Ethics in NLP: Beyond Biases

Karën Fort

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

Institut des sciences du Digital Management & Cognition
COMPOSANTE DE L'UNIVERSITÉ DE LORRAINE

UNIVERSITÉ DE LORRAINE

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

Carbon footprint

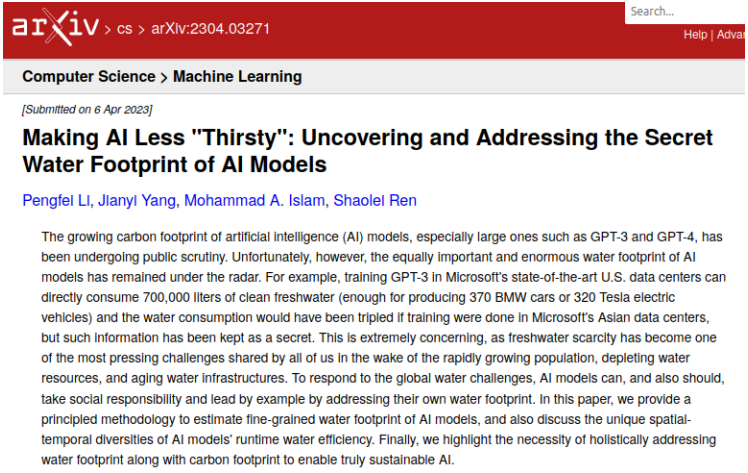
Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture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

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

[Strubell et al., 2019]

Note: this concerns only 1 source out of four [Bannour et al., 2021] ⇒ largely under-estimated

About water consumption



arXiv > cs > arXiv:2304.03271

Search...

Help | Adva

Computer Science > Machine Learning

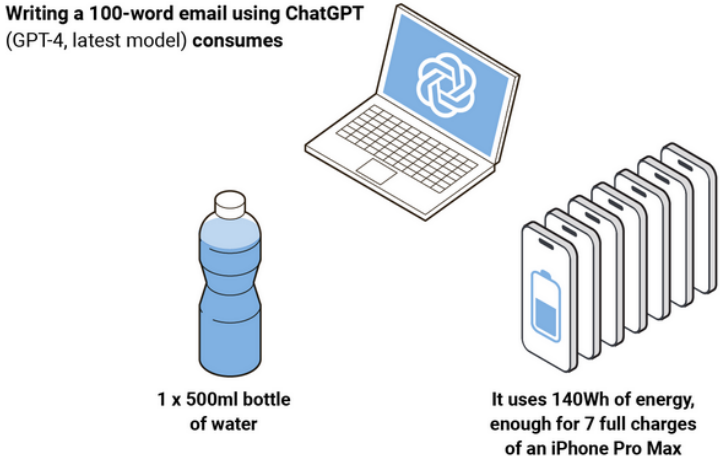
[Submitted on 6 Apr 2023]

Making AI Less "Thirsty": Uncovering and Addressing the Secret Water Footprint of AI Models

Pengfei Li, Jianyi Yang, Mohammad A. Islam, Shaolei Ren

The growing carbon footprint of artificial intelligence (AI) models, especially large ones such as GPT-3 and GPT-4, has been undergoing public scrutiny. Unfortunately, however, the equally important and enormous water footprint of AI models has remained under the radar. For example, training GPT-3 in Microsoft's state-of-the-art U.S. data centers can directly consume 700,000 liters of clean freshwater (enough for producing 370 BMW cars or 320 Tesla electric vehicles) and the water consumption would have been tripled if training were done in Microsoft's Asian data centers, but such information has been kept as a secret. This is extremely concerning, as freshwater scarcity has become one of the most pressing challenges shared by all of us in the wake of the rapidly growing population, depleting water resources, and aging water infrastructures. To respond to the global water challenges, AI models can, and also should, take social responsibility and lead by example by addressing their own water footprint. In this paper, we provide a principled methodology to estimate fine-grained water footprint of AI models, and also discuss the unique spatial-temporal diversities of AI models' runtime water efficiency. Finally, we highlight the necessity of holistically addressing water footprint along with carbon footprint to enable truly sustainable AI.

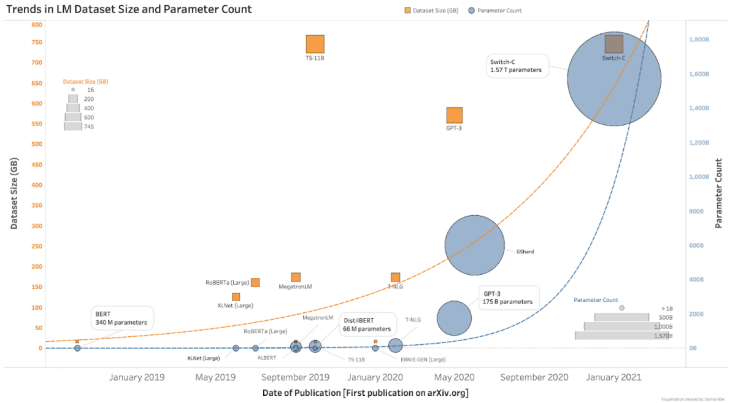
Water consumption: it's, in fact, much worse than expected!



https://www.thetimes.com/article/9167a8a8-96d1-4a68-9a13-824d862f627a

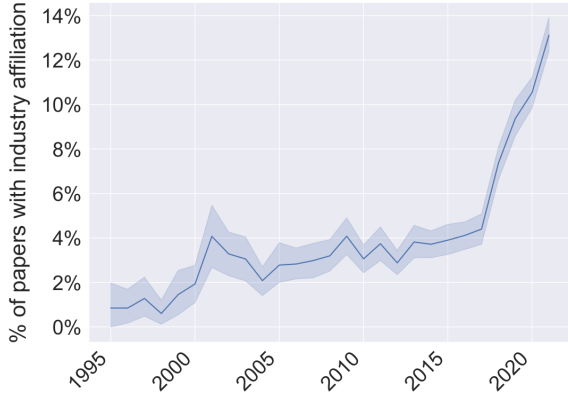
Models trained once and for all?

from a presentation of [Bender et al., 2021]

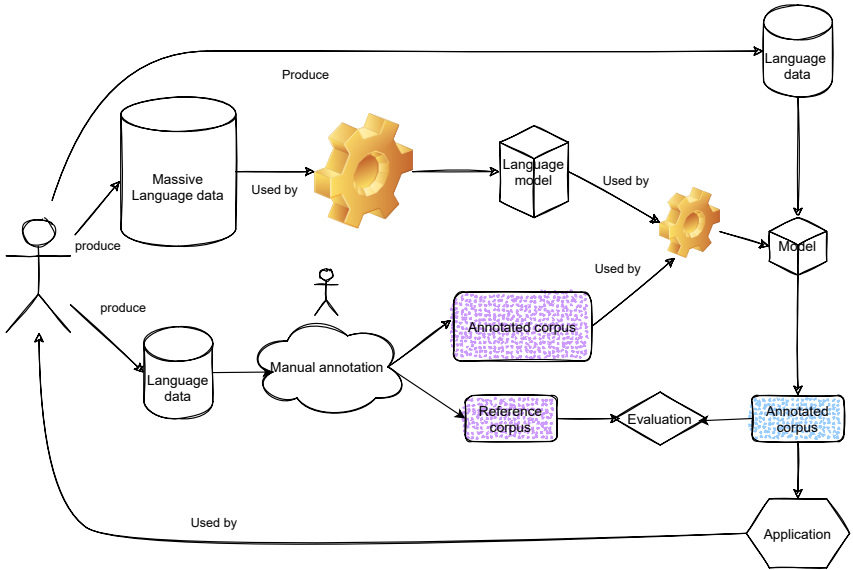


[Bender et al., 2021]

BigTech's presence in NLP [Abdalla et al., 2023]



Today's NLP


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

89 1.9K 8.7K

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

da·ta | \ˈdā-tə, ˈda- *also* ˈdä- \

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/

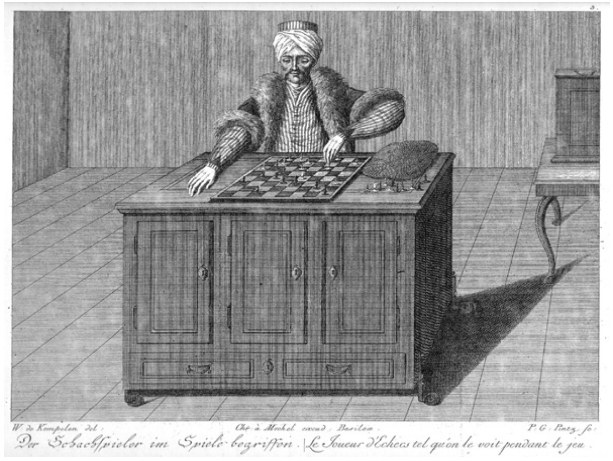
Sensitive Data: exceptions again again again

(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/

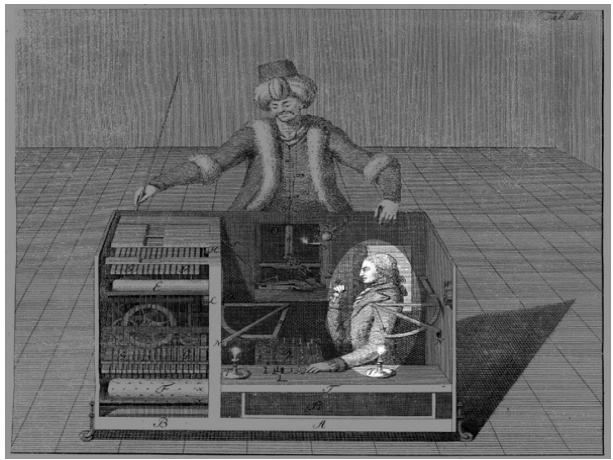
von Kempelen's "Mechanical Turc"

A mechanical chess player created by J. W. von Kempelen in 1770:



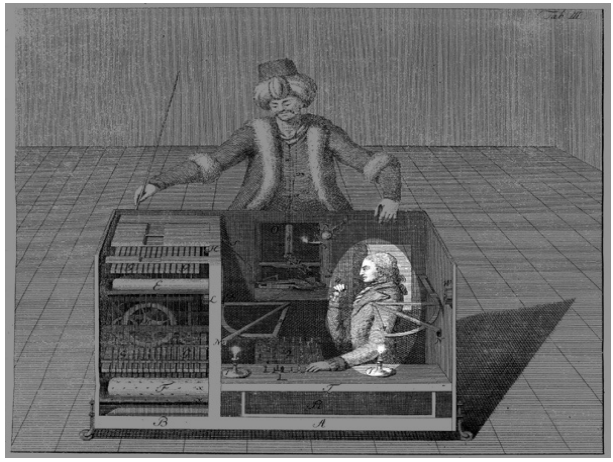
von Kempelen's "Mechanical Turc"

In fact, a chess master was hidden in the machine:



von Kempelen's "Mechanical Turc"

it's artificial **artificial** intelligence!



Amazon Mechanical Turk

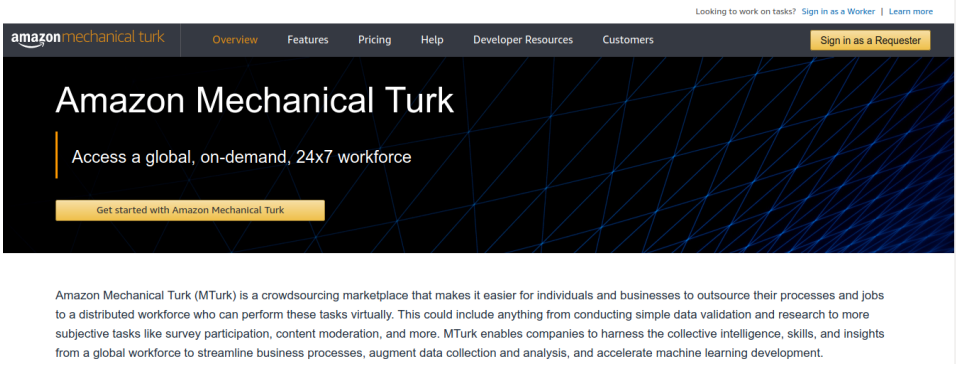
Amazon created for its own needs a

microworking crowdsourcing platform

and opens it to all in 2005 (taking X% of the transactions)

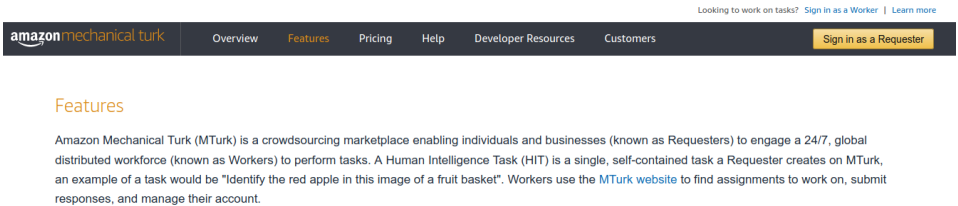
Amazon Mechanical Turk

MTurk



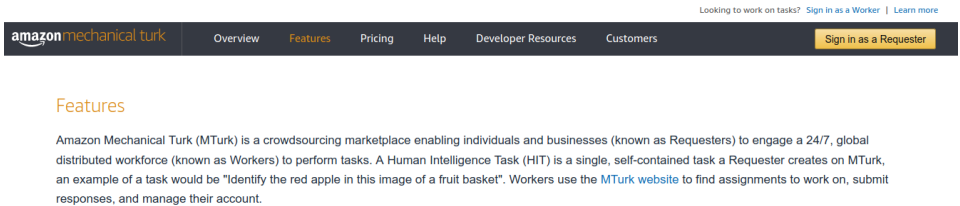
Amazon Mechanical Turk

MTurk is a crowdsourcing platform: the work is *outsourced* via the Web and done by numerous persons (the *crowd*), here the ~~Turkers~~ **workers**



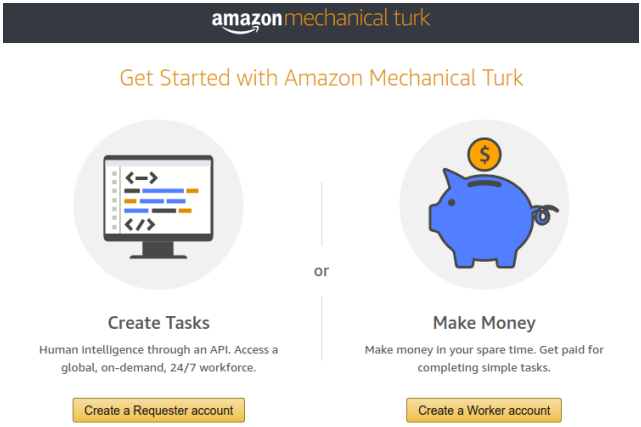
Amazon Mechanical Turk

MTurk is a crowdsourcing platform allowing to perform microwork: tasks are split into subtasks (HITs) and their execution is payed for by the **Requesters**



Amazon Mechanical Turk

MTurk is a crowdsourcing platform allowing to perform microwork: payed for.



Amazon Mechanical Turk

MTurk is a crowdsourcing platform allowing to perform microwork: payed for.

How are Workers paid?

Workers will be paid and Amazon Mechanical Turk (MTurk) fees will be charged when you approve submitted work. If you reject the work, the Worker is not paid and you are not charged the MTurk fees. MTurk Prepaid HITs are subject to Participation Agreement. You can review MTurk pricing here.

Some characteristics of AMT

Remuneration:

- by the task (*illegal* in France except some (rares) exceptions): less than $2/h
- no explicit relationship between the *workers* and the *Requesters*

Tasks:

- traditionally performed by salaried employees: transcription, translation (LDC, ELDA), etc

Typical HITs on AMT

Data Processing

Workers help companies understand and respond to different types of data by:

Editing and transcribing audio content

Translating content from one language to another

Rating the accuracy of search results

Categorizing information based on instructions

Data Verification and Clean-up

Companies with large online directories or catalogs use MTurk to identify duplicate entries and verify item details. Workers help clean and verify data by:

Removing duplicate content from business listings

Identifying incomplete or duplicate product listings in a catalog

Verifying restaurant details such as phone numbers or hours of operation

Converting unstructured data about locations into well-formed addresses

https://www.mturk.com/worker

AMT: a dream come true?

Cheap and Fast — But is it Good?
Evaluating Non-Expert Annotations for Natural Language Tasks

Rion Snow[†] **Brendan O'Connor**[‡] **Daniel Jurafsky**[§] **Andrew Y. Ng**[†]

[†]Computer Science Dept.
Stanford University
Stanford, CA 94305
{rion,ang}@cs.stanford.edu

[‡]Dolores Labs, Inc.
832 Capp St.
San Francisco, CA 94110
brendano@doloreslabs.com

[§]Linguistics Dept.
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

[Snow et al., 2008]

AMT: a dream come true?

Cheap and Fast — But is it Good?
Evaluating Non-Expert Annotations for Natural Language Tasks

Rion Snow[†] **Brendan O'Connor**[‡] **Daniel Jurafsky**[§] **Andrew Y. Ng**[†]

[†]Computer Science Dept.
Stanford University
Stanford, CA 94305
{rion,ang}@cs.stanford.edu

[‡]Dolores Labs, Inc.
832 Capp St.
San Francisco, CA 94110
brendano@doloreslabs.com

[§]Linguistics Dept.
Stanford University
Stanford, CA 94305
jurafsky@stanford.edu

[Snow et al., 2008]

It's cheap, fast, good
and it's a hobby for the *workers*!

AMT allows to reduce the annotation costs

Very (very) low remuneration $\Rightarrow$ low costs? Yes, but...

- costs of putting in place the interface
- costs of creating protections against spammers
- costs of validating and post-processing data

+ some tasks (for ex, translation from Pashto to English) generate costs which are similar to the usual translation costs, because of the lack of qualified workers [Novotney and Callison-Burch, 2010].

When Amazon takes its toll...

Amazon is doubling the fee it collects from "requesters," those seeking laborers to perform online tasks, to 20% beginning July 21. And for tasks requiring at least 10 people, Amazon will charge an additional 20%, a new fee.

[Wall Street Journal blog, June 23, 2015]

AMT allows to produce quality resources?

- allows to produce quality resources in some specific cases (for example, simple transcription)
- but:
  - the quality is insufficient when the task is complex (for example, summarization [Gillick and Liu, 2010])
  - the interface can generate some problems [Tratz and Hovy, 2010]
  - the *workers* can generate problems (cheaters, spammers)
  - by the task remuneration in itself generate problems [Kochhar et al., 2010]
- for some tasks, NLP tools now produce better results

HITs (*Human Intelligence Task*): simplified tasks

Impossible to train oneself on a task on AMT:

⇒ Simplification of tasks:

- a real textual entailment task (entailment, neutral, contradiction) gets reduced to 2 sentences and a question:
  "Would most people say that if the first sentence is true, then the second sentence must be true?" [Bowman et al., 2015]

AMT: a hobby for the *workers*?

[Ross et al., 2010, Ipeirotis, 2010] show that:

- *workers* are mostly motivated by **money** (91%):
  - 20% consider AMT as their primary source of income
  - 50% as their secondary source of income
  - the hobby aspect is important only for a (US) minority (30%)
- 20% of the *workers* spend more than 15h per week on AMT, and contribute to 80% of the tasks
- the average observed hourly wage is below $2 [Hara et al., 2019]

[Gupta et al., 2014]: due to the impossibility to train, an important amount of hidden work is performed by the *workers*

Who are the AMT workers?

https://demographics.mturk-tracker.com [Difallah et al., 2018]

Is AMT ethical or even legal? [Fort et al., 2011]

Ethics:

- no identification: no official link between the *Requesters* and the *workers* or among *workers*
- (almost) no possibility to unionize, to protest against the wrongdoings of *Requesters* or to take legal action against them
- no minimum wage ($<2/h on average [Hara et al., 2019])
- possibility to refuse to pay the *workers*

Is AMT ethical or even legal? [Fort et al., 2011]

How are Workers paid?

Workers will be paid and Amazon Mechanical Turk (MTurk) fees will be charged when you approve submitted work. If you reject the work, the Worker is not paid and you are not charged the MTurk fees. MTurk Prepaid HITs are subject to Participation Agreement. You can review MTurk pricing here.

Is AMT ethical or even legal? [Fort et al., 2011]

Law:

- `Amazon` licence agreement: the *workers* are considered as independent workers $\Rightarrow$ they are supposed to declare themselves as such and pay taxes and social benefit charges accordingly
- illusory, considering the level of remuneration

$\Rightarrow$ states lose a legitimate source of income

Depending on an external platform

Impossibility to control:

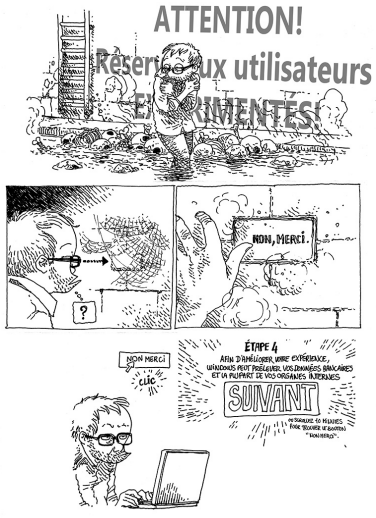
- costs
- working conditions
- selection of workers
- conditions of experiment

Making choices

- **Other types of crowdsourcing**:
  - Citizen Sciences: `LanguageArc`, a Linguistic Data Consortium (LDC)platform, allowing to create tasks and have them performed by voluntary users
  - Games With A Purpose: `ESP Game` [von Ahn, 2006], `Phrase Detectives` [Chamberlain et al., 2008], etc

- **Other solutions**:
  - unsupervised, semi supervised, weakly supervised approaches
  - pre-annotation
  - using existing resources (not well-known, forgotten)

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. [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:

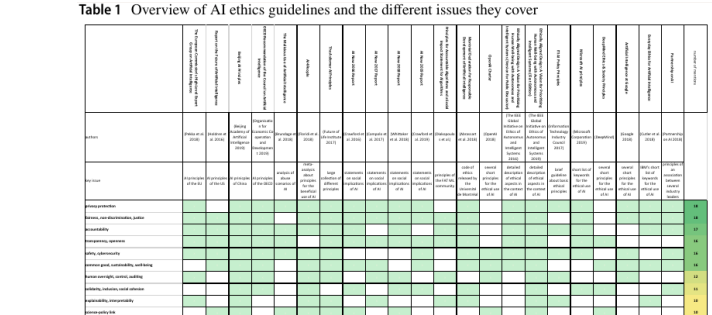
- opt in *vs* opt out
- importance of the default settings
- possibility to withdraw one's consent at anytime

https://www.grosbill.com/

Data and "informed" consent

Guidelines, guidelines everywhere!

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



[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)

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

Reading List

Please participate!

ACL ethics committee reading list

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.

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.


Bowman, S. R., Angeli, G., Potts, C., and Manning, C. D. (2015).
A large annotated corpus for learning natural language inference.
In Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 632–642, Lisbon, Portugal. Association for Computational Linguistics.


Chamberlain, J., Poesio, M., and Kruschwitz, U. (2008).
Phrase Detectives: a web-based collaborative annotation game.
In Proceedings of the International Conference on Semantic Systems (I-Semantics'08), Graz, Autriche.


Difallah, D., Filatova, E., and Ipeirotis, P. (2018).
Demographics and dynamics of mechanical turk workers.

In Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining, WSDM '18, pages 135–143, New York, NY, USA. Association for Computing Machinery.


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.


Gillick, D. and Liu, Y. (2010).
Non-expert evaluation of summarization systems is risky.
In Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, CSLDAMT '10, pages 148–151, Stroudsburg, PA, USA. Association for Computational Linguistics.


Gupta, N., Martin, D., Hanrahan, B. V., and O'Neill, J. (2014).
Turk-life in india.

In Proceedings of the 18th International Conference on Supporting Group Work, GROUP '14, pages 1–11, New York, NY, USA. ACM.


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.


Hara, K., Adams, A., Milland, K., Savage, S., Hanrahan, B. V., Bigham, J. P., and Callison-Burch, C. (2019).
Worker demographics and earnings on amazon mechanical turk: An exploratory analysis.
CHI EA '19, pages 1–6, New York, NY, USA. Association for Computing Machinery.


Ipeirotis, P. (2010).
The new demographics of mechanical turk.

http://behind-the-enemy-lines.blogspot.com/2010/03/new-demographics-of-mechanical-turk.html.


Kochhar, S., Mazzocchi, S., and Paritosh, P. (2010).
The anatomy of a large-scale human computation engine.
In Proceedings of Human Computation Workshop at the 16th ACM SIKDD Conference on Knowledge Discovery and Data Mining, KDD 2010, Washington D.C.


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 T
Actes de la 1e Ethique et TRaitemeNt Automatique des Langues (ETeRNAL'2015), Caen (France), pages 53–66, Caen, France.


Novotney, S. and Callison-Burch, C. (2010).
Cheap, fast and good enough: automatic speech recognition with non-expert transcription.

In Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), HLT'10, pages 207–215, Stroudsburg, PA, USA. Association for Computational Linguistics.


Ross, J., Irani, L., Silberman, M. S., Zaldivar, A., and Tomlinson, B. (2010).
Who are the crowdworkers?: shifting demographics in mechanical turk.
In Proceedings of the 28th of the international conference extended abstracts on Human factors in computing systems, CHI EA '10, pages 2863–2872, New York, NY, USA. ACM.


Snow, R., O'Connor, B., Jurafsky, D., and Ng., A. Y. (2008).
Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks.
In Proceedings of EMNLP 2008, pages 254–263, Waikiki, Honolulu, Hawaii.


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.


Tratz, S. and Hovy, E. (2010).
A taxonomy, dataset, and classifier for automatic noun compound interpretation.
In Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics, pages 678–687, Uppsala, Suède. Association for Computational Linguistics.


von Ahn, L. (2006).
Games with a purpose.
IEEE Computer Magazine, pages 96–98.