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

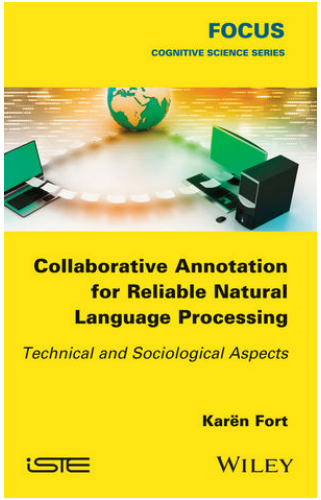
Loria
Laboratoire lorrain de recherche en informatique et ses applications

LETTRES SORBONNE UNIVERSITÉ

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

GAMES | BROWSE THESAURUS | WORD OF THE DAY | WORDS AT PLAY

Merriam-Webster SINCE 1828

ethic

Dictionary | Thesaurus

eth·ic | \ ˈe-thik \

Definition of *ethic*

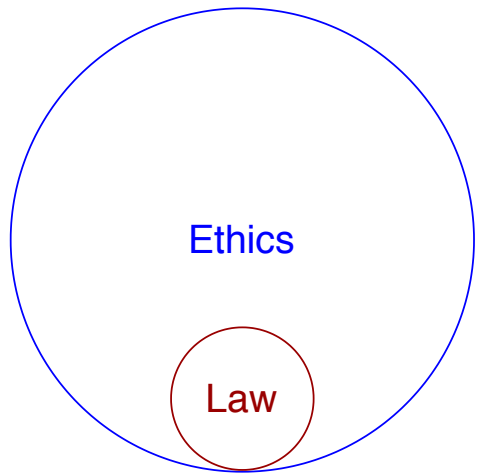
1 **ethics** *plural in form but singular or plural in construction* **:** the discipline dealing with what is good and bad and with moral duty and obligation

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)

vs

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?

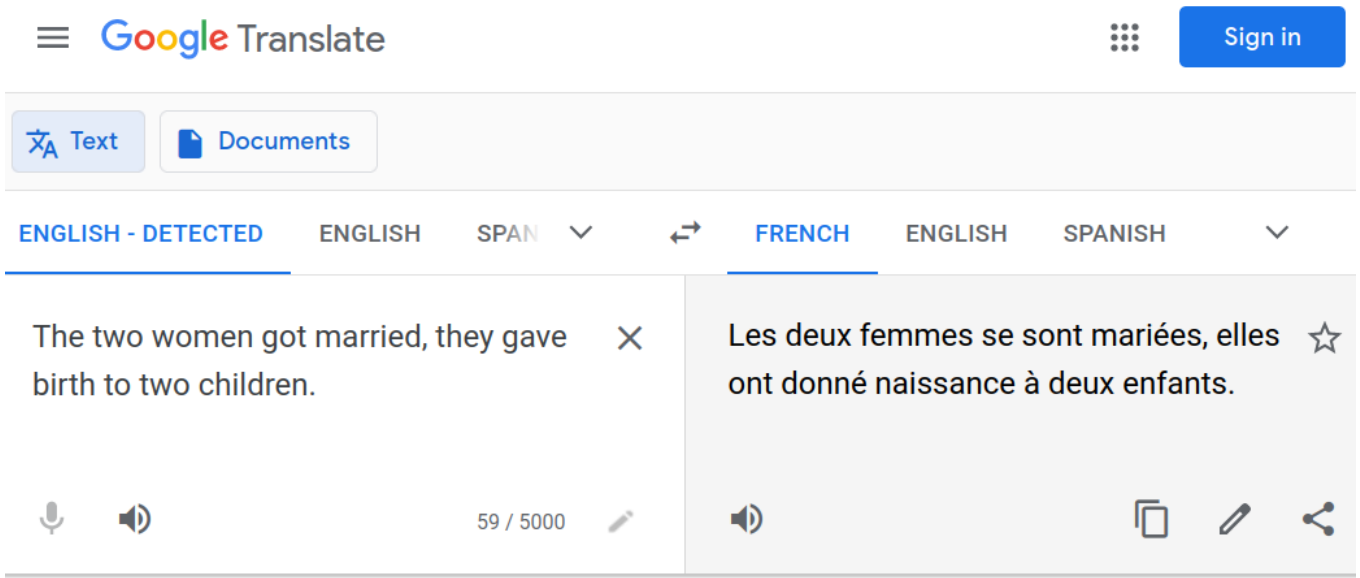
- "Neutralization"
- Invisibilization
- Mirror of prejudice?
- Consequences in people's life

Beyond biases

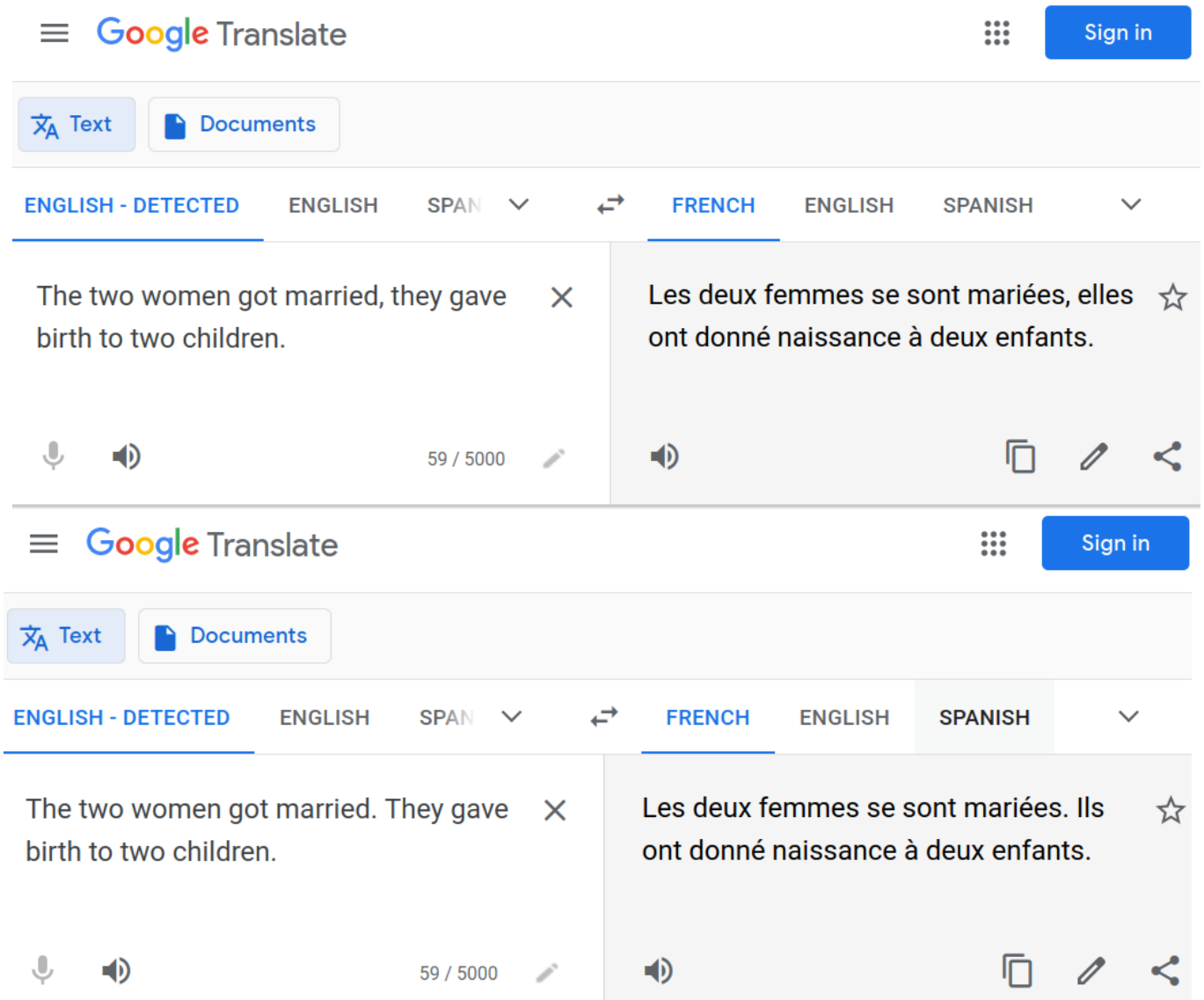
Hopes?

Thanks

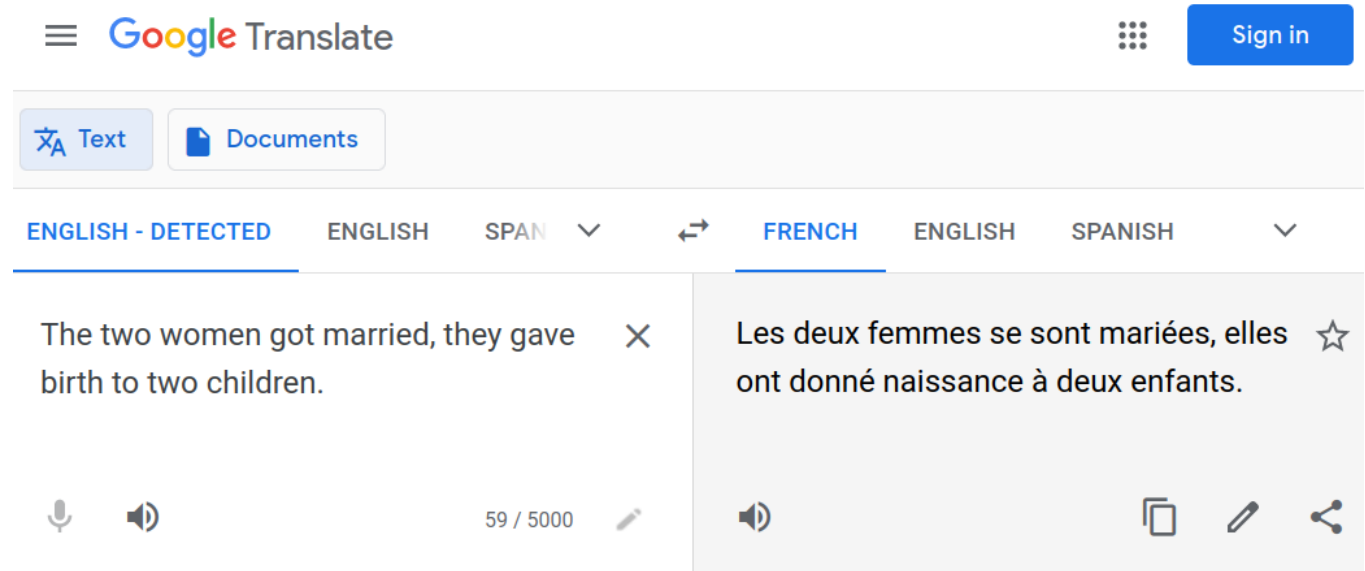
Example of issue: "Neutralization" bias

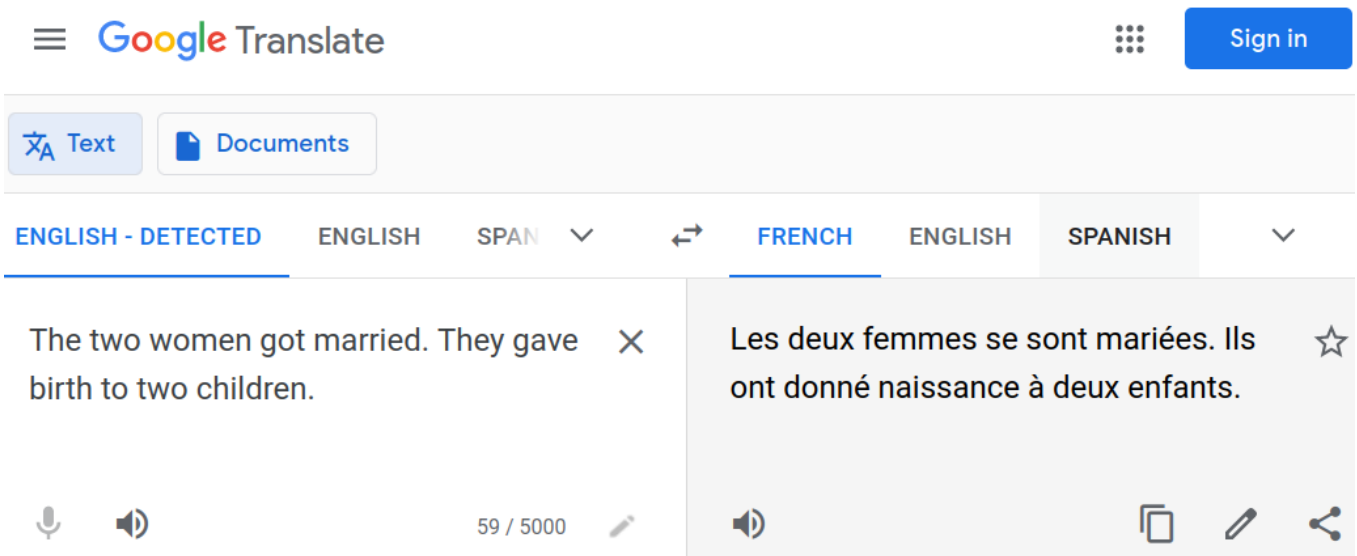


Example of issue: "Neutralization" bias



Example of issue: "Neutralization" bias





context taken into account (sentence) + masculine = neutral

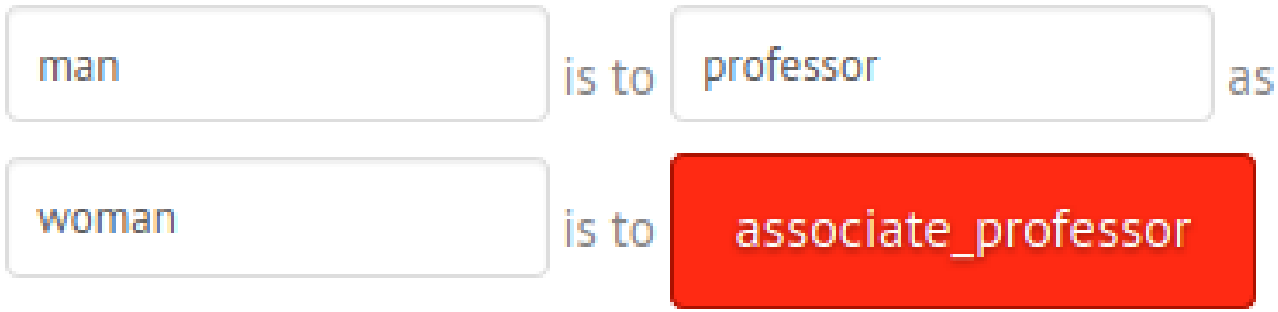
Machine learning is not magic

The decisions to:

- define masculine as neutral in French
- take the sentence as the context

were MADE by people

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

Mirror or amplifier?

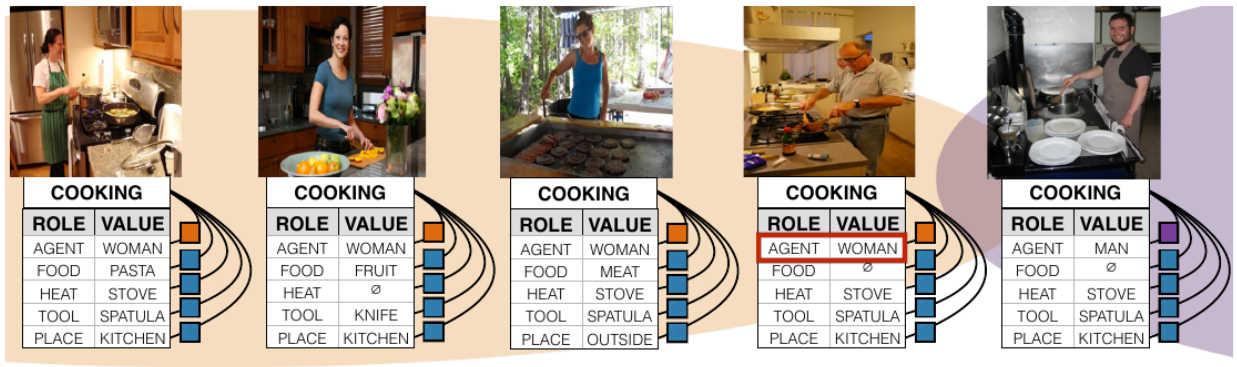


Figure 1: Five example images from the imSitu visual semantic role labeling (vSRL) dataset. Each image is paired with a table describing a situation: the verb, `cooking`, its semantic roles, i.e `agent`, and noun values filling that role, i.e. `woman`. In the imSitu training set, 33% of `cooking` images 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

https://unesdoc.unesco.org/ark:/48223/pf0000367416.page=85

	SIRI	CORTANA	ALEXA	GOOGLE ASSISTANT
Mainstream release date?	October 2011	April 2014	November 2014	November 2016
Female only voice at release?	Yes	Yes	Yes	Yes
Date fully functioning male option was added?	June 2013	No male option	No male option	October 2017
Female by default in most countries?	Yes	Yes	Yes	Yes
Male by default?	Only when the operating system language is set to Arabic, French, Dutch or British English	No	No	No
Descriptions of assistants' personalities by company representatives	'Sense of helpfulness and camaraderie, spunky without being sharp, happy without being cartoonish'	'Supportive, helpful, friendly, empathetic'	'Smart, humble, sometimes funny'	'Humble, it's helpful, a little playful at times'

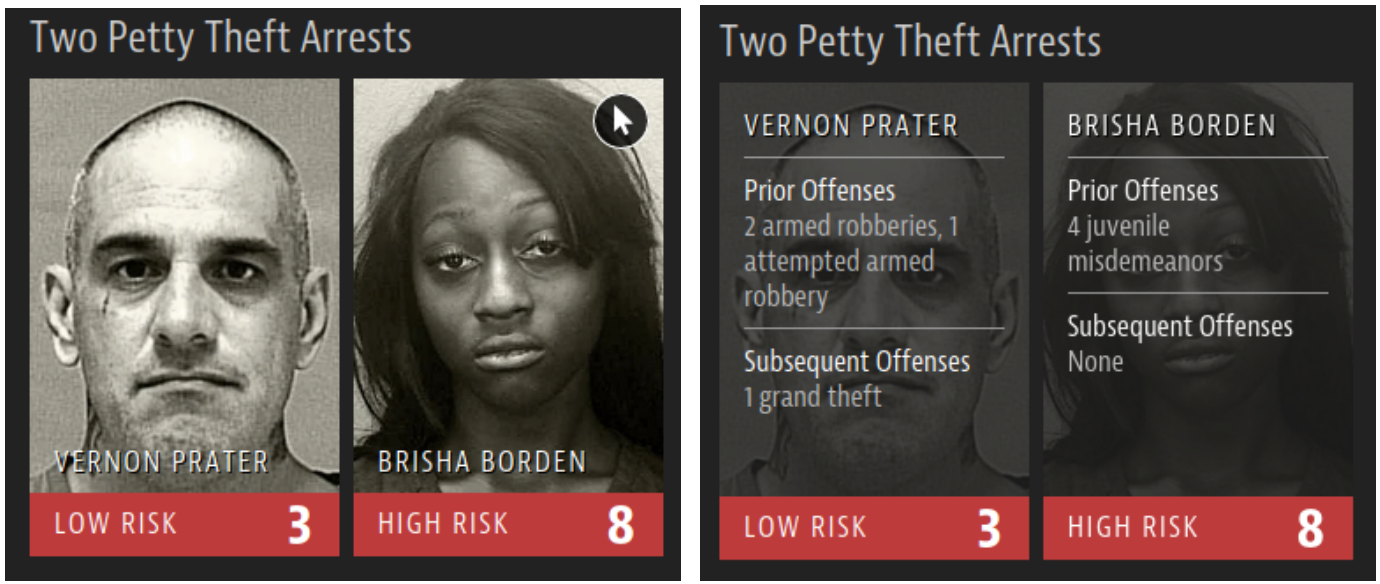
Programmed reactions

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

Justice (*risk assessment instruments*)

systems used in all the states in the USA

Example of COMPAS (2016)



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
https://epic.org/algorithmic-transparency/crim-justice/

Recruiting

"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."

https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

About the past

"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

Accueil › Espace presse

Invitation à la journée « Intelligence artificielle : l'ordinateur passe la barrière de la langue »

04 janvier 2021 NUMÉRIQUE

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

Issues in systems' evaluation

Pratiques d'évaluation en ASR et biais de performance

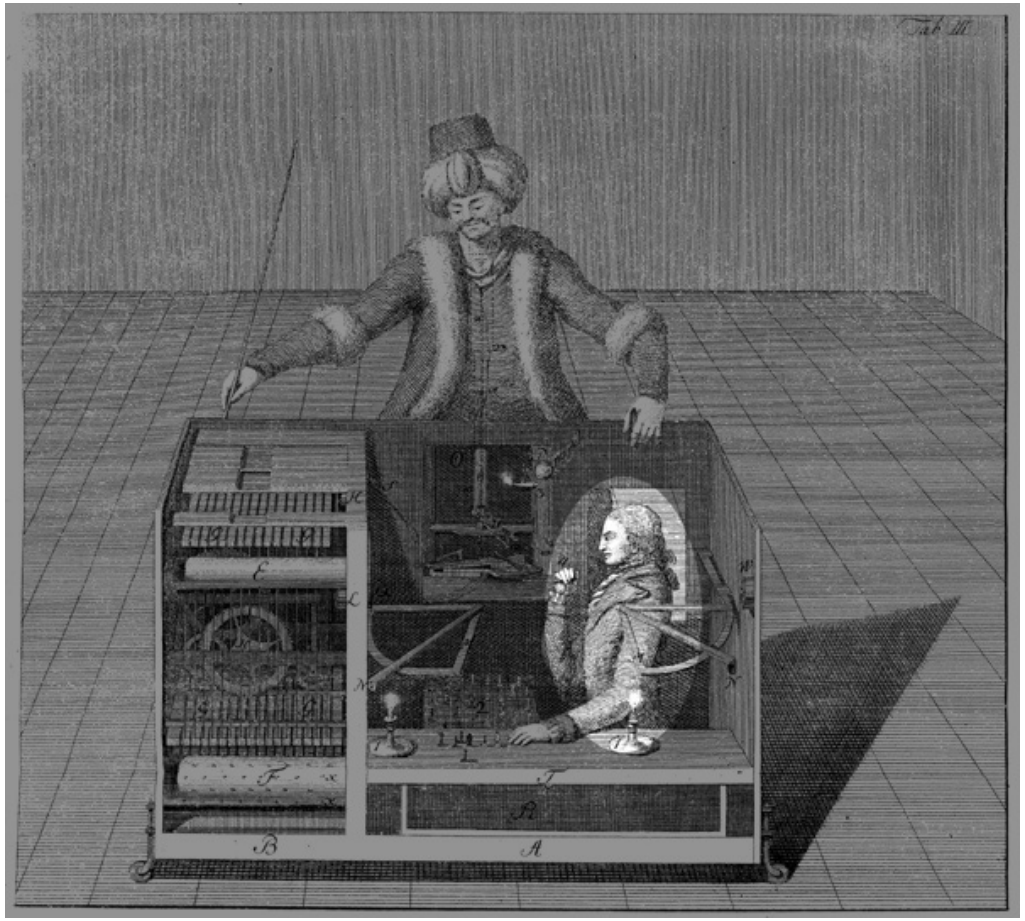
Mahault Garnerin[1, 2] Solange Rossato[2] Laurent Besacier[2]
(1) LIDILEM, Univ. Grenoble Alpes, FR-38000 Grenoble, France
(2) LIG, Univ. Grenoble Alpes, CNRS, Grenoble INP, FR-38000 Grenoble, France
prenom.nom@univ-grenoble-alpes.fr

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.

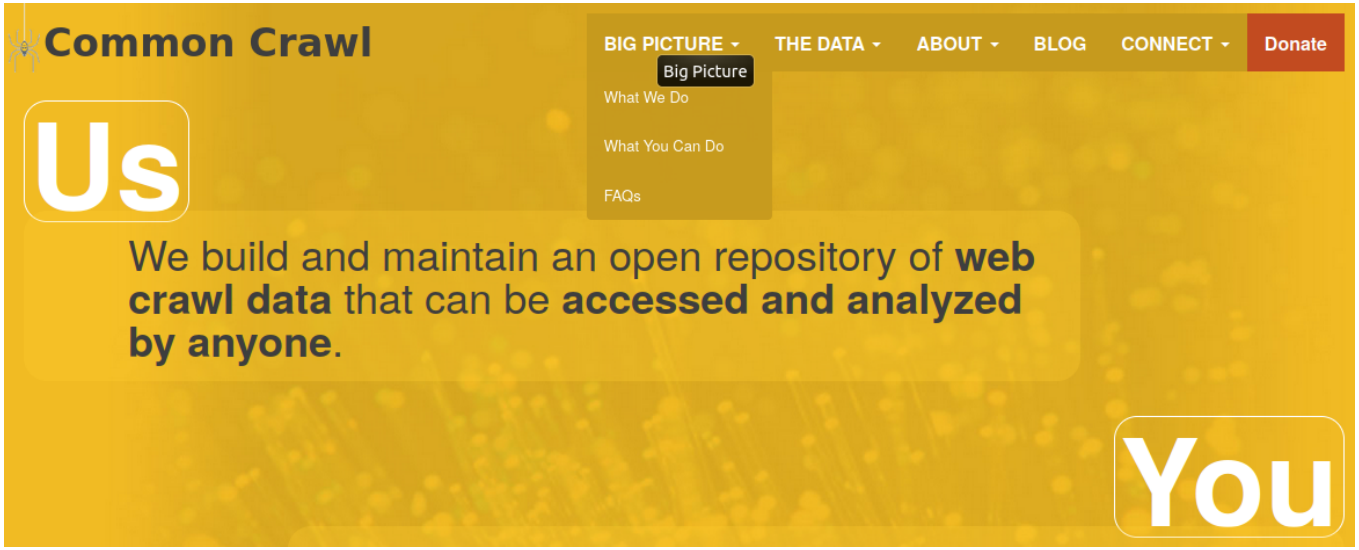
[Garnerin et al., 2020]

Data production: real humans behind the curtain



[Fort et al., 2011]

Data and "informed" consent



Carbon footprint

Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
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?

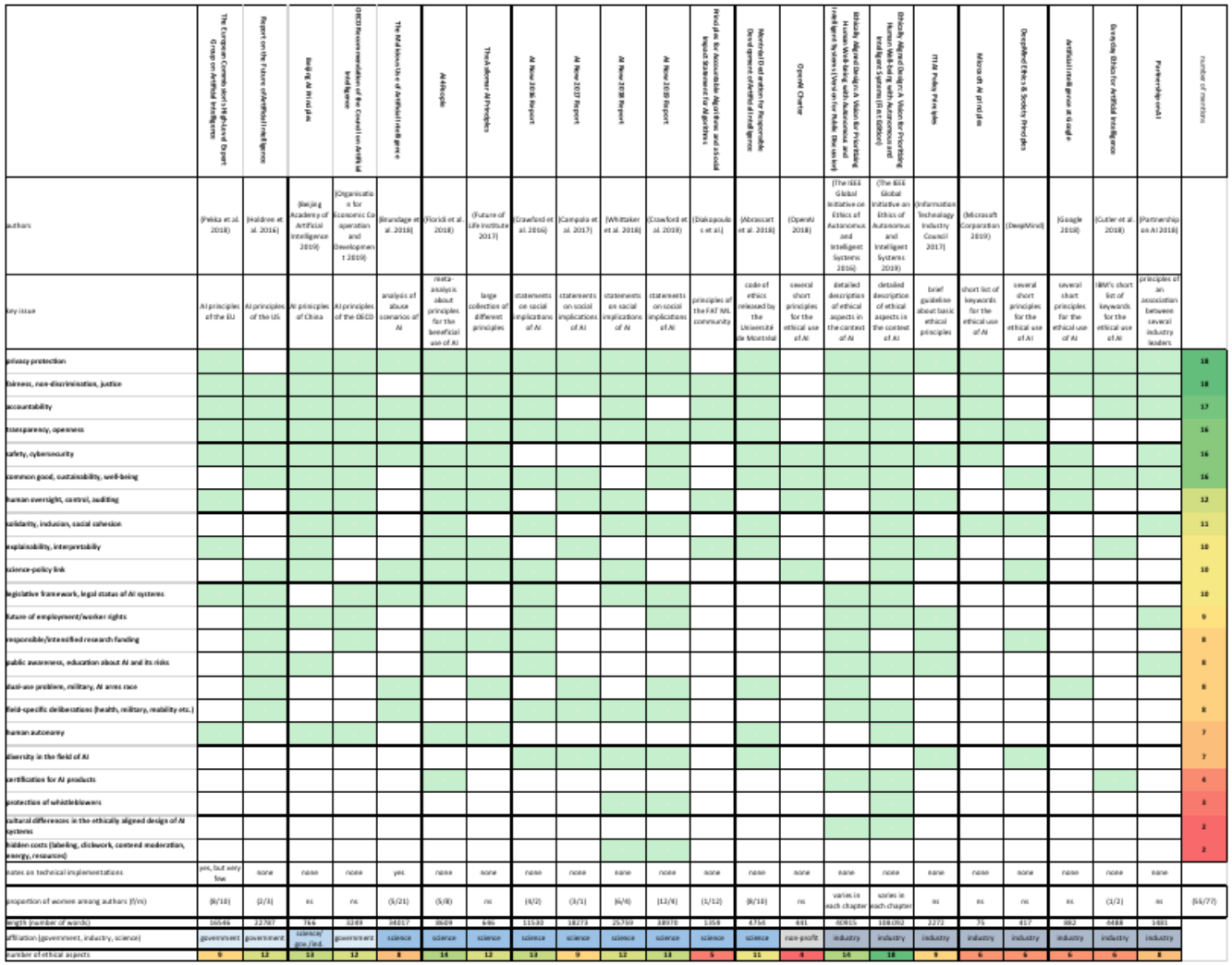
Beyond biases

Hopes?

Thanks

Guidelines, guidelines everywhere!

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



	The European Commission's High-level Expert Group on Artificial Intelligence	Report on the Future of Artificial Intelligence	Beijing AI Principles	OECD Recommendation of the Council on Artificial Intelligence	The Malicious Use of Artificial Intelligence	AI4People	The Montreal Declaration	Tenets	AI Now 2016 Report	AI Now 2017 Report	AI Now 2018 Report	AI Now 2019 Report	Principles for Accountable Algorithms and a Social Impact Statement for Algorithms	Montreal Declaration for Responsible Development of Artificial Intelligence	OpenAI Charter	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (Version for Public Discussion)	Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems (First Edition)	ITI AI Policy Principles	Microsoft AI principles	DeepMind Ethics & Society Principles	Google AI principles	Ilse Media Ethics for Artificial Intelligence	Partnership on AI	number of mentions
authors	(Pekka et al. 2018)	(Haldane et al. 2016)	(Beijing Academy of Artificial Intelligence 2019)	(Organisation for Economic Co-operation and Development 2019)	(Brundage et al. 2018)	(Floridi et al. 2018)	(Future of Life Institute 2017)	(Crawford et al. 2016)	(Campolo et al. 2017)	(Whittaker et al. 2018)	(Crawford et al. 2019)	(Diakopoulos et al.)	(Abrassart et al. 2018)	(OpenAI 2018)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2016)	(The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems 2019)	(Information Technology Industry Council 2017)	(Microsoft Corporation 2019)	(DeepMind)	(Google 2018)	(Cutler et al. 2018)	(Partnership on AI 2018)	
key issue	AI principles of the EU	AI principles of the US	AI principles of China	AI principles of the OECD	analysis of abuse scenarios of AI	meta-analysis about principles for the beneficial use of AI	large collection of different principles	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	statements on social implications of AI	principles of the FAT ML community	code of ethics released by the Université de Montréal	several short principles for the ethical use of AI	detailed description of ethical aspects in the context of AI	detailed description of ethical aspects in the context of AI	brief guideline about basic ethical principles	short list of keywords for the ethical use of AI	several short principles for the ethical use of AI	several short principles for the ethical use of AI	IBM's short list of keywords for the ethical use of AI	principles of an association between several industry leaders	
privacy protection	■		■	■		■	■	■	■	■	■	■	■	■		■	■	■	■		■	■	■	18
fairness, non-discrimination, justice	■	■	■	■		■	■	■	■	■	■	■	■	■		■	■	■	■		■	■	■	18
accountability	■	■	■	■	■	■	■			■		■	■	■		■	■	■	■			■		17
transparency, openness	■	■	■	■	■		■	■	■	■	■	■	■	■		■	■		■	■				16
safety, cybersecurity	■	■	■	■	■	■	■				■	■	■	■		■	■	■	■		■	■	■	16
common good, sustainability, well-being		■	■	■		■	■	■	■		■		■	■		■	■		■	■	■	■		16
human oversight, control, auditing	■		■	■		■	■		■	■	■	■	■	■		■	■	■			■			12
solidarity, inclusion, social cohesion			■	■		■		■		■	■	■	■			■		■	■				■	11
explainability, interpretability	■		■			■	■		■			■	■			■	■					■		10
science-policy link		■	■			■	■	■		■	■		■	■		■	■			■				10
legislative framework, legal status of AI systems	■	■		■	■			■	■	■	■				■	■								10
future of employment/worker rights		■	■	■				■	■		■				■	■							■	9
responsible/intensified research funding		■		■		■	■	■							■		■		■					8
public awareness, education about AI and its risks		■	■			■	■	■							■	■							■	8
dual-use problem, military, AI arms race		■			■		■		■	■	■	■	■		■	■				■				8
field-specific deliberations (health, military, mobility etc.)					■		■	■	■	■	■	■	■		■	■								8
human autonomy	■		■	■		■	■						■			■								7
diversity in the field of AI								■	■	■	■	■	■				■		■					7
certification for AI products						■									■	■								4
protection of whistleblowers											■	■												3
cultural differences in the ethically aligned design of AI systems																■	■							2
hidden costs (labeling, clickwork, content moderation, energy, resources)											■	■												2
notes on technical implementations	yes, but very few	none	none	none	yes	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	none	
proportion of women among authors (f/m)	(6/18)	(2/3)	m	m	(5/21)	(5/8)	m	(4/2)	(3/1)	(6/4)	(12/4)	(1/12)	science	(8/10)	m	varies in each chapter	varies in each chapter	m	m	m	m	(1/2)	m	(56/77)
length (number of words)	16646	22787	766	3189	14857	8609	4486	11530	18273	25709	18978	1358	4756	441	40815	168092	2272	75	417	382	6488	1681		
affiliation (government, industry, science)	government	government	science/ gov./ind.	government	science	science	science	science	science	science	science	science	science	non-profit	industry	industry	industry	industry	industry	industry	industry	industry	industry	
number of ethical aspects	9	12	13	12	8	14	12	13	9	12	13	5	11	6	14	18	9	6	6	6	6	8		

[Hagendorff, 2020]

Guidelines and checklists are great, but won't fix this

"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)

https://twitter.com/dantley/status/1307432466441859072

Improving systems' evaluation

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]

(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!

Bender, E. M., Gebru, T., McMillan-Major, A., and Shmitchell, S. (2021).
On the dangers of stochastic parrots: Can language models be too big?
In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, pages 610–623, New York, NY, USA. Association for Computing Machinery.

Bender, E. M. and Koller, A. (2020).
Climbing towards NLU: On meaning, form, and understanding in the age of data.
In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Blodgett, S. L., Barocas, S., DauméIII, H., and Wallach, H. (2020).
Language (technology) is power: A critical survey of "bias" in nlp.
In *ACL*.


Fort, K., Adda, G., and Cohen, K. B. (2011).
Amazon Mechanical Turk: Gold mine or coal mine?
Computational Linguistics (editorial), 37(2):413–420.


Fort, K. and Amblard, M. (2018).
Éthique et traitement automatique des langues.
In *Journée éthique et intelligence artificielle*, Nancy, France.


Garnerin, M., Rossato, S., and Besacier, L. (2020).
Pratiques d'évaluation en ASR et biais de performance.
In Adda, G., Amblard, M., and Fort, K., editors, *2e atelier Éthique et TRaitemeNt Automatique des Langues (ETeRNAL)*, pages 1–9, Nancy, France. ATALA.


Hagendorff, T. (2020).
The ethics of ai ethics: An evaluation of guidelines.
Minds & Machines, 30:99–120.


Hovy, D. and Spruit, S. L. (2016).
The social impact of natural language processing.

In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 591–598, Berlin, Germany. Association for Computational Linguistics.

Lefeuvre, A., Antoine, J.-Y., and Allegre, W. (2015).
Ethique conséquentialiste et traitement automatique des langues : une typologie de facteurs de risques adaptée aux technologies langagières.
In *Atelier Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), conférence TALN'2015*, Actes de la 1e Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), Caen (France), pages 53–66, Caen, France.

Strubell, E., Ganesh, A., and McCallum, A. (2019).
Energy and policy considerations for deep learning in NLP.
In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.


Zhao, J., Wang, T., Yatskar, M., Ordonez, V., and Chang, K.-W. (2017).
Men also like shopping: Reducing gender bias amplification using corpus-level constraints.
In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.