Ethics in NLP: Beyond Biases

Karën Fort

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

LETTRES SORBONNE UNIVERSITÉ

Loria Laboratoire lorrain de recherche en informatique et ses applications

UNIVERSITÉ DE LORRAINE

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

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

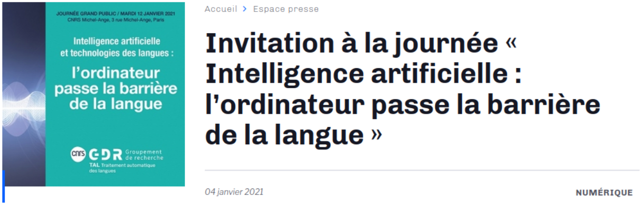
- Advertizing *vs* publishing
- Artificial artificial intelligence
- Environmental impact (in a nutshell)

”All your data are belong to us”

What about guidelines?

To finish

"Overselling" research results

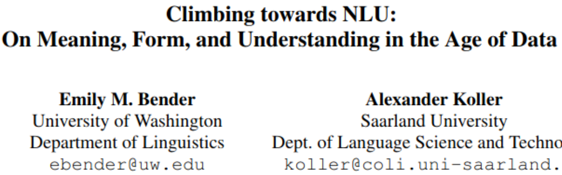

JOURNÉE GRAND PUBLIC / MARDI 12 JANVIER 2021
CNRS Michel-Ange, 3 rue Michel-Ange, Paris
Intelligence artificielle
et technologies des langues :
l'ordinateur
passe la barrière
de la langue
cnrs GDR Groupement de recherche
TAL Traitement automatique des langues

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

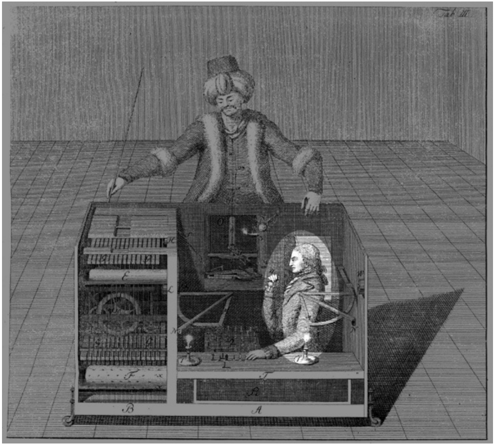
- Advertizing *vs* publishing
- Artificial artificial intelligence
- Environmental impact (in a nutshell)

"All your data are belong to us"

What about guidelines?

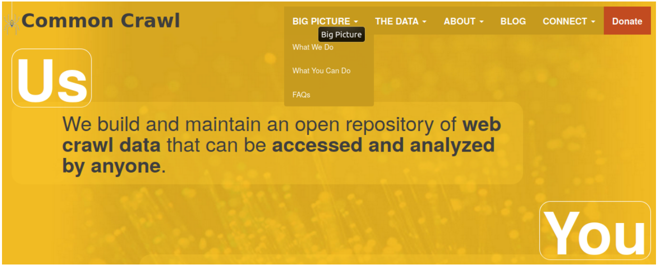
To finish

Data production: real humans behind the curtain



[Fort et al., 2011]

Data and ”informed” consent



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

Beyond biases

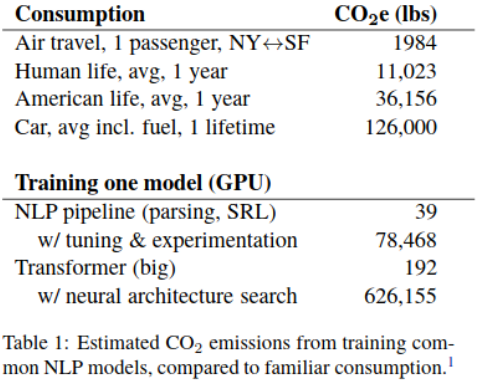
- Advertizing *vs* publishing
- Artificial artificial intelligence
- Environmental impact (in a nutshell)

"All your data are belong to us"

What about guidelines?

To finish

Carbon footprint

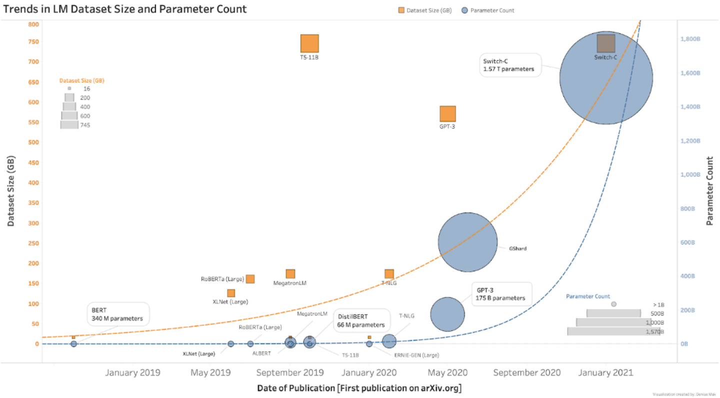


Consumption	**CO$_2$e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

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

[Strubell et al., 2019]

Models trained once and for all?



[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

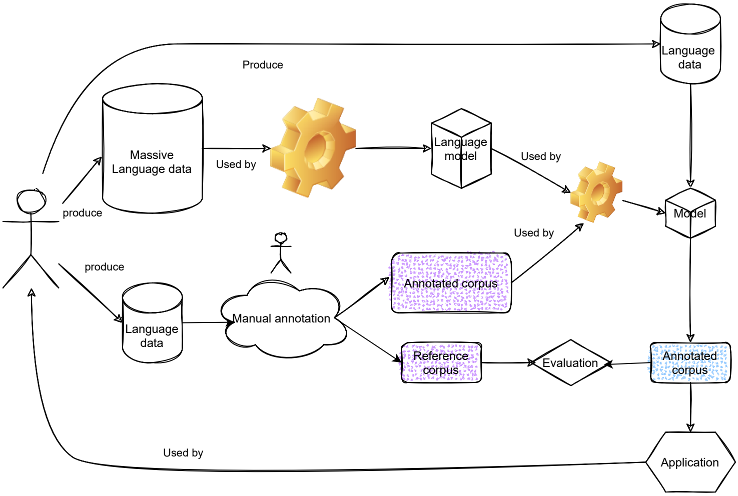
”All your data are belong to us”

- Data in NLP
- Definition
- What Happens to Data?
- Back to Consent

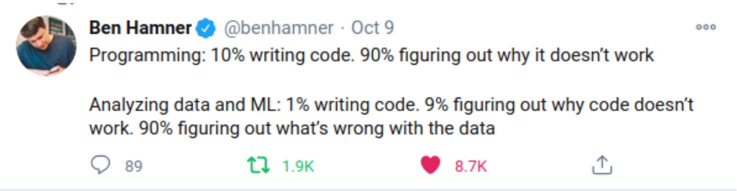
What about guidelines?

To finish

Today's NLP


Produce
Language data
Massive Language data
Used by
Language model
Used by
produce
Model
Used by
produce
Annotated corpus
Manual annotation
Language data
Reference corpus
Evaluation
Annotated corpus
Used by
Application

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

Beyond biases

”All your data are belong to us”

Data in NLP

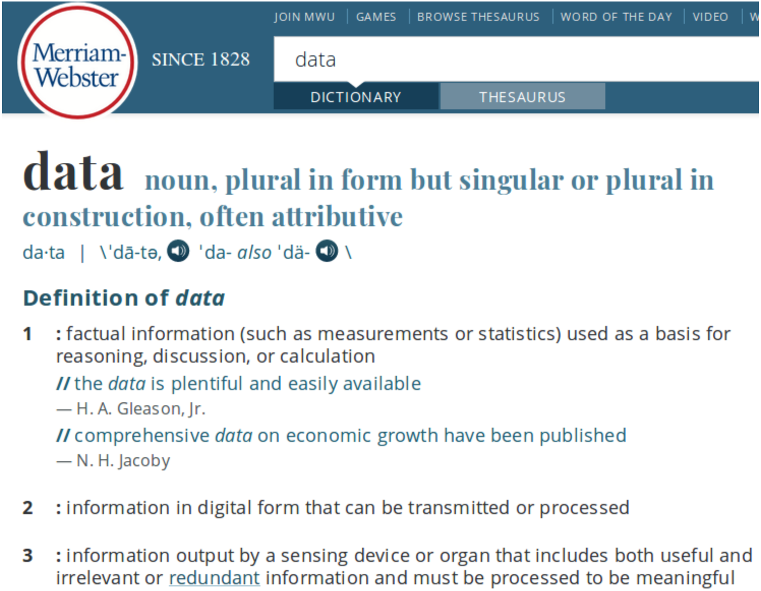
Definition

What Happens to Data?

Back to Consent

What about guidelines?

To finish



data 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.

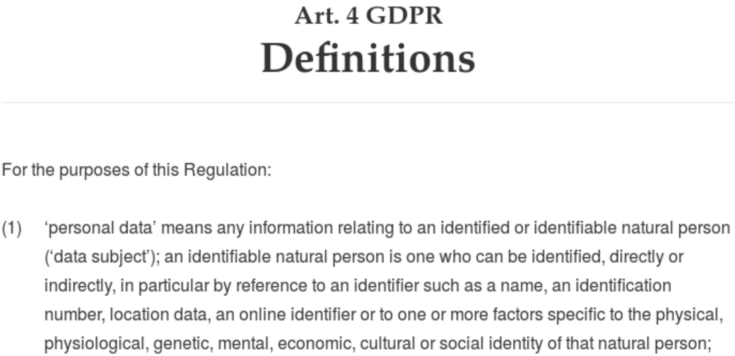
// comprehensive *data* on economic growth have been published

— 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

Personal Data



Art. 4 GDPR
Definitions

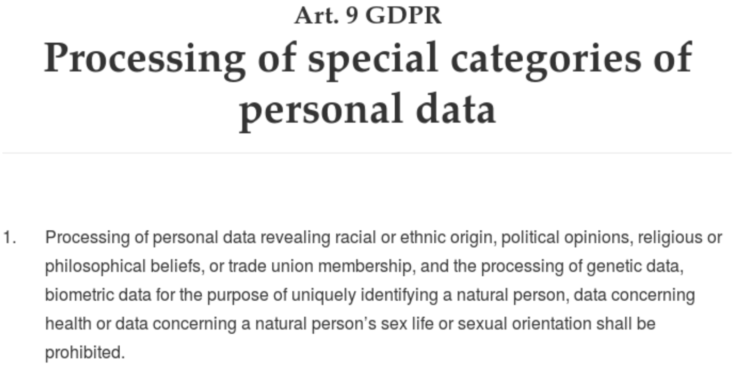
For the purposes of this Regulation:

(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 ?



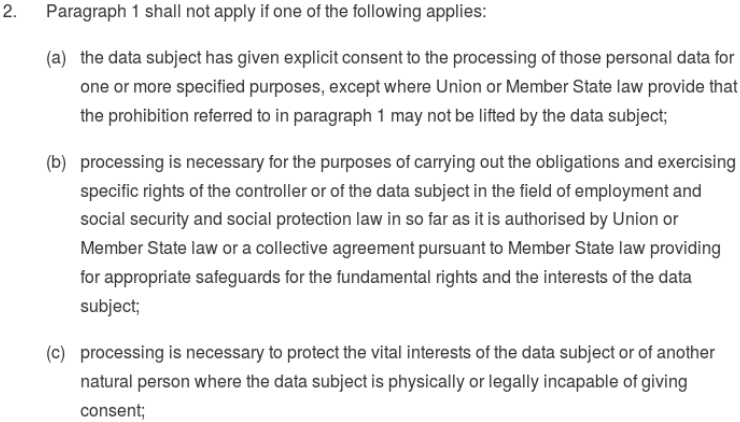
Art. 9 GDPR

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:

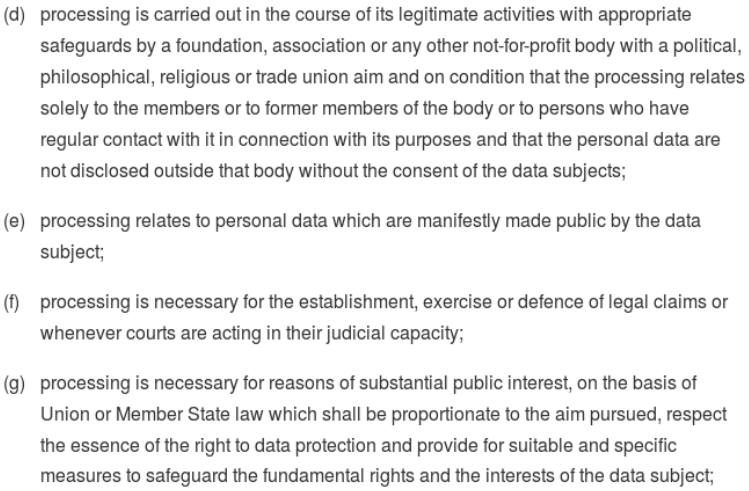
   (a) the data subject has given explicit consent to the processing of those personal data for one or more specified purposes, except where Union or Member State law provide that the prohibition referred to in paragraph 1 may not be lifted by the data subject;

   (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;

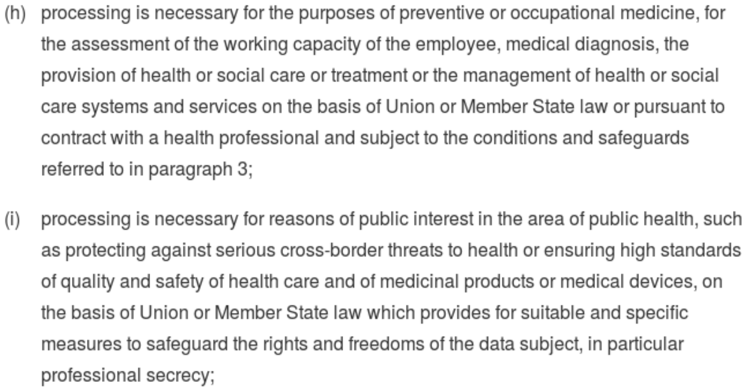
(e) processing relates to personal data which are manifestly made public by the data subject;

(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

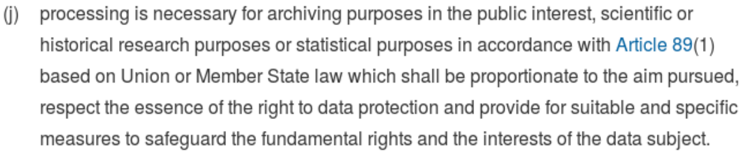


(h) processing is necessary for the purposes of preventive or occupational medicine, for the assessment of the working capacity of the employee, medical diagnosis, the provision of health or social care or treatment or the management of health or social care systems and services on the basis of Union 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

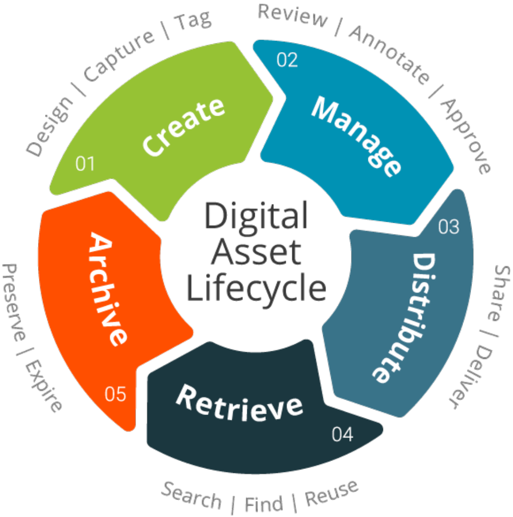
”All your data are belong to us”

- Data in NLP
- Definition
- What Happens to Data?
- Back to Consent

What about guidelines?

To finish

Data Lifecycle



Haztowichp - CC BY-SA

Beyond biases

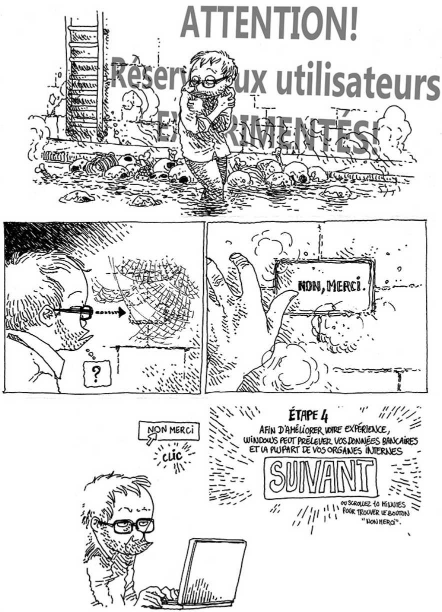
”All your data are belong to us”

- Data in NLP
- Definition
- What Happens to Data?
- Back to Consent

What about guidelines?

To finish

What Consent Means (or not), by ©Boulet

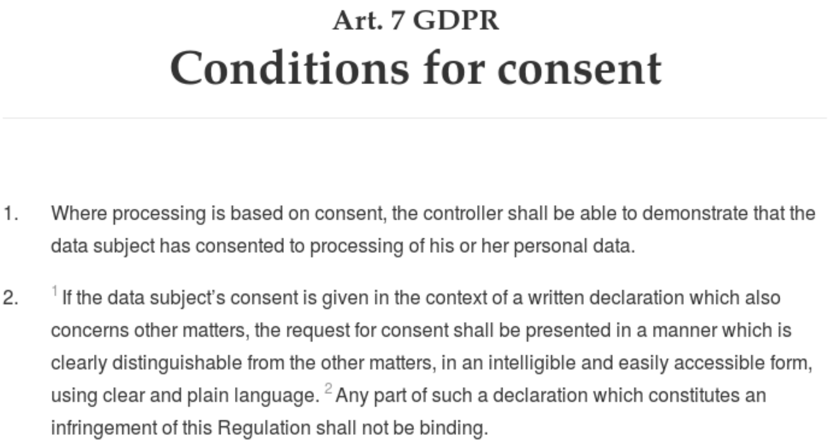

ATTENTION!
Réservé aux utilisateurs EXPÉRIMENTÉS!
?
NON, MERCI.
NON MERCI
CLIC
ÉTAPE 4
AFIN D'AMÉLIORER VOTRE EXPÉRIENCE, WINDOWS PEUT PRÉLEVER VOS DONNÉES BANCAIRES ET LA PLUPART DE VOS ORGANES INTERNES
SUIVANT
OU SCROLLEZ 10 MINUTES POUR TROUVER LE BOUTON "NON MERCI".

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

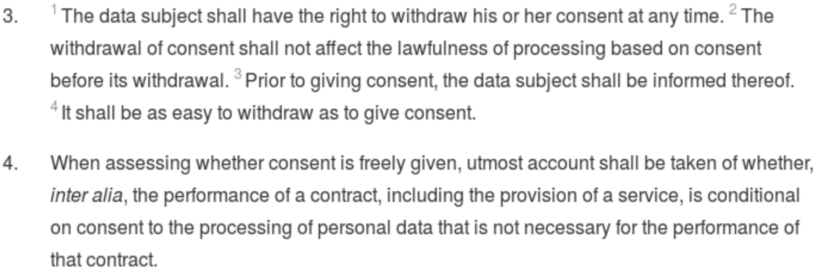
Conditions for consent

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. [1] 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. [2] 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)



3. [1] The data subject shall have the right to withdraw his or her consent at any time. [2] The withdrawal of consent shall not affect the lawfulness of processing based on consent before its withdrawal. [3] Prior to giving consent, the data subject shall be informed thereof. [4] It shall be as easy to withdraw as to give consent.

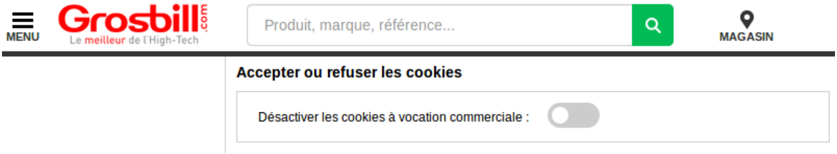
4. When assessing whether consent is freely given, utmost account shall be taken of whether, *inter alia*, the performance of a contract, including the provision of a service, is conditional on consent to the processing of personal data that is not necessary for the performance of that contract.

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



Accepter ou refuser les cookies

Désactiver les cookies à vocation commerciale : ○

https://www.grosbill.com/

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!

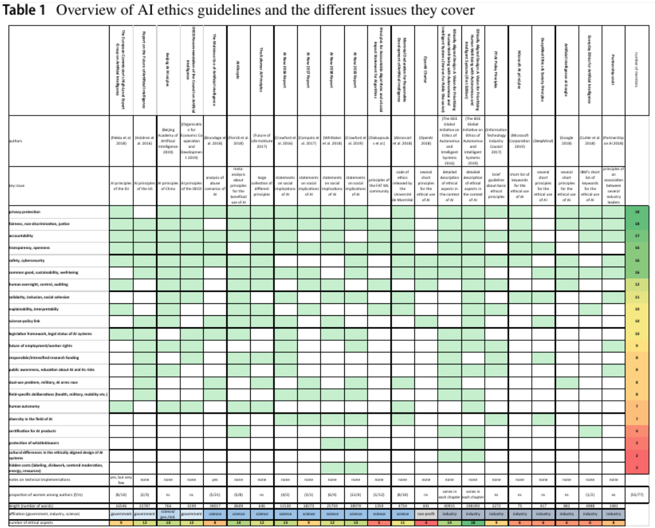


Table 1 Overview of AI ethics guidelines and the different issues they cover

	The European 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 Asilomar AI Principles	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	Montréal 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	Artificial Intelligence at Google	Everyday Ethics for Artificial Intelligence	Partnership on AI	number of mentions
authors	(Pekka et al. 2018)	(Holdren 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, interpretabiliy	🟩		🟩			🟩	🟩		🟩			🟩	🟩			🟩	🟩				🟩		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, contend 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	
proportion of women among authors (f/m)	(8/10)	(2/3)	ns	ns	(5/21)	(5/8)	ns	(4/2)	(3/1)	(6/4)	(12/4)	(1/12)	(8/10)	ns	varies in each chapter	varies in each chapter	ns	ns	ns	ns	(1/2)	ns	(55/77)
length (number of words)	16546	22787	766	3249	34017	8609	646	11530	18273	25759	38970	1359	4754	441	40915	108092	2272	75	417	882	4488	1481	
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	
number of ethical aspects	9	12	13	12	8	14	12	13	9	12	13	5	11	4	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]*

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 argumented

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."*

- 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.

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
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.