



# Ethics in NLP: Beyond Biases

Karën Fort

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

Malta – May 6th, 2022

## Beyond biases

Advertizing vs publishing

Artificial artificial intelligence

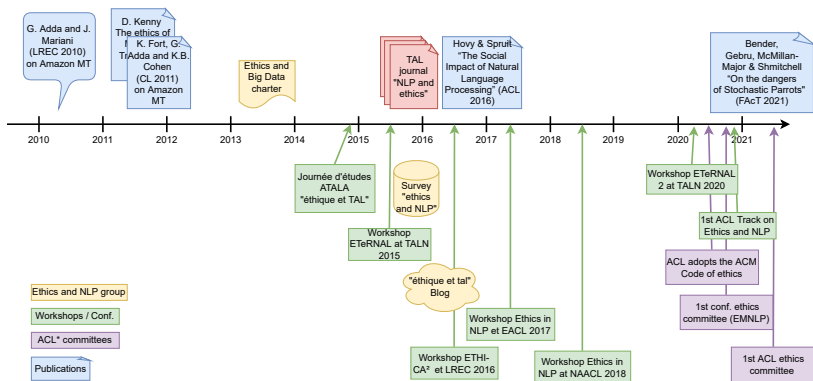
Environmental impact (in a nutshell)

"All your data are belong to us"

What about guidelines?

To finish

# Where are we (ethically speaking)?



## 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)

## Beyond biases

Advertizing vs publishing

Artificial artificial intelligence

Environmental impact (in a nutshell)

"All your data are belong to us"

What about guidelines?

To finish

# ”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

## Beyond biases

Advertizing vs publishing

**Artificial artificial intelligence**

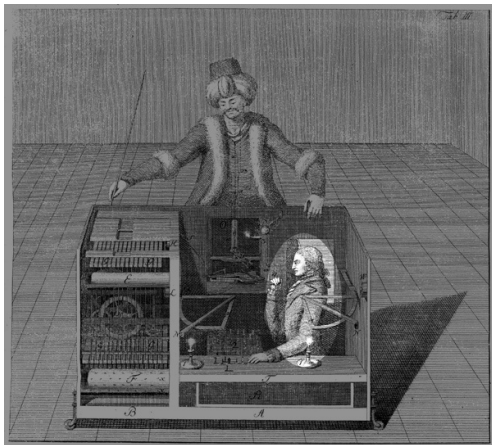
Environmental impact (in a nutshell)

"All your data are belong to us"

What about guidelines?

To finish

# Data production: real humans behind the curtain



[Fort et al., 2011]



# Data and "informed" consent

The image shows a screenshot of the Common Crawl website. The background is a solid yellow color with a subtle pattern of small, lighter yellow dots. At the top left, the text "Common Crawl" is displayed in a bold, black, sans-serif font. To the right of this, a horizontal navigation bar contains several menu items: "BIG PICTURE -", "THE DATA -", "ABOUT -", "BLOG", "CONNECT -", and "Donate". The "BIG PICTURE -" item is highlighted with a dark yellow background, and a dropdown menu is visible below it, containing the items "Big Picture", "What We Do", "What You Can Do", and "FAQs". On the left side of the page, the word "Us" is written in a large, white, rounded font inside a white-bordered rounded rectangle. In the center of the page, there is a white-bordered rounded rectangle containing the text: "We build and maintain an open repository of **web crawl data** that can be **accessed and analyzed** by anyone." On the right side of the page, the word "You" is written in a large, white, rounded font inside a white-bordered rounded rectangle.

## Beyond biases

Advertizing vs publishing

Artificial artificial intelligence

Environmental impact (in a nutshell)

"All your data are belong to us"

What about guidelines?

To finish

# Carbon footprint

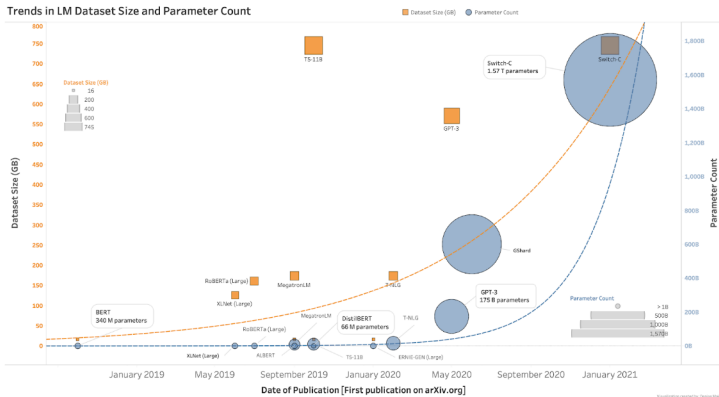
<b>Consumption</b>	<b>CO<sub>2</sub>e (lbs)</b>
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
<b>Training one model (GPU)</b>	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO<sub>2</sub> emissions from training common NLP models, compared to familiar consumption.<sup>1</sup>

[Strubell et al., 2019]

# Models trained once and for all?

Trends in LM Dataset Size and Parameter Count



[Bender et al., 2021]

## Beyond biases

"All your data are belong to us"

- Data in NLP

- Definition

- What Happens to Data?

- Back to Consent

What about guidelines?

To finish

## Beyond biases

"All your data are belong to us"

Data in NLP

Definition

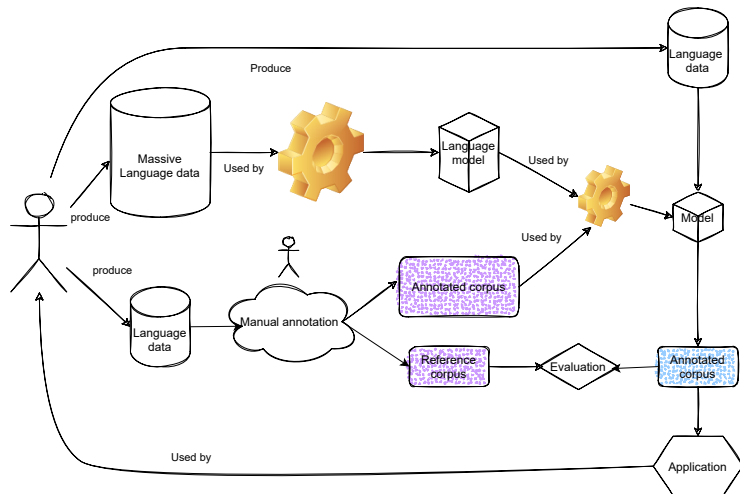
What Happens to Data?

Back to Consent

What about guidelines?

To finish

# Today's NLP



# Why it's important!



**Ben Hamner**  @benhamner · Oct 9



Programming: 10% writing code. 90% figuring out why it doesn't work

Analyzing data and ML: 1% writing code. 9% figuring out why code doesn't work. 90% figuring out what's wrong with the data



89



1.9K



8.7K





Beyond biases

"All your data are belong to us"

Data in NLP

**Definition**

What Happens to Data?

Back to Consent

What about guidelines?

To finish



SINCE 1828

JOIN MWU | GAMES | BROWSE THESAURUS | WORD OF THE DAY | VIDEO | W

data

DICTIONARY

THESAURUS

# data

noun, plural in form but singular or plural in construction, often attributive

da-ta | \ 'dā-tə,  'da- also 'dä-  \

## Definition of *data*

- 1 : factual information (such as measurements or statistics) used as a basis for reasoning, discussion, or calculation  
*//* the *data* is plentiful and easily available  
— H. A. Gleason, Jr.  
*//* comprehensive *data* on economic growth have been published  
— N. H. Jacoby
- 2 : information in digital form that can be transmitted or processed
- 3 : information output by a sensing device or organ that includes both useful and irrelevant or redundant information and must be processed to be meaningful

## Art. 4 GDPR

# Definitions

---

For the purposes of this Regulation:

- (1) 'personal data' means any information relating to an identified or identifiable natural person ('data subject'); an identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person;

<https://gdpr-info.eu/art-4-gdpr/>

# Sensitive Data

specifically protected ?

Art. 9 GDPR

## Processing of special categories of personal data

---

1. Processing of personal data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation shall be prohibited.

<https://gdpr-info.eu/art-9-gdpr/>

## Sensitive Data: exceptions

2. Paragraph 1 shall not apply if one of the following applies:
  - (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject;
  - (b) processing is necessary for the purposes of carrying out the obligations and exercising specific rights of the controller or of the data subject in the field of employment and social security and social protection law in so far as it is authorised by Union or Member State law or a collective agreement pursuant to Member State law providing for appropriate safeguards for the fundamental rights and the interests of the data subject;
  - (c) processing is necessary to protect the vital interests of the data subject or of another natural person where the data subject is physically or legally incapable of giving consent;

<https://gdpr-info.eu/art-9-gdpr/>

## Sensitive Data: exceptions again

- (d) processing is carried out in the course of its legitimate activities with appropriate safeguards by a foundation, association or any other not-for-profit body with a political, philosophical, religious or trade union aim and on condition that the processing relates solely to the members or to former members of the body or to persons who have regular contact with it in connection with its purposes and that the personal data are not disclosed outside that body without the consent of the data subjects;
- (e) processing relates to personal data which are manifestly made public by the data subject;
- (f) processing is necessary for the establishment, exercise or defence of legal claims or whenever courts are acting in their judicial capacity;
- (g) processing is necessary for reasons of substantial public interest, on the basis of Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject;

<https://gdpr-info.eu/art-9-gdpr/>

## Sensitive Data: exceptions again again

- (h) processing is necessary for the purposes of preventive or occupational medicine, for the assessment of the working capacity of the employee, medical diagnosis, the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union or Member State law or pursuant to contract with a health professional and subject to the conditions and safeguards referred to in paragraph 3;
- (i) processing is necessary for reasons of public interest in the area of public health, such as protecting against serious cross-border threats to health or ensuring high standards of quality and safety of health care and of medicinal products or medical devices, on the basis of Union or Member State law which provides for suitable and specific measures to safeguard the rights and freedoms of the data subject, in particular professional secrecy;

<https://gdpr-info.eu/art-9-gdpr/>

## Sensitive Data: exceptions again again again

- (j) processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes in accordance with [Article 89\(1\)](#) based on Union or Member State law which shall be proportionate to the aim pursued, respect the essence of the right to data protection and provide for suitable and specific measures to safeguard the fundamental rights and the interests of the data subject.

<https://gdpr-info.eu/art-9-gdpr/>



Beyond biases

**"All your data are belong to us"**

Data in NLP

Definition

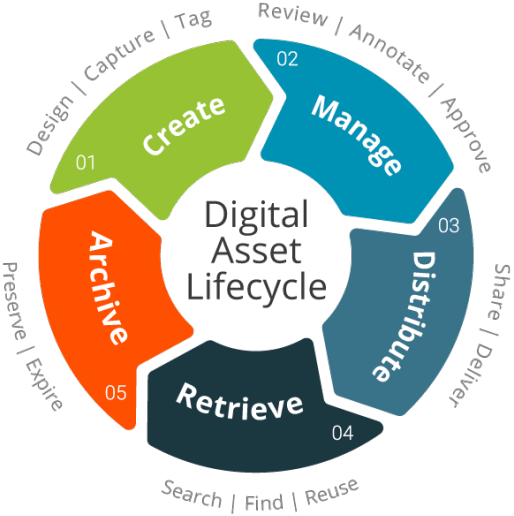
**What Happens to Data?**

Back to Consent

What about guidelines?

To finish

# Data Lifecycle



## Beyond biases

"All your data are belong to us"

Data in NLP

Definition

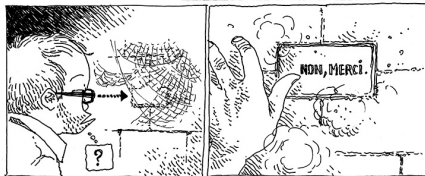
What Happens to Data?

Back to Consent

What about guidelines?

To finish

# What Consent Means (or not), by ©Boulet



# Informed Consent

The Nuremberg Code (1947) states that consent can be voluntary **only if**:

- ▶ participants are **able** to consent
- ▶ they are **free from coercion**
- ▶ they **comprehend** the risks and benefits involved

# Art. 7 GDPR: Conditions for consent (1/2)

## Art. 7 GDPR Conditions for consent

---

1. Where processing is based on consent, the controller shall be able to demonstrate that the data subject has consented to processing of his or her personal data.
2. <sup>1</sup> If the data subject's consent is given in the context of a written declaration which also concerns other matters, the request for consent shall be presented in a manner which is clearly distinguishable from the other matters, in an intelligible and easily accessible form, using clear and plain language. <sup>2</sup> Any part of such a declaration which constitutes an infringement of this Regulation shall not be binding.

<https://gdpr-info.eu/art-7-gdpr/>

## Art. 7 GDPR: Conditions for consent (2/2)

- <sup>1</sup> The data subject shall have the right to withdraw his or her consent at any time. <sup>2</sup> The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. <sup>3</sup> Prior to giving consent, the data subject shall be informed thereof. <sup>4</sup> It shall be as easy to withdraw as to give consent.
- When assessing whether consent is freely given, utmost account shall be taken of whether, *inter alia*, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.

<https://gdpr-info.eu/art-7-gdpr/>

# Consequences in Practice

There is **no** consent if no decision is made:

- ▶ opt in vs opt out
- ▶ importance of the default settings
- ▶ possibility to withdraw one's consent at anytime



The screenshot shows the top navigation bar of the Grosbill website. On the left is a 'MENU' icon. The logo 'Grosbill.com' is prominently displayed in red, with the tagline 'Le meilleur de l'High-Tech' below it. A search bar contains the placeholder text 'Produit, marque, référence...' and a green search button. To the right is a 'MAGASIN' icon with a location pin. Below the navigation bar is a cookie consent banner with the title 'Accepter ou refuser les cookies'. The banner contains the text 'Désactiver les cookies à vocation commerciale :' followed by a toggle switch that is currently turned off.

<https://www.grosbill.com/>



# Exercise

Exercise your "ethical" brain

Go on the [Facebook homepage](#) (disconnect if you are connected) and evaluate to what extent the consent of a new user is informed.

Beyond biases

"All your data are belong to us"

**What about guidelines?**

Beware of guidelines

Ethics guidelines for trustworthy AI

To finish

Beyond biases

"All your data are belong to us"

**What about guidelines?**

Beware of guidelines

Ethics guidelines for trustworthy AI

To finish

# Guidelines, guidelines everywhere!

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

	The European Code of Ethics Guidelines on Robotics (2017)	European Code of Ethics on Robotics (2017)	OECD Guidelines for the Governance of Artificial Intelligence (2019)	IEEE Global AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)	AI Ethics Principles (2019)																										
Transparency																																																						
Privacy																																																						
Autonomy																																																						
Accountability																																																						
Non-discrimination																																																						
Human well-being																																																						
Human rights																																																						
Human dignity																																																						
Human agency																																																						
Human diversity																																																						
Human autonomy																																																						
Human agency																																																						
Human dignity																																																						
Human autonomy																																																						
Human diversity																																																						
Human agency																																																						
Human dignity																																																						
Human autonomy																																																						
Human diversity																																																						

[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]*

# Beyond Guidelines

Guidelines and checklists are attractive:

- ▶ simple
- ▶ illusion of exhaustiveness

But they are far from enough:

*" Neither the risk analysis informed by engineering practice, nor the socially informed engineering practice can be replaced by the other." [Gurses et al., 2011]*

# Making the Most of Guidelines

1. start thinking/discussing **without** them
2. use them as a complement
3. do not limit your thinking because you checked all the list in the grid

## Some guidelines I recommend

1. AI HLEG Ethics guidelines for trustworthy AI (EN or FR or ...)
2. The consequentialist grid of analysis [Lefeuvre et al., 2015] (FR)
3. CERNA Machine learning ethics report (FR and EN)
4. CCNE Chatbots ethics report (FR)



Beyond biases

"All your data are belong to us"

**What about guidelines?**

Beware of guidelines

Ethics guidelines for trustworthy AI

To finish

# Ethical guidelines for trustworthy AI

4 ethical principles:

1. Respect for human autonomy
2. Prevention of harm
3. Fairness
4. Explicability

+ tensions between them and decisions made should be documented and argued

## Respect for human autonomy

*"AI systems should not unjustifiably subordinate, coerce, deceive, manipulate, condition or herd humans. Instead, they should be designed to augment, complement and empower human cognitive, social and cultural skills. The allocation of functions between humans and AI systems should follow human-centric design principles and leave meaningful opportunity for human choice. **This means securing human oversight over work processes in AI systems.**"*

*"the less oversight a human can exercise over an AI system, the more extensive testing and stricter governance is required"*

## Prevention of harm

*"AI systems should neither cause nor exacerbate harm or otherwise adversely affect human beings. This entails the protection of human dignity as well as mental and physical integrity. AI systems and the environments in which they operate must be safe and secure. [...] Particular attention must also be paid to situations where AI systems can cause or exacerbate adverse impacts due to asymmetries of power or information, such as between employers and employees, businesses and consumers or governments and citizens. "*

## Fairness

*"The development, deployment and use of AI systems must be fair. [...] Moreover, the use of AI systems should never lead to people being deceived or unjustifiably impaired in their freedom of choice. Additionally, fairness implies that AI practitioners should respect the principle of proportionality between means and ends, and consider carefully how to balance competing interests and objectives. The procedural dimension of fairness entails the ability to contest and seek effective redress against decisions made by AI systems and by the humans operating them. In order to do so, **the entity accountable for the decision must be identifiable, and the decision-making processes should be explicable.**"*

# Explicability

*"This means that **processes need to be transparent**, the capabilities and purpose of AI systems openly communicated, and decisions – to the extent possible – explainable to those directly and indirectly affected. [...] The degree to which explicability is needed is highly dependent on the context and the severity of the consequences if that output is erroneous or otherwise inaccurate."*

Beyond biases

"All your data are belong to us"

What about guidelines?

**To finish**

WYHTR: What You Have To Remember



- ▶ data is everywhere in NLP
- ▶ data lifecycle and ethical hotspots
- ▶ consent, consent, consent




# Reading List

Please participate!

ACL ethics committee reading list



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.



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.



Gurses, S., Troncoso, C., and Diaz, C. (2011).

Engineering privacy by design.

In Computers, Privacy & Data Protection.



Hagendorff, T. (2020).

The ethics of ai ethics: An evaluation of guidelines.

Minds & Machines, 30:99–120.



Lefevre, 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

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.