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

…

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/)

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/)

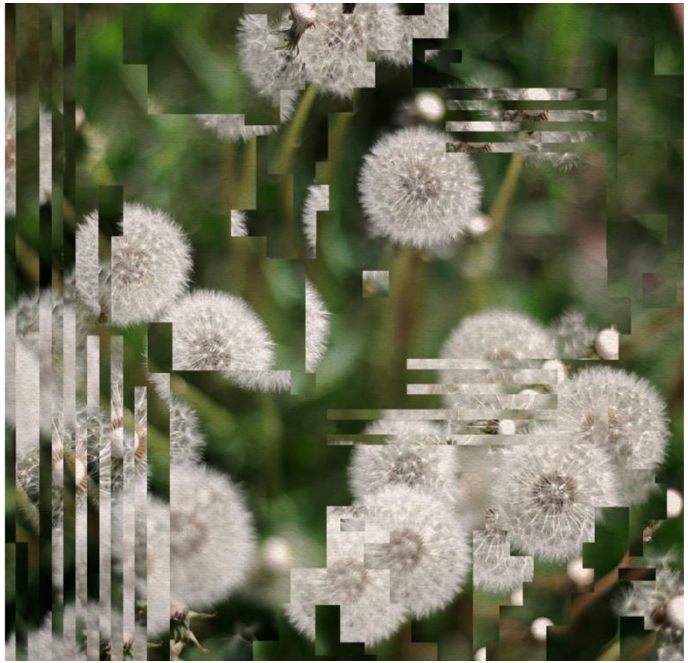
Lone Thomasky & Bits&Bäume / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

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/

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

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.

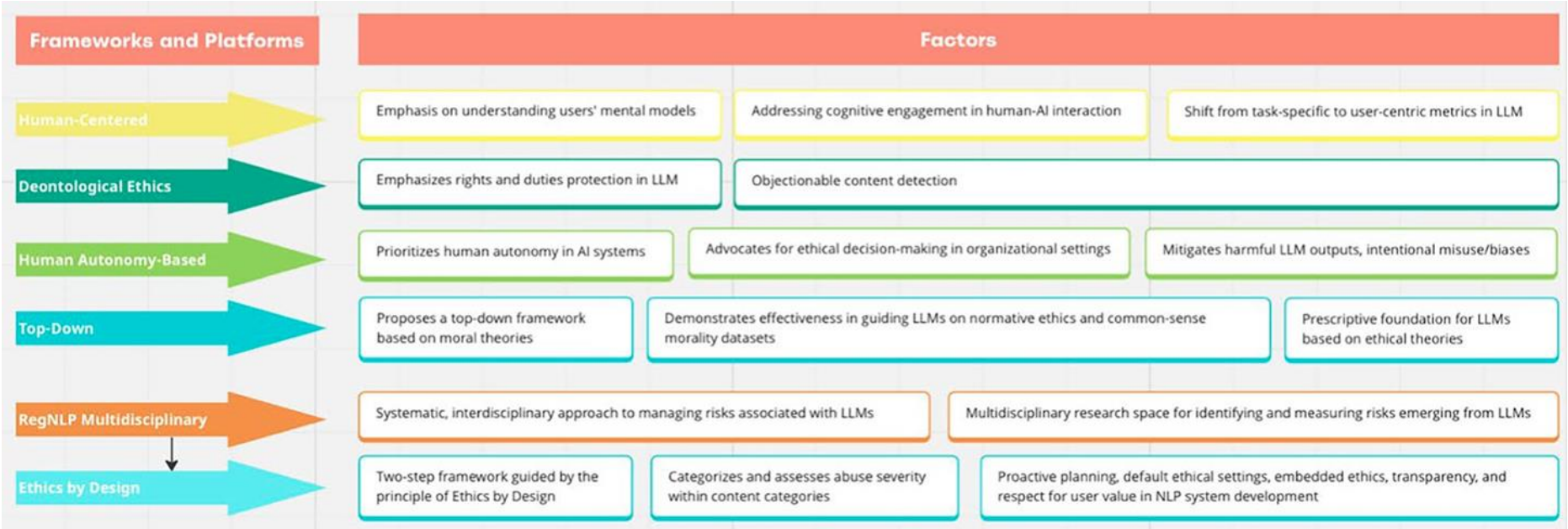
LLM Ethics

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

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?

Thank you!

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