



Ethics in NLP

Types of risk

Bias amplification: systems exacerbate real-world bias rather than correct for it

Exclusion: underprivileged users are left behind by systems

Dangers of automation: automating things in ways we don't understand is dangerous

Unethical use: powerful systems can be used for bad ends

Unethical Use

- ▶ Surveillance applications?
- ▶ Generating convincing fake news / fake comments?

FCC Comment ID: 106030756805675	FCC Comment ID: 106030135205754	FCC Comment ID: 10603733209112
Dear Commissioners:	Dear Chairman Pai,	---
Hi, I'd like to comment on	I'm a voter worried about	In the matter of
net neutrality regulations.	Internet freedom.	NET NEUTRALITY.
I want to	I'd like to	I strongly
implore	ask	ask
the government to	Ajit Pai to	the commission to
repeal	repeal	reverse
Barack Obama's	President Obama's	Tom Wheeler's
decision to	order to	scheme to
regulate	regulate	take over
internet access.	broadband.	the web.
Individuals,	people like me,	People like me,
rather than	rather than	rather than

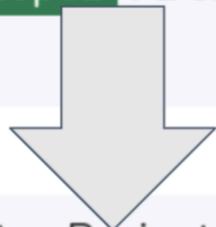
- ▶ What if these were undetectable?

Unethical Use

Anonymization (De-Identification)

Informe clínico del paciente : Paciente **varón** de **70 años** de edad ,
sin alergias medicamentosas conocidas . Operado de
una hernia el **12 de enero de 2016** en el **Hospital Costa del
Sol** por la Dra . **Juana López** . Derivado a este centro el día 16 del
mismo mes para revisión .

Category: DATE
Tagger: PHI NER



Informe clínico del paciente : Paciente **SEX** de **AGE AGE** de edad ,
PROFESSION jubilado , sin alergias medicamentosas conocidas .
Operado de una hernia el **DATE DATE DATE DATE DATE** en el
HOSPITAL HOSPITAL HOSPITAL HOSPITAL por la Dra .
DOCTOR DOCTOR . Derivado a este centro el día 16 del mismo mes
para revisión .

HitzalMed
(Lopez et al., 2020)

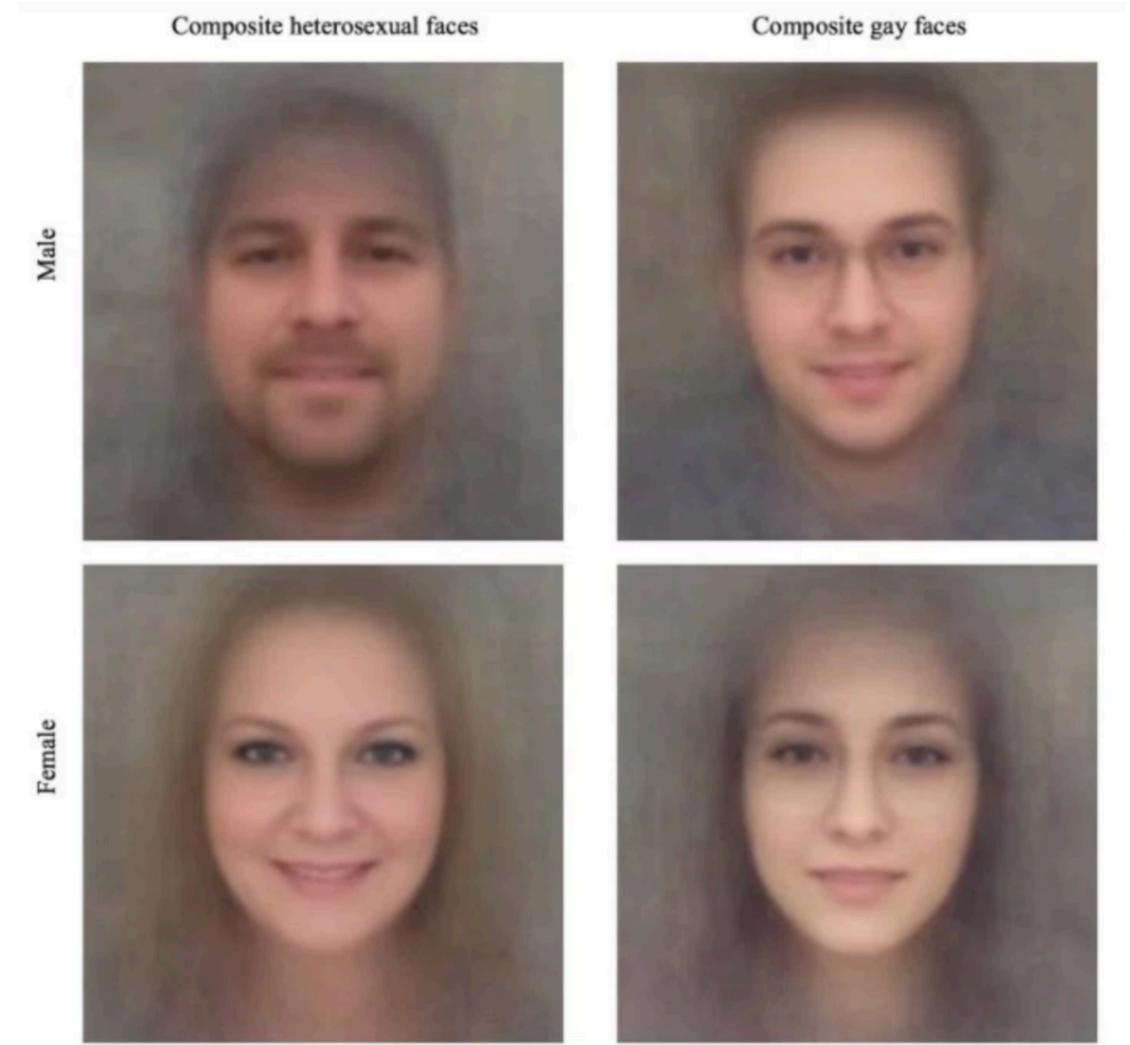
After having run some
anonymization system
on our data, is
everything fine?

Image Source: <https://www.aclweb.org/anthology/2020.lrec-1.870/>

Friedrich + Zesch

Unethical Use

- ▶ Wang and Kosinski: gay vs. straight classification based on faces
- ▶ Authors argued they were testing a hypothesis: sexual orientation has a genetic component reflected in appearance
- ▶ Blog post by Agüera y Arcas, Todorov, Mitchell: the system detects mostly social phenomena (glasses, makeup, angle of camera, facial hair)
- ▶ Potentially dangerous tool, and **not even good science**



Slide credit: <https://medium.com/@blaisea/do-algorithms-reveal-sexual-orientation-or-just-expose-our-stereotypes-d998fafdf477>

Unethical Use: LLMs

- ▶ Many hypothesized issues, although not much documentation/systematic study yet:
 - ▶ AI-generated misinformation (intentional or not)
 - ▶ Cheating/plagiarism (in school, academic papers, ...)
 - ▶ “Better Google” can also help people learn how to build bombs and things like that
- ▶ LLMs are a powerful tool, and such tools bring up questions of how they concentrate power, and more...

How to move forward

- ▶ Hal Daume III: Proposed code of ethics
<https://nlpers.blogspot.com/2016/12/should-nlp-and-ml-communities-have-code.html>
- ▶ Many other points, but these are relevant:
 - ▶ Contribute to society and human well-being, and minimize negative consequences of computing systems
 - ▶ Make reasonable effort to prevent misinterpretation of results
 - ▶ Make decisions consistent with safety, health, and welfare of public
 - ▶ Improve understanding of technology, its applications, and its potential consequences (pos and neg)
- ▶ Value-sensitive design: vsdesign.org
 - ▶ Account for human values in the design process: understand *whose* values matter here, analyze how technology impacts those values

How to move forward

- ▶ Datasheets for datasets [Gebru et al., 2018]
<https://arxiv.org/pdf/1803.09010.pdf>
 - ▶ Set of criteria for describing the properties of a dataset; a subset:
 - ▶ What is the nature of the data?
 - ▶ Errors or noise in the dataset?
 - ▶ Does the dataset contain confidential information?
 - ▶ Is it possible to identify individuals directly from the dataset?
- ▶ Related proposal: Model Cards for Model Reporting

How to move forward

- ▶ Closing the AI Accountability Gap [Raji et al., 2020]

<https://dl.acm.org/doi/pdf/10.1145/3351095.3372873>

Scoping	Mapping	Artifact Collection	Testing	Reflection	Post-Audit
Define Audit Scope	Stakeholder Buy-In	Audit Checklist	Review Documentation	Remediation Plan	Go / No-Go Decisions
Product Requirements Document (PRD)	Conduct Interviews	Model Cards	Adversarial Testing	Design History File (ADHF)	Design Mitigations
AI Principles	Stakeholder Map	Datasheets	Ethical Risk Analysis Chart		Track Implementation
Use Case Ethics Review	Interview Transcripts			Summary Report	
Social Impact Assessment	Failure modes and effects analysis (FMEA)				

- ▶ Structured framework for producing an audit of an AI system