

Risk of Perpetuating Racism

- Ignoring race can lead to reinforcement of racist ideas.

Approaches for Racially Inclusive Design

- Diverse languages and accents.
- Anthropocentric robots reflecting human racial diversity.
- Racial and cultural diversity on the design team.

Indigenous Relationality

- Concept of "making kin with machines" for respectful relations with robots as “other-than-human” relations.

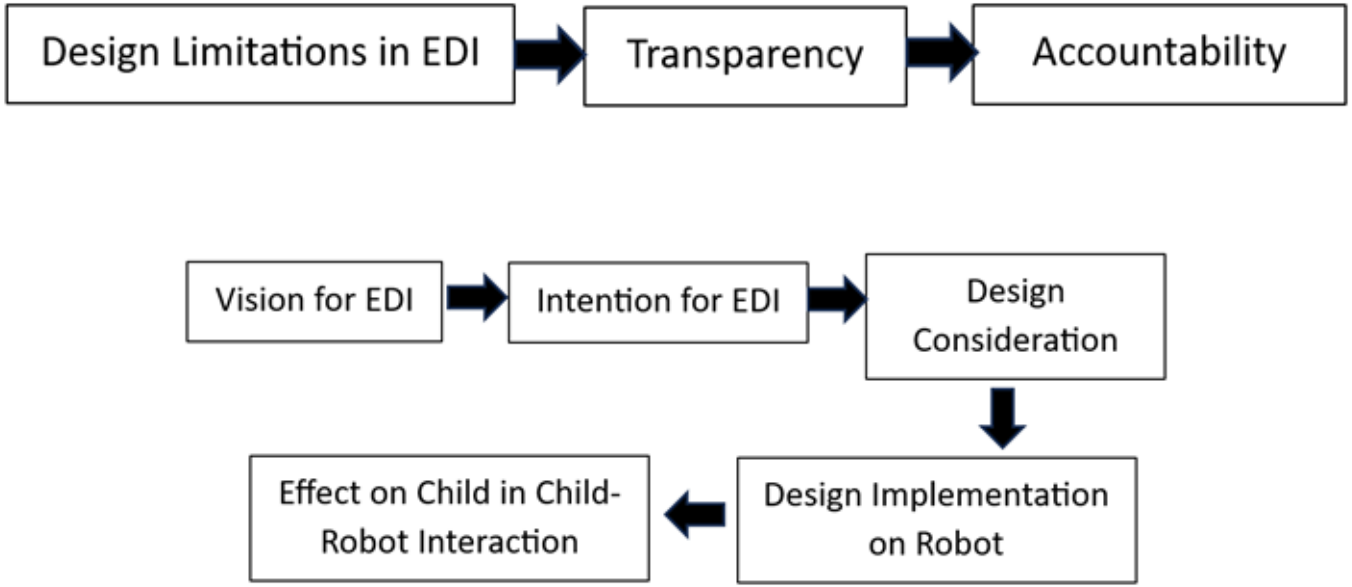
Vision for the Future

- Need for racially and culturally diverse representation in robot design.

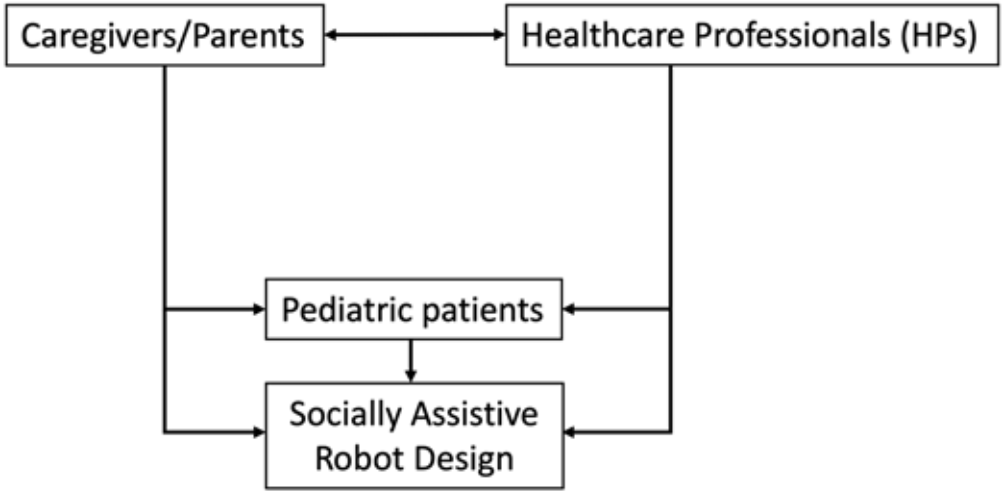
Disability

- Diversity of Disabilities
  - Hearing, visual, mobility limitations, and neurodiversity.
- Specialized Design Considerations
  - Different disabilities require unique design accommodations.
- Sætra et al.'s Framework
  - "Explanation", "Neutralization", and "Queering".
    - "Explanation": Transparency in design to avoid isolation.
    - "Neutralization": Design for typical and atypical behaviors.
    - "Queering": Design that includes neurodivergent behaviors.
- Critique of "Assumption of Normalization"
  - Need for avoiding designs that aim to normalize children into assumed behaviors.

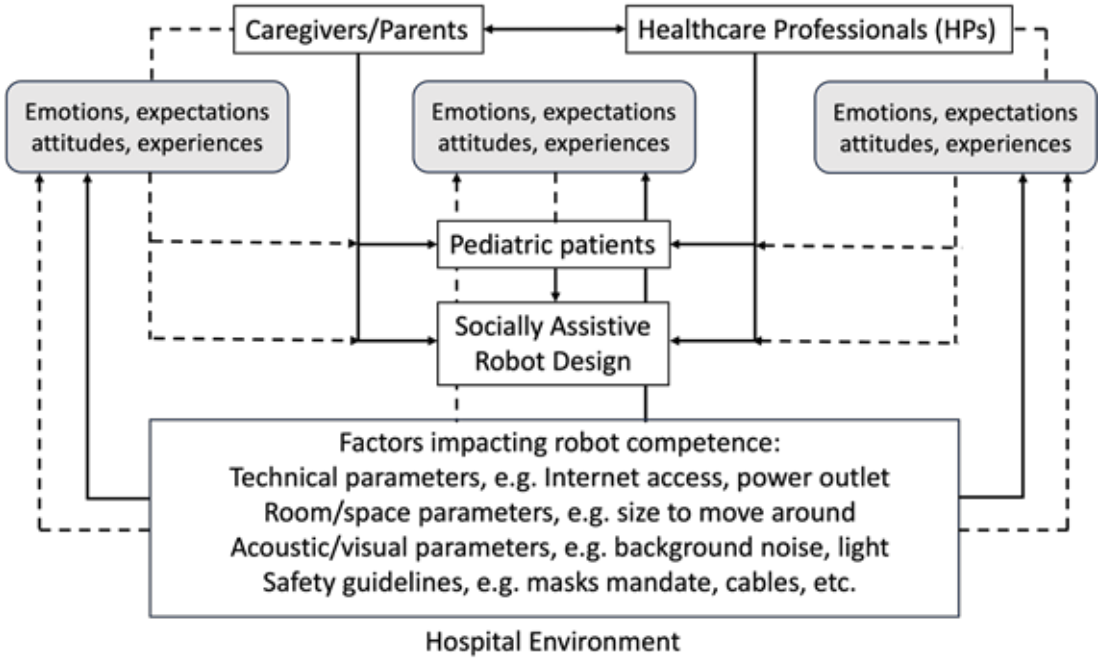
Recommendations for EDI-Based Design and Development



Recommendations for EDI-Based Design and Development



Recommendations for EDI-Based Design and Development





Thank you!

Reihaneh Golpayegani & Digit / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/