Éthique et TAL : au-delà des biais

Karën Fort

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

ETAL 2023 – Marseille

LETTRES SORBONNE UNIVERSITÉ

cnrs

Loria Laboratoire lorrain de recherche en informatique et ses applications

Au-delà des biais

- Publicité *vs* publication
- L'intelligence artificielle artificielle
- L'impact environnemental
- Le temps long
- Les conflits d'intérêts

"All your data are belong to us"

Bonnes pratiques

On en parle aussi

- l'explicabilité, mais pas l'interprétabilité [Rudin, 2019]
- le *dual use* [Hovy and Spruit, 2016], mais pas la ligne rouge à ne pas franchir
- la diversité linguistique [Joshi et al., 2020], mais encore trop peu des besoins des locuteurs [Bird, 2020]
- la documentation des données [Couillault et al., 2014] [Gebru et al., 2021] et des modèles [Mitchell et al., 2019], mais pas vraiment des droits sur les données

Très peu d'approches systémiques sur l'éthique dans le TAL

- [Lefeuvre et al., 2015] (FR) : une grille conséquentialiste complète pour une évaluation des recherches en TAL et de leurs applications
- [Fort and Amblard, 2018] (FR) : un (embryon de) vue déontologique et systémique de l'éthique dans le TAL
- [Bender et al., 2021] : à propos des dangers des gros modèles de langues

Des résultats de recherche à nuancer

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

L'überisation du travail (*microworking crowdsourcing*)

- *Crowdsourcing* est devenu synonyme de microtravail
- plus personne ne remet en question son utilisation en TAL
- alors que les problèmes soulevés dans [Fort et al., 2011] n'ont pas disparu :

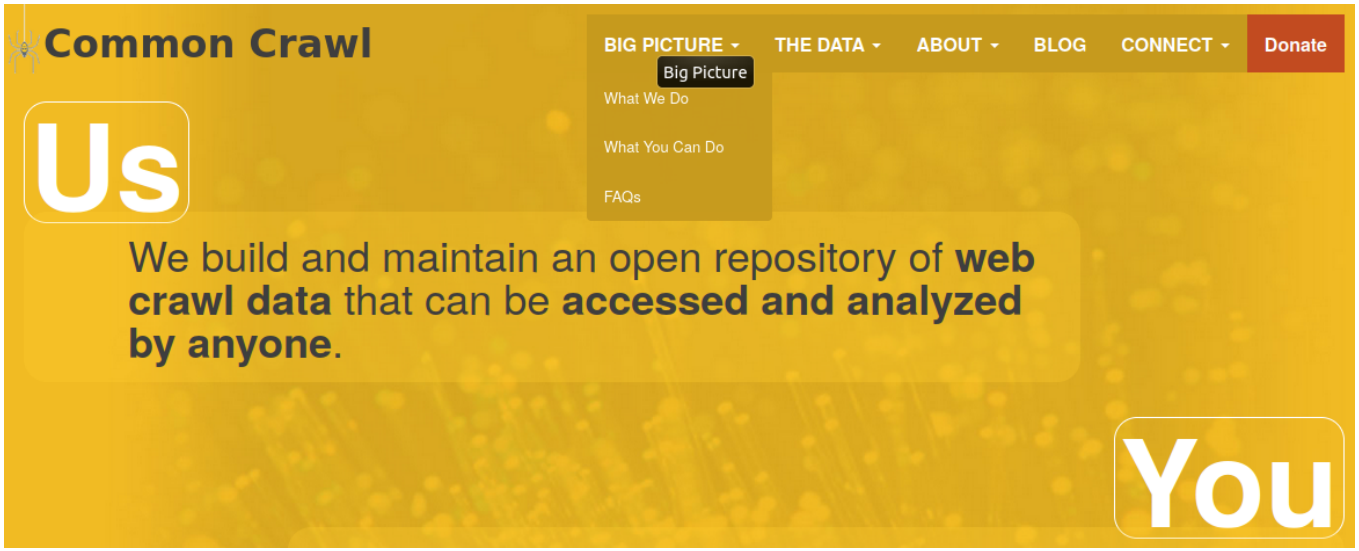
Le lien unissant un chauffeur et Uber reconnu « contrat de travail »

Le conducteur VTC avait saisi la justice en juin 2017, après que la plate-forme eut « désactivé son compte ». L'arrêt de la cour d'appel de Paris renvoie ce dossier aux prud'hommes.

Le Monde avec AFP

Publié le 11 janvier 2019 à 00h16 - Mis à jour le 11 janvier 2019 à 06h32 · Lecture 2 min.

Le consentement (éclairé)



Au-delà des biais

- Publicité *vs* publication
- L'intelligence artificielle artificielle
- L'impact environnemental
- Le temps long
- Les conflits d'intérêts

"All your data are belong to us"

Bonnes pratiques

L'empreinte carbone [Strubell et al., 2019]

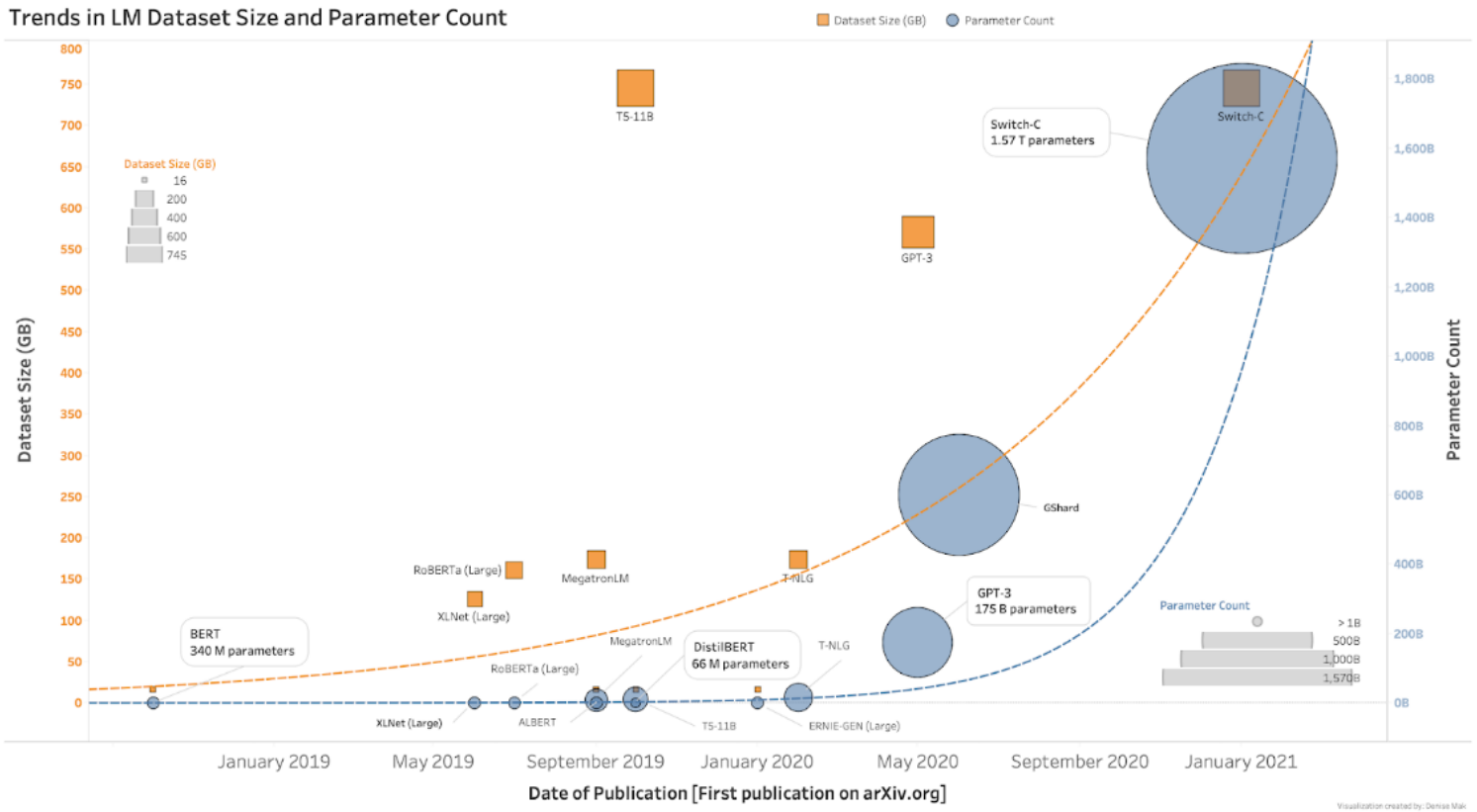
Consumption	**CO_2e (lbs)**
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture search	626,155

Table 1: Estimated CO_2 emissions from training common NLP models, compared to familiar consumption.[1]

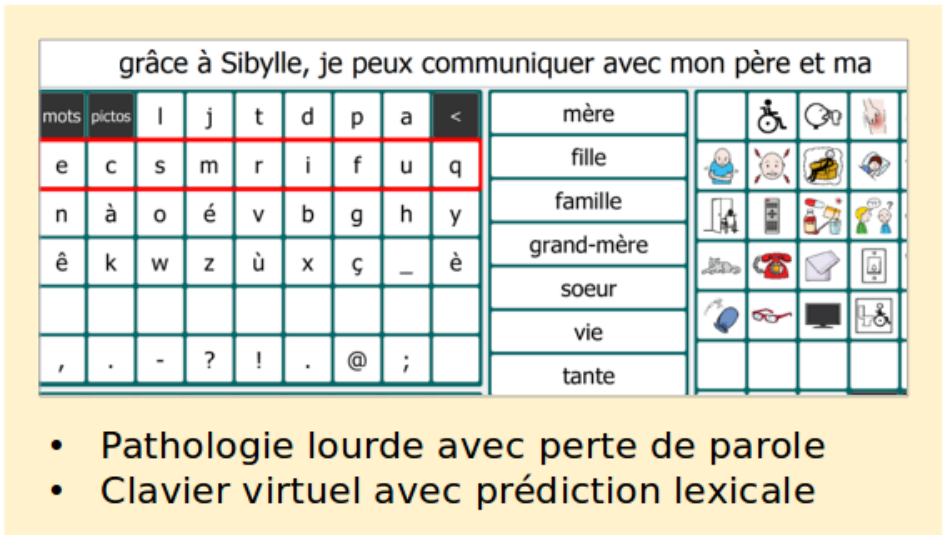
Note : ces mesures ne concernent qu'une source d'émission C02 sur quatre [Bannour et al., 2021] ⇒ largement sous-estimée

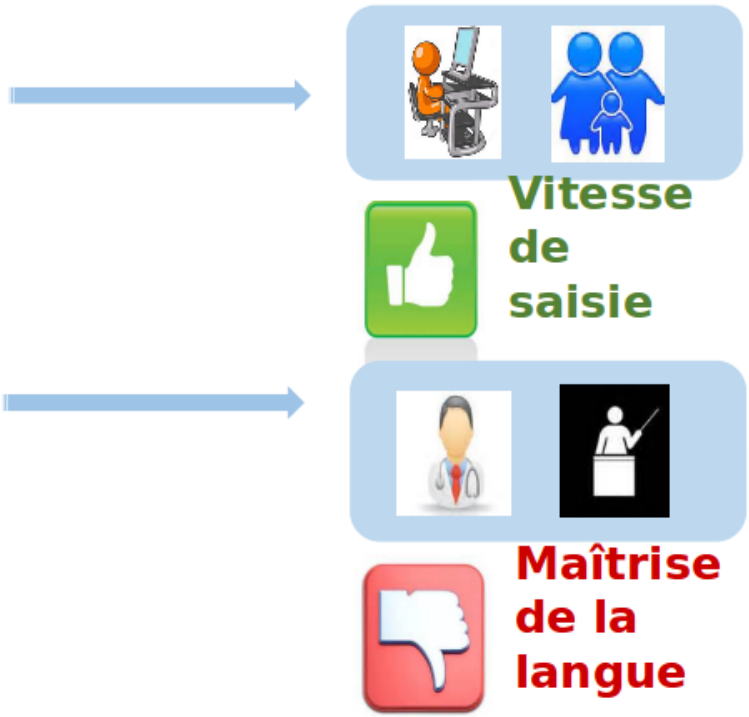
L'argument du modèle unique...

schéma issu d'une présentation de l'article [Bender et al., 2021]



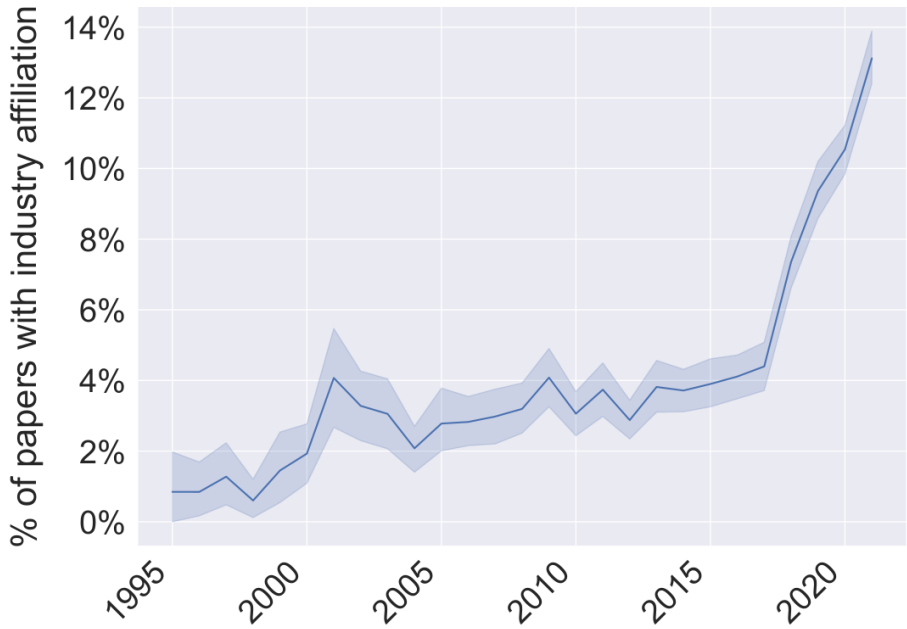
Le temps long



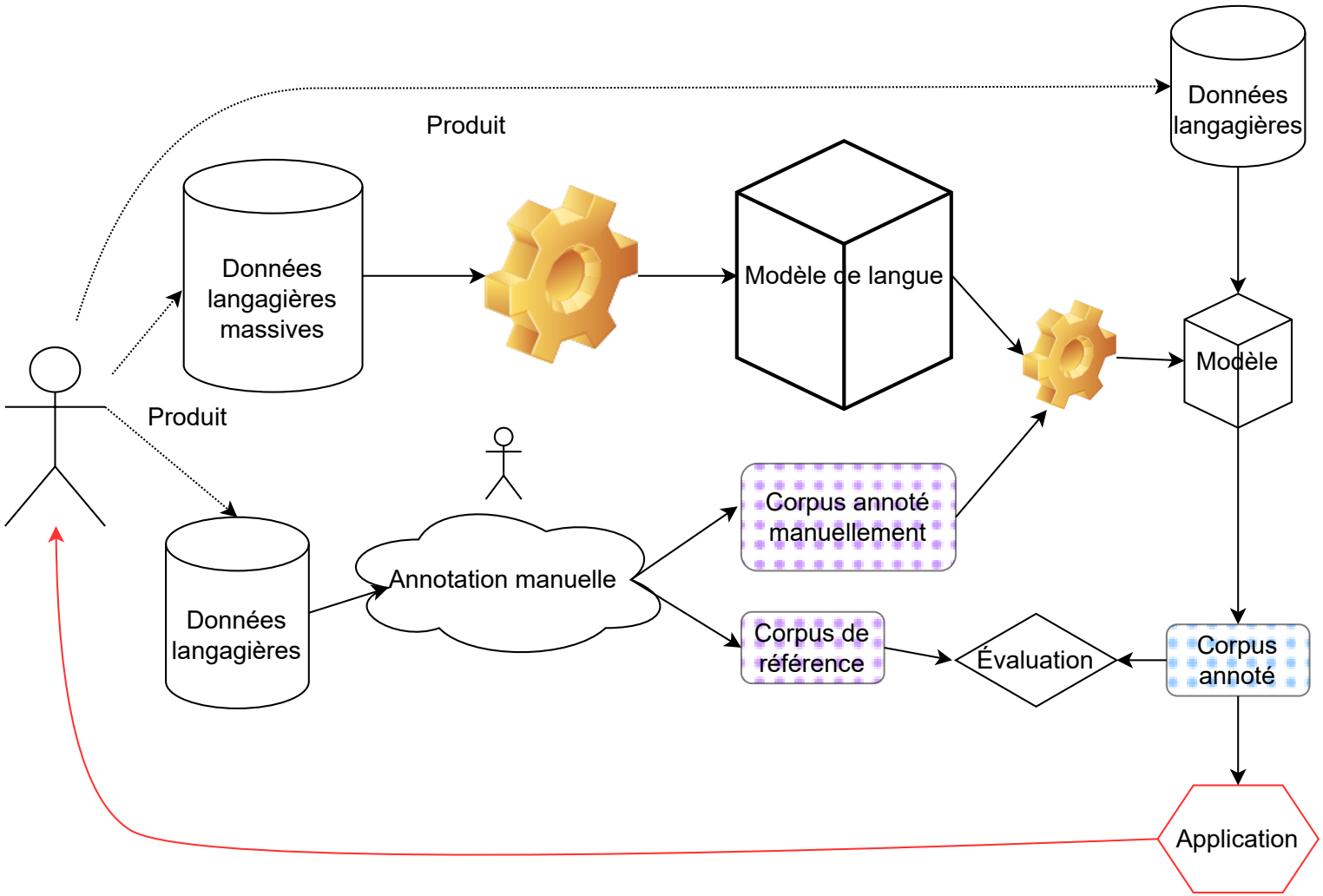


[Antoine and Lefeuvre, 2014]

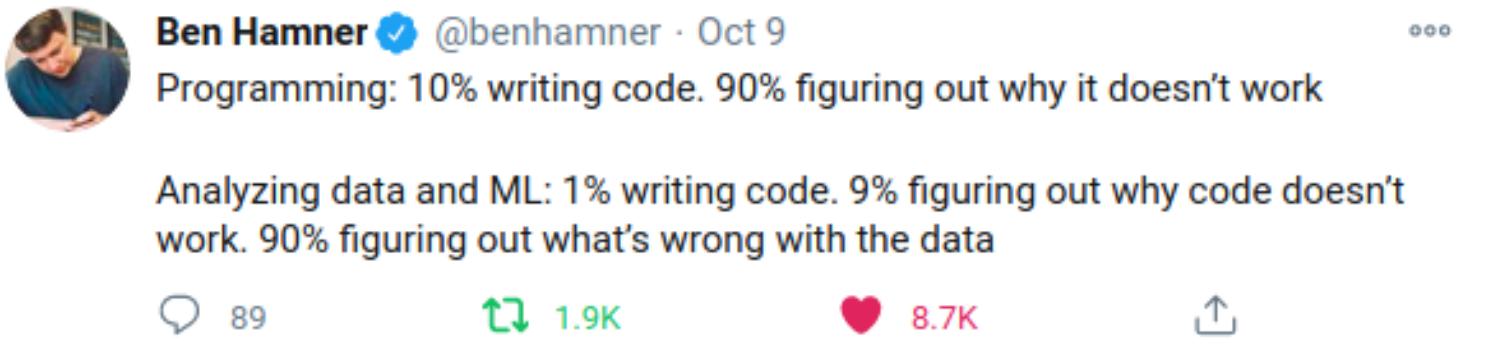
L'omniprésence des BigTech [Abdalla et al., 2023]



Le TAL aujourd'hui

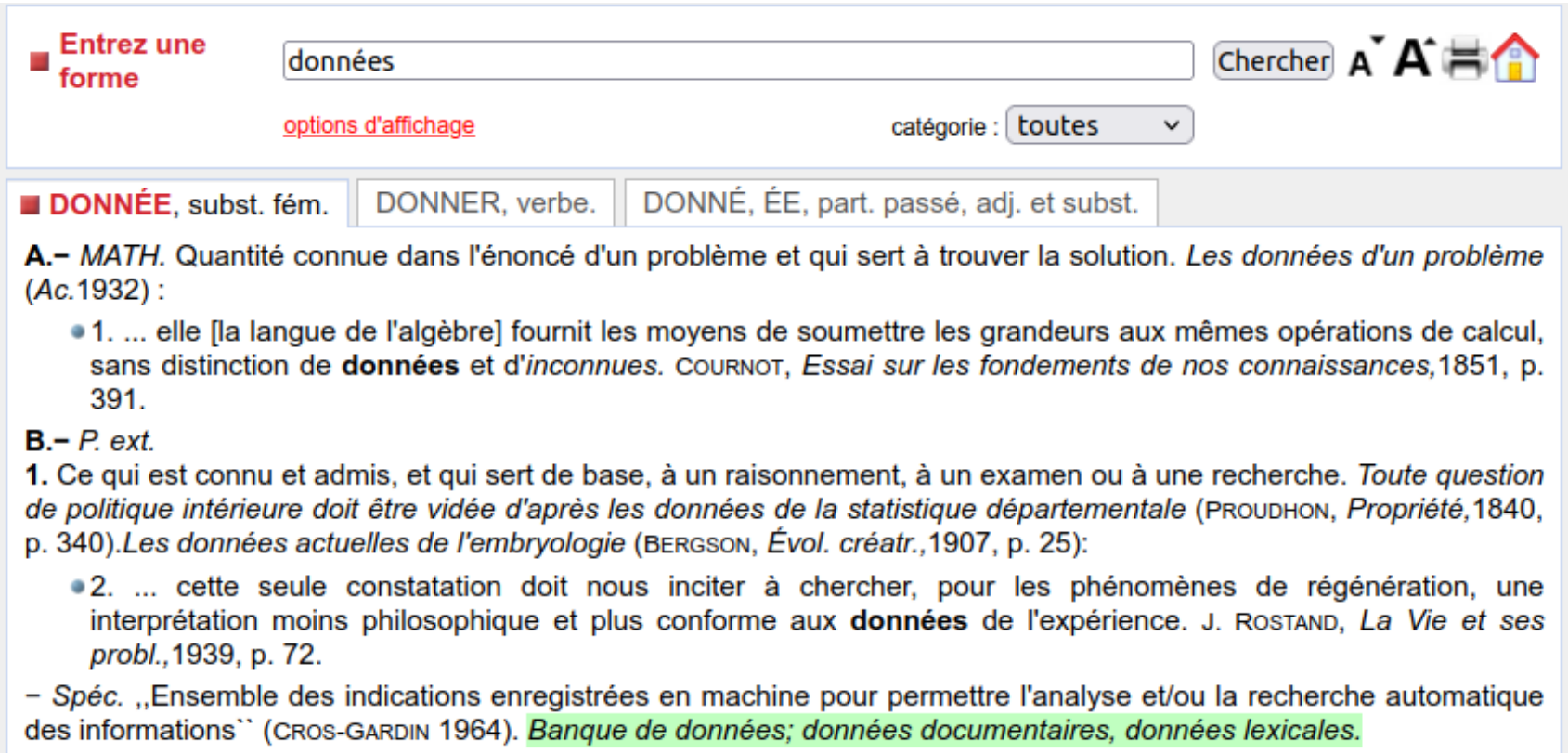


Pourquoi c'est important (rappel) !



Définition

https://www.cnrtl.fr/lexicographie/donn%C3%A9es



Entrez une forme

données

Chercher

options d'affichage

catégorie : toutes

DONNÉE, subst. fém. | DONNER, verbe. | DONNÉ, ÉE, part. passé, adj. et subst.

A.– *MATH.* Quantité connue dans l'énoncé d'un problème et qui sert à trouver la solution. *Les données d'un problème* (*Ac.*1932) :

- 1. ... elle [la langue de l'algèbre] fournit les moyens de soumettre les grandeurs aux mêmes opérations de calcul, sans distinction de **données** et d'*inconnues*. COURNOT, *Essai sur les fondements de nos connaissances,*1851, p. 391.

B.– *P. ext.*

1. Ce qui est connu et admis, et qui sert de base, à un raisonnement, à un examen ou à une recherche. *Toute question de politique intérieure doit être vidée d'après les données de la statistique départementale* (PROUDHON, *Propriété,*1840, p. 340).*Les données actuelles de l'embryologie* (BERGSON, *Évol. créatr.,*1907, p. 25):

- 2. ... cette seule constatation doit nous inciter à chercher, pour les phénomènes de régénération, une interprétation moins philosophique et plus conforme aux **données** de l'expérience. J. ROSTAND, *La Vie et ses probl.,*1939, p. 72.

– *Spéc.* ,,Ensemble des indications enregistrées en machine pour permettre l'analyse et/ou la recherche automatique des informations`` (CROS-GARDIN 1964). *Banque de données; données documentaires, données lexicales.*

Données à caractère personnel

Article 4 - Définitions

Aux fins du présent règlement, on entend par :

1. «données à caractère personnel», toute information se rapportant à une personne physique identifiée ou identifiable (ci-après dénommée «personne concernée») ; est réputée être une «personne physique identifiable» une personne physique qui peut être identifiée, directement ou indirectement, notamment par référence à un identifiant, tel qu'un nom, un numéro d'identification, des données de localisation, un identifiant en ligne, ou à un ou plusieurs éléments spécifiques propres à son identité physique, physiologique, génétique, psychique, économique, culturelle ou sociale;

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1#Article4

Données à caractère personnel spécifiquement protégées

Article 9 - Traitement portant sur des catégories particulières de données à caractère personnel

1. Le traitement des données à caractère personnel qui révèle l'origine raciale ou ethnique, les opinions politiques, les convictions religieuses ou philosophiques ou l'appartenance syndicale, ainsi que le traitement des données génétiques, des données biométriques aux fins d'identifier une personne physique de manière unique, des données concernant la santé ou des données concernant la vie sexuelle ou l'orientation sexuelle d'une personne physique sont interdits.

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2

Données à caractère personnel spécifiquement protégées : exceptions

a) la personne concernée a donné son consentement explicite au traitement de ces données à caractère personnel pour une ou plusieurs finalités spécifiques, sauf lorsque le droit de l'Union ou le droit de l'État membre prévoit que l'interdiction visée au paragraphe 1 ne peut pas être levée par la personne concernée;

b) le traitement est nécessaire aux fins de l'exécution des obligations et de l'exercice des droits propres au responsable du traitement ou à la personne concernée en matière de droit du travail, de la sécurité sociale et de la protection sociale, dans la mesure où ce traitement est autorisé par le droit de l'Union, par le droit d'un État membre ou par une convention collective conclue en vertu du droit d'un État membre qui prévoit des garanties appropriées pour les droits fondamentaux et les intérêts de la personne concernée;

c) le traitement est nécessaire à la sauvegarde des intérêts vitaux de la personne concernée ou d'une autre personne physique, dans le cas où la personne concernée se trouve dans l'incapacité physique ou juridique de donner son consentement;

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2

Données à caractère personnel spécifiquement protégées : exceptions (encore)

d) le traitement est effectué, dans le cadre de leurs activités légitimes et moyennant les garanties appropriées, par une fondation, une association ou tout autre organisme à but non lucratif et poursuivant une finalité politique, philosophique, religieuse ou syndicale, à condition que ledit traitement se rapporte exclusivement aux membres ou aux anciens membres dudit organisme ou aux personnes entretenant avec celui-ci des contacts réguliers en liaison avec ses finalités et que les données à caractère personnel ne soient pas communiquées en dehors de cet organisme sans le consentement des personnes concernées;

e) le traitement porte sur des données à caractère personnel qui sont manifestement rendues publiques par la personne concernée;

f) le traitement est nécessaire à la constatation, à l'exercice ou à la défense d'un droit en justice ou chaque fois que des juridictions agissent dans le cadre de leur fonction juridictionnelle;

g) le traitement est nécessaire pour des motifs d'intérêt public important, sur la base du droit de l'Union ou du droit d'un État membre qui doit être proportionné à l'objectif poursuivi, respecter l'essence du droit à la protection des données et prévoir des mesures appropriées et spécifiques pour la sauvegarde des droits fondamentaux et des intérêts de la personne concernée;

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2

Données à caractère personnel spécifiquement protégées : exceptions (encore)

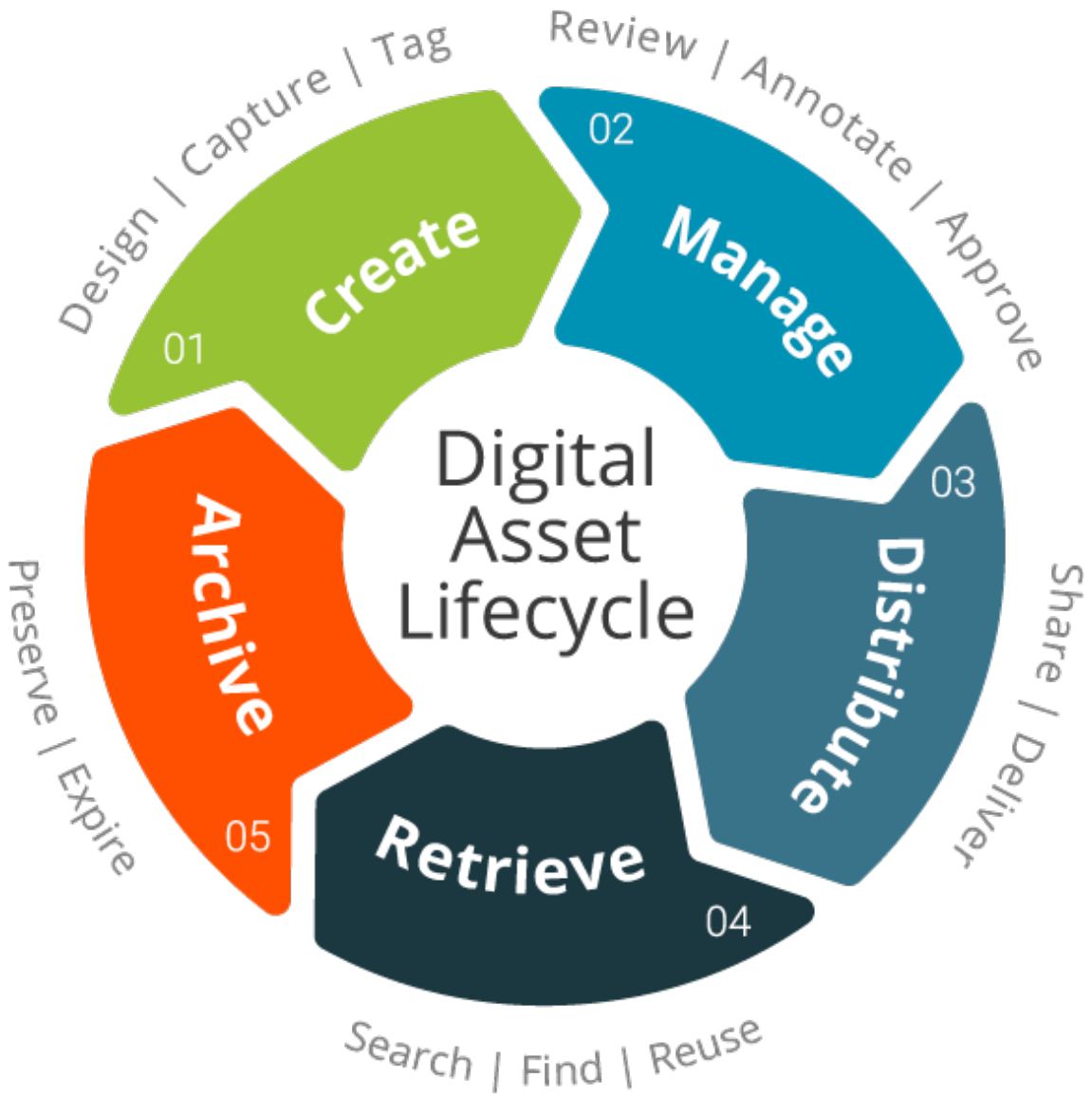
h) le traitement est nécessaire aux fins de la médecine préventive ou de la médecine du travail, de l'appréciation de la capacité de travail du travailleur, de diagnostics médicaux, de la prise en charge sanitaire ou sociale, ou de la gestion des systèmes et des services de soins de santé ou de protection sociale sur la base du droit de l'Union, du droit d'un État membre ou en vertu d'un contrat conclu avec un professionnel de la santé et soumis aux conditions et garanties visées au paragraphe 3;

i) le traitement est nécessaire pour des motifs d'intérêt public dans le domaine de la santé publique, tels que la protection contre les menaces transfrontalières graves pesant sur la santé, ou aux fins de garantir des normes élevées de qualité et de sécurité des soins de santé et des médicaments ou des dispositifs médicaux, sur la base du droit de l'Union ou du droit de l'État membre qui prévoit des mesures appropriées et spécifiques pour la sauvegarde des droits et libertés de la personne concernée, notamment le secret professionnel;

j) le traitement est nécessaire à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques, conformément à l'article 89, paragraphe 1, sur la base du droit de l'Union ou du droit d'un État membre qui doit être proportionné à l'objectif poursuivi, respecter l'essence du droit à la protection des données et prévoir des mesures appropriées et spécifiques pour la sauvegarde des droits fondamentaux et des intérêts de la personne concernée.

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2

Cycle de vie des données



Traitement des données à caractère personnel (1/2)

Article 5 - Principes relatifs au traitement des données à caractère personnel

1. Les données à caractère personnel doivent être :

> a) traitées de manière licite, loyale et transparente au regard de la personne concernée (licéité, loyauté, transparence);
>
> b) collectées pour des finalités déterminées, explicites et légitimes, et ne pas être traitées ultérieurement d'une manière incompatible avec ces finalités; le traitement ultérieur à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques n'est pas considéré, conformément à l'article 89, paragraphe 1, comme incompatible avec les finalités initiales (limitation des finalités);
>
> c) adéquates, pertinentes et limitées à ce qui est nécessaire au regard des finalités pour lesquelles elles sont traitées (minimisation des données);

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1#Article5

Traitement des données à caractère personnel (2/2)

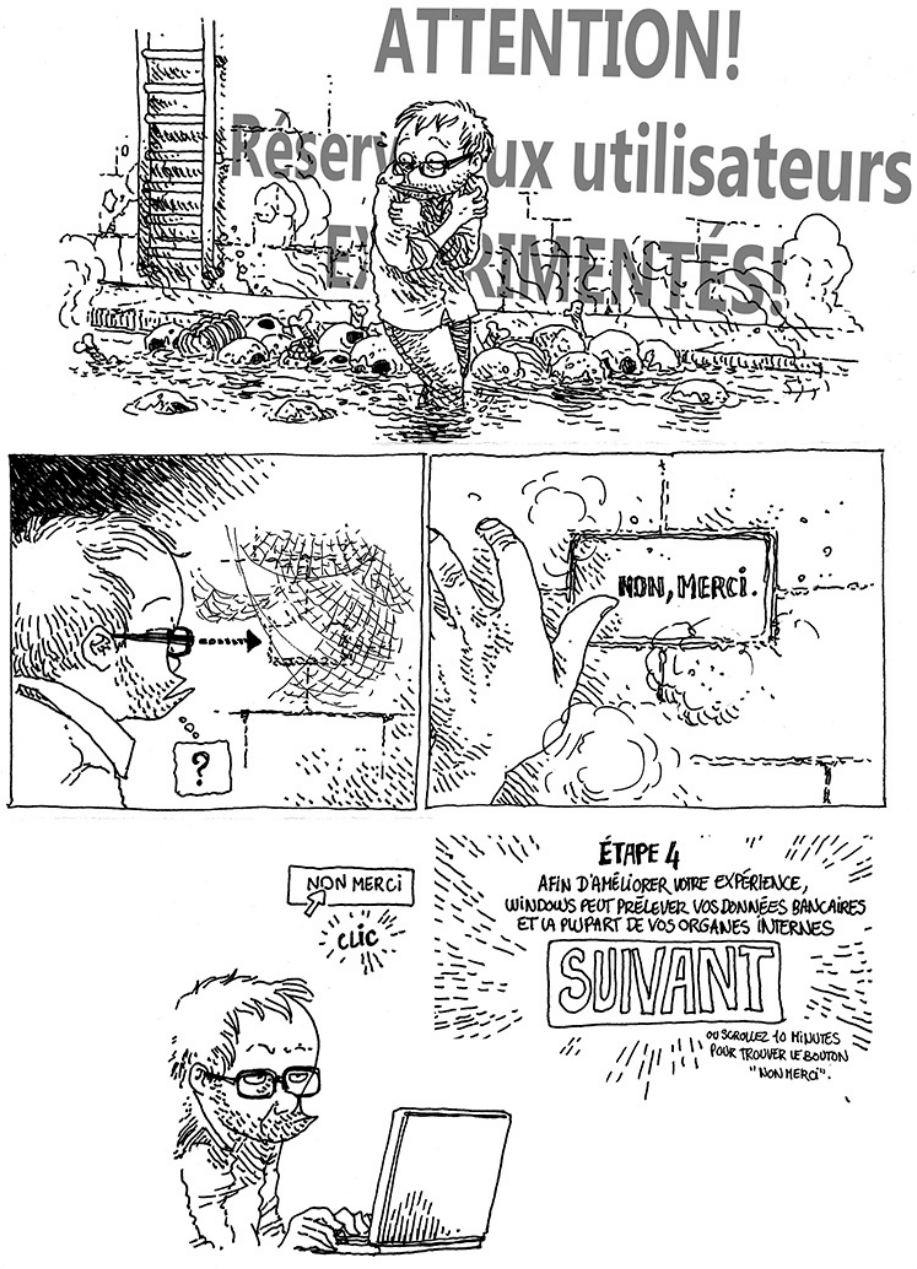
d) exactes et, si nécessaire, tenues à jour; toutes les mesures raisonnables doivent être prises pour que les données à caractère personnel qui sont inexactes, eu égard aux finalités pour lesquelles elles sont traitées, soient effacées ou rectifiées sans tarder (exactitude);

e) conservées sous une forme permettant l'identification des personnes concernées pendant une durée n'excédant pas celle nécessaire au regard des finalités pour lesquelles elles sont traitées; les données à caractère personnel peuvent être conservées pour des durées plus longues dans la mesure où elles seront traitées exclusivement à des fins archivistiques dans l'intérêt public, à des fins de recherche scientifique ou historique ou à des fins statistiques conformément à l'article 89, paragraphe 1, pour autant que soient mises en œuvre les mesures techniques et organisationnelles appropriées requises par le présent règlement afin de garantir les droits et libertés de la personne concernée (limitation de la conservation);

f) traitées de façon à garantir une sécurité appropriée des données à caractère personnel, y compris la protection contre le traitement non autorisé ou illicite et contre la perte, la destruction ou les dégâts d'origine accidentelle, à l'aide de mesures techniques ou organisationnelles appropriées (intégrité et confidentialité);

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre1#Article5

Ce que consentir (ou pas) signifie, par ©Boulet



Le consentement : volontaire

Le code de Nuremberg (1947) indique que le consentement ne peut être volontaire que si :

- les participants sont **capables** de consentir
- ils ne sont soumis à **aucune forme de coercition**
- ils **comprennent** les risques et bénéfices encourus

Le consentement selon le RGPD

Article 7 - Conditions applicables au consentement

1. Dans les cas où le traitement repose sur le consentement, le responsable du traitement est en mesure de démontrer que la personne concernée a donné son consentement au traitement de données à caractère personnel la concernant.

2. Si le consentement de la personne concernée est donné dans le cadre d'une déclaration écrite qui concerne également d'autres questions, la demande de consentement est présentée sous une forme qui la distingue clairement de ces autres questions, sous une forme compréhensible et aisément accessible, et formulée en des termes clairs et simples. Aucune partie de cette déclaration qui constitue une violation du présent règlement n'est contraignante.

3. La personne concernée a le droit de retirer son consentement à tout moment. Le retrait du consentement ne compromet pas la licéité du traitement fondé sur le consentement effectué avant ce retrait. La personne concernée en est informée avant de donner son consentement. Il est aussi simple de retirer que de donner son consentement.

4. Au moment de déterminer si le consentement est donné librement, il y a lieu de tenir le plus grand compte de la question de savoir, entre autres, si l'exécution d'un contrat, y compris la fourniture d'un service, est subordonnée au consentement au traitement de donn[illegible] caractère personnel qui n'est pas nécessaire à l'exécution dudit contrat.

https://www.cnil.fr/fr/reglement-europeen-protection-donnees/chapitre2

Le consentement en pratique

Il n'y a pas consentement si l'utilisateur ne prend pas de décision :

- opt in *vs* opt out
- importance des paramètres par défaut
- possibilité de retirer son consentement n'importe quand

https://www.grosbill.com/

Au-delà des biais

"All your data are belong to us"

Bonnes pratiques

- Éthique de l'apprentissage machine
- Recommandations du High-Level Experts Group (HLEG) on IA
- Le piège

Des propositions intéressantes

- AI HLEG : recommandation du groupe d'experts européen sur l'IA (EN)
- Une grille d'analyse conséquentialiste [Lefeuvre et al., 2015] (FR)
- CERNA : éthique de la recherche en apprentissage machine (FR)
- CCNE : enjeux éthiques des agents conversationnels (FR)

CERNA

= Commission de réflexion sur l'Éthique de la Recherche en sciences et technologies du Numérique d'Allistene

- créée en 2012, n'existe plus aujourd'hui (CCNE)
- plusieurs acteurs de la recherche publique : Inria, CEA, CNRS, etc.
- a organisé plusieurs journées d'études (conférences) sur différents sujets :
  - les robots
  - l'apprentissage machine (*machine learning*)
  - etc.
- groupes de travail, produisant des rapports

Éthique de l'apprentissage machine : six thèmes

1. les données
2. l'autonomie
3. l'explicabilité
4. la prise de décision
5. le consentement
6. la responsabilité

Guide de bonnes pratiques éthiques pour une IA de confiance

https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

4 principes éthiques :

1. respect de l'autonomie humaine
2. prévention de toute atteinte
3. équité
4. explicabilité

+ tensions entre eux : les décisions prises doivent être documentées et argumentées

Respect de l'autonomie humaine

https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

"En l'absence de justification, les systèmes d'IA ne devraient pas subordonner, contraindre, tromper, manipuler, conditionner ni régenter des êtres humains. Au contraire, les systèmes d'IA devraient être conçus afin d'augmenter, de compléter et de favoriser les compétences cognitives, sociales et culturelles. La répartition des tâches entre êtres humains et systèmes d'IA devrait suivre des principes de conception centrés sur l'humain et donner à l'être humain une possibilité réelle de poser des choix. En d'autres termes, ***il convient de veiller à la supervision et au contrôle humains sur les processus de travail des systèmes d'IA****."*

"moins un être humain peut exercer de contrôle sur un système d'IA, plus il faut approfondir les essais et renforcer la gouvernance."

Prévention de toute atteinte

https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf

"Les systèmes d'IA ne devraient ni porter atteinte, ni aggraver toute atteinte portée, ni nuire aux êtres humains d'une quelconque autre manière. Cela englobe la protection de la dignité humaine ainsi que de l'intégrité mentale et physique. Les systèmes d'IA et les environnements dans lesquels ils évoluent doivent être sûrs et sécurisés. [...] Il convient également d'accorder une attention particulière aux situations dans lesquelles les systèmes d'IA peuvent entraîner ou aggraver des incidences négatives du fait d'asymétries de pouvoir ou d'information, par exemple entre les employeurs et les travailleurs, entre les entreprises et les consommateurs ou entre les pouvoirs publics et les citoyens."

Équité

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

"La mise au point, le déploiement et l'utilisation de systèmes d'IA doivent être équitables. [...] En outre, l'utilisation de systèmes d'IA ne devrait jamais avoir pour conséquence de tromper les utilisateurs (finaux) ou de limiter leur liberté de choix. L'équité implique en outre que les professionnels de l'IA devraient respecter le principe de proportionnalité entre la fin et les moyens, et examiner de manière attentive la manière de trouver un équilibre entre des intérêts et des objectifs en concurrence. Le volet procédural de l'équité suppose la capacité de contester les décisions prises par des systèmes d'IA et par les êtres humains qui les utilisent, ainsi que celle d'introduire un recours efficace à l'encontre de ces décisions. Pour ce faire, ***l'entité responsable de la décision doit pouvoir être identifiée, et le processus de prise de décisions devrait pouvoir être expliqué.***"

Explicabilité

https://ec.europa.eu/digital-single-market/en/news/ethics-guidelines-trustworthy-ai

*"Cela signifie que les **processus doivent être transparents**, que les capacités et la finalité des systèmes d'IA doivent être communiquées ouvertement, et que les décisions – dans la mesure du possible – doivent pouvoir être expliquées aux personnes directement et indirectement concernées. [...] La mesure dans laquelle l'explicabilité est nécessaire dépend fortement du contexte et de la gravité des conséquences si ce résultat est erroné ou imprécis d'une autre manière."*

Au-delà des biais

"All your data are belong to us"

Bonnes pratiques

- Éthique de l'apprentissage machine
- Recommandations du High-Level Experts Group (HLEG) on IA
- Le piège

Une multiplication des bonnes pratiques

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
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)
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
privacy protection	x		x	x	x	x	x	x	x	x	x
fairness, non-discrimination, justice	x	x	x	x		x	x	x	x	x	x
accountability	x	x	x	x	x	x	x	x		x	
transparency, openness	x	x	x	x	x		x	x	x	x	x
safety, cybersecurity	x	x	x	x	x	x	x				x
common good, sustainability, well-being		x	x	x		x	x	x	x		x
human oversight, control, auditing	x		x	x		x	x		x	x	
solidarity, inclusion, social cohesion			x	x		x		x		x	x
explainability, interpretabiliy	x		x			x	x		x		
science-policy link		x	x		x	x	x	x		x	
legislative framework, legal status of AI systems	x	x		x	x	x		x		x	x
future of employment/worker rights		x	x	x				x			x
responsible/intensified research funding		x		x		x	x	x			
public awareness, education about AI and its risks		x	x			x		x			
dual-use problem, military, AI arms race		x			x		x		x		x
field-specific deliberations (health, military, mobility etc.)		x			x			x	x	x	x
human autonomy	x		x	x		x	x				
diversity in the field of AI								x	x	x	x
certification for AI products						x					
protection of whistleblowers										x	x
cultural differences in the ethically aligned design of AI systems											
hidden costs (labeling, clickwork, contend moderation, energy, resources)										x	x
notes on technical implementations	yes, but very few	none	none	none	yes	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)
length (number of words)	16546	22787	766	3249	34017	8609	646	11530	18273	25759	38970
affiliation (government, industry, science)	government	government	science/gov./ind.	government	science	science	science	science	science	science	science
number of ethical aspects	9	12	13	12	8	14	12	13	9	12	13

	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	(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	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		x		x	x	x	x		x	x	x	18
fairness, non-discrimination, justice	x	x		x	x		x		x	x	x	18
accountability	x	x		x	x	x	x			x	x	17
transparency, openness	x	x		x	x		x	x				16
safety, cybersecurity		x	x	x	x	x	x		x		x	16
common good, sustainability, well-being		x	x	x	x			x	x	x	x	16
human oversight, control, auditing	x		x		x	x			x			12
solidarity, inclusion, social cohesion		x			x		x	x			x	11
explainability, interpretabiliy	x	x			x	x				x		10
science-policy link			x		x			x				10
legislative framework, legal status of AI systems				x	x							10
future of employment/worker rights				x	x	x					x	9
responsible/intensified research funding				x		x		x				8
public awareness, education about AI and its risks				x	x	x					x	8
dual-use problem, military, AI arms race		x		x					x			8
field-specific deliberations (health, military, mobility etc.)				x	x							8
human autonomy		x			x							7
diversity in the field of AI		x				x		x				7
certification for AI products				x	x					x		4
protection of whistleblowers					x							3
cultural differences in the ethically aligned design of AI systems				x	x							2
hidden costs (labeling, clickwork, contend moderation, energy, resources)												2
notes on technical implementations	none	none	none	none	none	none	none	none	none	none	none	
proportion of women among authors (f/m)	(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)	1359	4754	441	40915	108092	2272	75	417	882	4488	1481	
affiliation (government, industry, science)	science	science	non-profit	industry	industry	industry	industry	industry	industry	industry	industry	
number of ethical aspects	5	11	4	16	18	9	6	6	6	6	8	

[Hagendorff, 2020]

Les bonnes pratiques ne résoudront pas le problème

"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]

Rappel : au-delà des bonnes pratiques

Les checklists et les grilles sont attirantes :

- elles sont simples
- elles donnent une illusion d'exhaustivité

Mais elles sont loin d'être suffisantes :

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

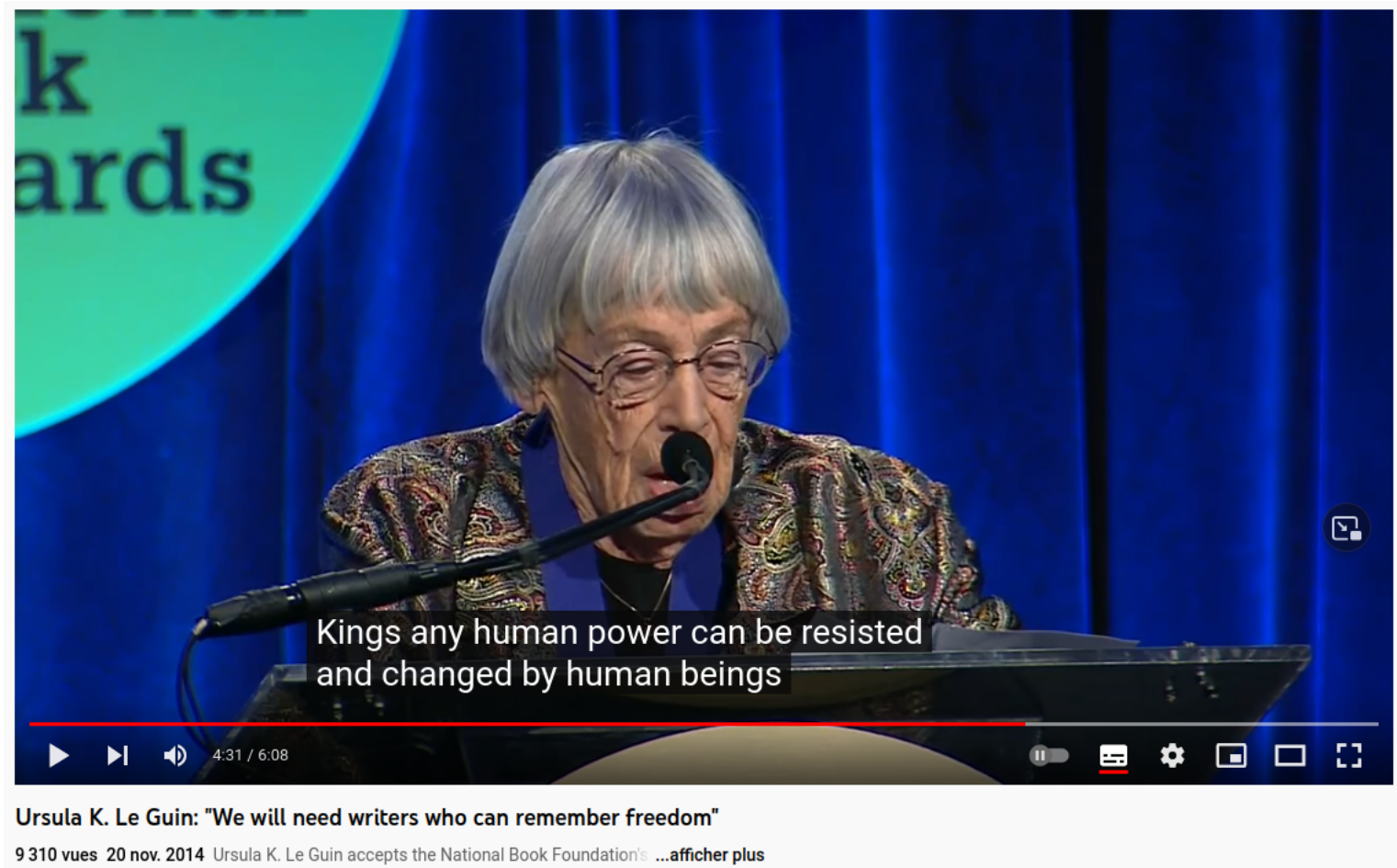
Profiter des grilles sans oublier leurs limites

1. commencer à réfléchir/discuter sans elles
2. les utiliser en complément
3. ne pas limiter sa réflexion parce qu'on a tout vérifier sur la liste/grille

Pour aller plus loin

- bibliographie collaborative du comité d'éthique d'ACL : https://github.com/acl-org/ethics-reading-list/blob/main/README.md

Une question de temps et de pouvoir



https://www.youtube.com/watch?v=Et9Nf-rsALk

Merci pour votre attention et bon loto !

Bingo d'excuses pour ne pas faire d'éthique en TAL

Si je ne le fais pas, quelqu'un d'autre le fera	Qui êtes-vous pour décider ?	L'éthique, c'est culturel	Il y a aussi des utilisations positives
La relecture éthique, c'est de la censure	La science est neutre	Pourtant, vous êtes venus en avion	Les gens aussi sont biaisés
On ne peut pas tout prévoir	Ces travailleurs sont contents avec 0,05$	Il n'y a pas d'alternative	Ne ralentissez pas le progrès
Vous voulez revenir à la bougie ?	La relecture éthique, c'est de l'impérialisme	Les données étaient accessibles sur le Web	Arrêtez de faire de la politique

Emily M. Bender & Karën Fort 2022

https://twitter.com/KarnFort1/status/1532699418067193857

Abdalla, M., Wahle, J. P., Ruas, T., Névéol, A., Ducel, F., Mohammad, S. M., and Fort, K. (2023).
The Elephant in the Room : Analyzing the Presence of Big Tech in Natural Language Processing Research.
In
61st Annual Meeting of the Association for Computational Linguistics
Toronto, Canada.

Antoine, J.-Y. and Lefeuvre, A. (2014).
Pour une réflexion éthique sur les conséquences de l'usage des NTIC : le cas des aides techniques (à composante langagière ou non) aux personnes handicapées.
In Actes de la journée ATALA Éthique et TAL.

Bannour, N., Ghannay, S., Névéol, A., and Ligozat, A.-L. (2021).
Evaluating the carbon footprint of NLP methods : a survey and analysis of existing tools.
In EMNLP, Workshop SustaiNLP, Punta Cana, Dominican Republic.

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.

Bird, S. (2020).
Decolonising speech and language technology.
In Proceedings of the 28th International Conference on Computational Linguistics, pages 3504–3519, Barcelona, Spain

(Online). International Committee on Computational Linguistics.


Couillault, A., Fort, K., Adda, G., and De Mazancourt, H. (2014).
Evaluating Corpora Documentation with regards to the Ethics and Big Data Charter.
In
International Conference on Language Resources and Evaluation (LRE
Reykjavik, Islande.


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.


Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., III, H. D., and Crawford, K. (2021).

Datasheets for datasets.
Commun. ACM, 64(12) :86–92.


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.


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.


Joshi, P., Santy, S., Budhiraja, A., Bali, K., and Choudhury, M. (2020).

The state and fate of linguistic diversity and inclusion in the NLP world.
In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 6282–6293, Online. 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'
Actes de la 1e Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), Caen (France), pages 53–66, Caen, France.


Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., and Gebru, T. (2019).
Model cards for model reporting.

In Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pages 220–229, New York, NY, USA. Association for Computing Machinery.

Rudin, C. (2019).
Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead.
Nature Machine Intelligence, 1 :206–215.

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.