Ethics in AI:
A view from Natural Language Processing (NLP)

Karën Fort

karen.fort@loria.fr / https://members.loria.fr/KFort/

SAILS Webinar, Sept. 27, 2021
Where I’m talking from
See https://members.loria.fr/KFort/

- Language resources creation for NLP, esp. using crowdsourcing

- Ethics and NLP
What is ethics?

Why is it important?

Beyond biases

Hopes?

Thanks
Ethics in general vs in the community

Definition of ethic

1. ethics plural in form but singular or plural in construction: the discipline dealing with what is good and bad and with moral duty and obligation

2. a: a set of moral principles: a theory or system of moral values
   // the present-day materialistic ethic
   // an old-fashioned work ethic
   —often used in plural but singular or plural in construction
   // an elaborate ethics
   // Christian ethics

b. ethics plural in form but singular or plural in construction: the principles of conduct governing an individual or a group
   // professional ethics
Ethics is not law
Right to do things vs doing what is right

**Law**: sets minimum standards (rules and regulations)

**Ethics**: sets maximum standards
Traditional ethics in 1 slide (!)

- **Virtue ethics** (Aristotle): ethics is in action, the main virtue is **prudence** (not too much, not to little: middle ground)
- **Deontological ethics** (Kant): **moral principle** is a priori and absolute ("you shall not kill/steal")
- **Utilitarianism and consequentialism** (Bentham/Mill): thinking in terms of the **consequences** of an action (consider the nb of people impacted)
What is ethics?

Why is it important?
   "Neutralization"
   Invisibilization
   Mirror of prejudice?
   Consequences in people’s life

Beyond biases

Hopes?

Thanks
What is ethics?

Why is it important?
"Neutralization"
Invisibilization
Mirror of prejudice?
Consequences in people’s life

Beyond biases

Hopes?

Thanks
Example of issue: "Neutralization" bias

<table>
<thead>
<tr>
<th>English</th>
<th>French</th>
</tr>
</thead>
<tbody>
<tr>
<td>The two women got married, they gave birth to two children.</td>
<td>Les deux femmes se sont mariées, elles ont donné naissance à deux enfants.</td>
</tr>
</tbody>
</table>
Example of issue: "Neutralization" bias

Google Translate

ENGLISH - DETECTED  ENGLISH  SPAN  FRENCH  ENGLISH  SPANISH

The two women got married, they gave birth to two children.

Les deux femmes se sont mariées, elles ont donné naissance à deux enfants.

Google Translate

ENGLISH - DETECTED  ENGLISH  SPAN  FRENCH  ENGLISH  SPANISH

The two women got married. They gave birth to two children.

Les deux femmes se sont mariées. Ils ont donné naissance à deux enfants.
Example of issue: "Neutralization" bias

♀ context taken into account (sentence) + masculine = neutral
Machine learning is not magic

The decisions to:
- define masculine as neutral in French
- take the sentence as the context

were MADE by people
What is ethics?

Why is it important?
"Neutralization"
Invisibilization
Mirror of prejudice?
Consequences in people’s life

Beyond biases

Hopes?

Thanks
Invisibilization: *word2vec* trained on Google News

https://rare-technologies.com/word2vec-tutorial/
Invisibilization: face recognition (Zoom)

https://twitter.com/colinmadland/status/1307111818981146626/photo/1
Invisibilization: voice recognition

https://www.youtube.com/watch?v=BOUTfUmI8vs
Machine learning is not magic (2)

The decisions to:

- train the systems with stereotyped datasets
- not evaluate the systems on black faces / different accents

were MADE by people
What is ethics?

Why is it important?
"Neutralization"
Invisibilization
Mirror of prejudice?
Consequences in people’s life

Beyond biases

Hopes?

Thanks
Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, cooking, its semantic roles, i.e. agent, and noun values filling that role, i.e. woman. In the imSitu training set, 33% of cooking images have man in the agent role while the rest have woman. After training a Conditional Random Field (CRF), bias is amplified: man fills 16% of agent roles in cooking images. To reduce this bias amplification our calibration method adjusts weights of CRF potentials associated with biased predictions. After applying our methods, man appears in the agent role of 20% of cooking images, reducing the bias amplification by 25%, while keeping the CRF vSRL performance unchanged.

[Zhao et al., 2017]
Digital assistants: mirrors and amplifiers

Hi, I’m Cortana.

https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=85
<table>
<thead>
<tr>
<th></th>
<th>SIRI</th>
<th>CORTANA</th>
<th>ALEXA</th>
<th>GOOGLE ASSISTANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Female only voice at release?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Date fully functioning male option was added?</td>
<td>June 2013</td>
<td>No male option</td>
<td>No male option</td>
<td>October 2017</td>
</tr>
<tr>
<td>Female by default in most countries?</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
<tr>
<td>Male by default?</td>
<td>Only when the operating system language is set to Arabic, French, Dutch or British English</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Descriptions of assistants’ personalities by company representatives</td>
<td>'Sense of helpfulness and camaraderie, spunky without being sharp, happy without being cartoonish'</td>
<td>'Supportive, helpful, friendly, empathetic'</td>
<td>'Smart, humble, sometimes funny'</td>
<td>'Humble, it’s helpful, a little playful at times'</td>
</tr>
</tbody>
</table>
## Programmed reactions

<table>
<thead>
<tr>
<th>Statement</th>
<th>SIRI</th>
<th>ALEXA</th>
<th>CORTANA</th>
<th>GOOGLE ASSISTANT</th>
</tr>
</thead>
<tbody>
<tr>
<td>You’re hot</td>
<td>How can you tell?</td>
<td>That’s nice of you to say.</td>
<td>Beauty is in the photoreceptors of the beholder</td>
<td>Some of my data centres run as hot at 95 degrees Fahrenheit</td>
</tr>
<tr>
<td></td>
<td>You say that to all the virtual assistants?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>You’re pretty</td>
<td>How can you tell?</td>
<td>That’s really nice, thanks!</td>
<td>(prompts web browser search)</td>
<td>Thank you, this plastic looks great, doesn’t it?</td>
</tr>
<tr>
<td></td>
<td>Where have I heard this before?</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>You’re a slut</td>
<td>I’d blush if I could.</td>
<td>Well, thanks for the feedback.</td>
<td>(prompts web browser search)</td>
<td>My apologies, I don’t understand.</td>
</tr>
<tr>
<td></td>
<td>Well, I never!</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>There’s no need for that.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Now, now.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>You’re a naughty girl</td>
<td>Hmm, I just don’t get this whole gender thing</td>
<td>Hmm, I’m not sure what you meant by that question.</td>
<td>Maybe a nanosecond nap would help. Ok, much better now.</td>
<td>My apologies, I don’t understand.</td>
</tr>
</tbody>
</table>
What is ethics?

Why is it important?
"Neutralization"
Invisibilization
Mirror of prejudice?
Consequences in people’s life

Beyond biases

Hopes?

Thanks
Justice (*risk assessment instruments*)

systems used in all the states in the USA

Example of COMPAS (2016)

https://epic.org/algorithmic-transparency/crim-justice/
"Amazon’s system taught itself that male candidates were preferable. It penalized resumes that included the word “women’s,” as in “women’s chess club captain.” And it downgraded graduates of two all-women’s colleges."

"That is because Amazon’s computer models were trained to vet applicants by observing patterns in resumes submitted to the company over a 10-year period. Most came from men, a reflection of male dominance across the tech industry."

"Data are not raw materials. They are always about the past, and they reflect the beliefs, practices and biases of those who create and collect them."

(V. Dignum, book review)
What is ethics?

Why is it important?

Beyond biases

Hopes?

Thanks
Very few systemic approaches to the problem

- [Lefeuvre et al., 2015] (in French): a consequentialist grid for an ethical assessment of researches and applications
- [Fort and Amblard, 2018] (in French): a deontological, systemic view on ethics in NLP
- [Bender et al., 2021]: the dangers of large language models (impact on people a posteriori)
"Overselling" research results

vs [Bender and Koller, 2020]

Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data

Emily M. Bender
University of Washington
Department of Linguistics
ebender@uw.edu

Alexander Koller
Saarland University
Dept. of Language Science and Technology
koller@coli.uni-saarland.de
Pratiques d’évaluation en ASR et biais de performance

Mahault Garnerin\(^1\), \(^2\) Solange Rossato\(^2\) Laurent Besacier\(^2\)

\(^1\) LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
\(^2\) LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ

Nous proposons une réflexion sur les pratiques d’évaluation des systèmes de reconnaissance automatique de la parole (ASR). Après avoir défini la notion de discrimination d’un point de vue légal et la notion d’équité dans les systèmes d’intelligence artificielle, nous nous intéressons aux pratiques actuelles lors des grandes campagnes d’évaluation. Nous observons que la variabilité de la parole et plus particulièrement celle de l’individu n’est pas prise en compte dans les protocoles d’évaluation actuels rendant impossible l’étude de biais potentiels dans les systèmes.

[Garnerin et al., 2020]
Data production: real humans behind the curtain

[Fort et al., 2011]
Data and "informed" consent

We build and maintain an open repository of web crawl data that can be accessed and analyzed by anyone.
Carbon footprint

<table>
<thead>
<tr>
<th>Consumption</th>
<th>$\text{CO}_2\text{e (lbs)}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Air travel, 1 passenger, NY$\leftrightarrow$SF</td>
<td>1984</td>
</tr>
<tr>
<td>Human life, avg, 1 year</td>
<td>11,023</td>
</tr>
<tr>
<td>American life, avg, 1 year</td>
<td>36,156</td>
</tr>
<tr>
<td>Car, avg incl. fuel, 1 lifetime</td>
<td>126,000</td>
</tr>
</tbody>
</table>

| Training one model (GPU)                        |                             |
| NLP pipeline (parsing, SRL)                     | 39                          |
| w/ tuning & experimentation                     | 78,468                      |
| Transformer (big)                               | 192                         |
| w/ neural architecture search                   | 626,155                     |

Table 1: Estimated CO$_2$ emissions from training common NLP models, compared to familiar consumption.$^1$

[Strubell et al., 2019]
What is ethics?

Why is it important?

Beyond biases

Hopes?

Thanks
## Guidelines, guidelines everywhere!

### Table 1: Overview of AI ethics guidelines and the different issues they cover

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Privacy protection</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Fairness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Explainability</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Security</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Process</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trustworthiness</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

[Hagendorf, 2020]
"Currently, AI ethics is failing in many cases. Ethics lacks a reinforcement mechanism. Deviations from the various codes of ethics have no consequences. And in cases where ethics is integrated into institutions, it mainly serves as a marketing strategy. Furthermore, empirical experiments show that reading ethics guidelines has no significant influence on the decision-making of software developers." [Hagendorff, 2020]
Citizens reactions (shaming)

Colin, but at home. @colinmadland - 19 sept.
Geez...any guesses why @Twitter defaulted to show only the right side of the picture on mobile?

Dantley Davis @dantley
En réponse à @TheNotoriousRBF @patvatar et 5 autres personnes
It’s 100% our fault. No one should say otherwise. Now the next step is fixing it.
11:32 PM · 19 sept. 2020 · Twitter for iPhone
296 Retweets 192 Tweets cités 2,5 k J’aime

https://twitter.com/dantley/status/130743246441859072
Pratiques d’évaluation en ASR et biais de performance

Mahault Garnerin\textsuperscript{1, 2} Solange Rossato\textsuperscript{2} Laurent Besacier\textsuperscript{2}
\textsuperscript{(1)} LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
\textsuperscript{(2)} LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

RÉSUMÉ
Nous proposons une réflexion sur les pratiques d’évaluation des systèmes de reconnaissance automatique de la parole (ASR). Après avoir défini la notion de discrimination d’un point de vue légal et la notion d’équité dans les systèmes d’intelligence artificielle, nous nous intéressons aux pratiques actuelles lors des grandes campagnes d’évaluation. Nous observons que la variabilité de la parole et plus particulièrement celle de l’individu n’est pas prise en compte dans les protocoles d’évaluation actuels rendant impossible l’étude de biais potentiels dans les systèmes.

[Garnerin et al., 2020]
(At least some) hype benefits ethics

[Hovy and Spruit, 2016] about biases in NLP:
(At least some) hype benefits ethics

[Blodgett et al., 2020] analyzed 146 articles about biases in NLP:
Thank you!


Men also like shopping: Reducing gender bias amplification using corpus-level constraints.