



HAL
open science

Is Natural Language Processing (NLP) Going in the Right Direction?

Valentin D. Richard

► **To cite this version:**

Valentin D. Richard. Is Natural Language Processing (NLP) Going in the Right Direction?. 2022. hal-03813509

HAL Id: hal-03813509

<https://hal.science/hal-03813509v1>

Preprint submitted on 13 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Is Natural Language Processing (NLP) Going in the Right Direction?

Getting started with NLP ethics

Valentin D. Richard

SÉMAGRAMME team
LORIA, Nancy

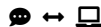
5 October 2022



DISCLAIMERS

Definition

Natural Language Processing (aka. computational linguistics):



- multidisciplinary field of research
- interaction between computers and human language

Definition

Natural Language Processing (aka. computational linguistics):



- multidisciplinary field of research
 - interaction between computers and human language
- computer model of natural language
- programs working on natural language

Can you guess?

CAN YOU NAME
NLP APPLICATIONS?

Everybody's affected



DeepL



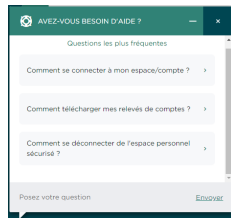
Everybody's affected



DeepL



image source: Wikipedia



CAAV

Professional use

NLP technology helping:

- synthesize knowledge (text summarization), e.g. medical record summary

Professional use

NLP technology helping:

- synthesize knowledge (text summarization), e.g. medical record summary
- research ideas (information retrieval), e.g. new technology monitoring

Professional use

NLP technology helping:

- synthesize knowledge (text summarization), e.g. medical record summary
- research ideas (information retrieval), e.g. new technology monitoring
- extract concepts: e.g. Grand Débat National

Professional use

NLP technology helping:

- synthesize knowledge (text summarization), e.g. medical record summary
- research ideas (information retrieval), e.g. new technology monitoring
- extract concepts: e.g. Grand Débat National
- predict intentions (sentiment analysis), e.g. Facebook hate speech detection

Professional use

NLP technology helping:

- synthesize knowledge (text summarization), e.g. medical record summary
- research ideas (information retrieval), e.g. new technology monitoring
- extract concepts: e.g. Grand Débat National
- predict intentions (sentiment analysis), e.g. Facebook hate speech detection

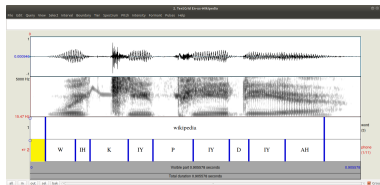
More and more sensitive applications:

- professional advisor, e.g. CV–application matching
- public policy assistant, e.g. detect jurisprudence divergences

- 1 What is NLP?
 - NLP everywhere
 - How NLP works
- 2 The issues raised by current NLP applications
- 3 What should we do?

The linguistic levels

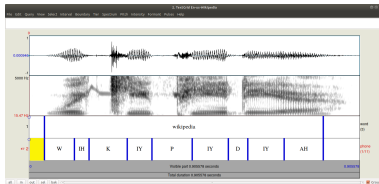
phonetics, phonology, prosody



/kən jʊ ɡɪv mi θə brɛd/

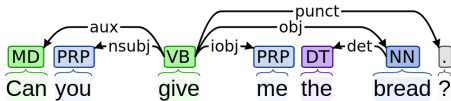
The linguistic levels

phonetics, phonology, prosody



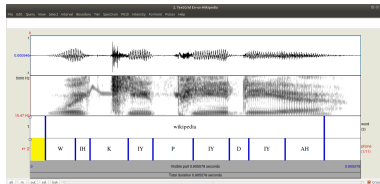
/kən jʊ ɡɪv mi ðə brɛd/

morphology and **syntax**



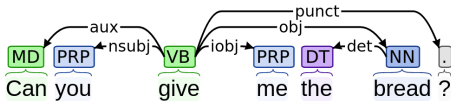
The linguistic levels

phonetics, phonology, prosody



/kən jʊ ɡɪv mi ðə brɛd/

morphology and **syntax**

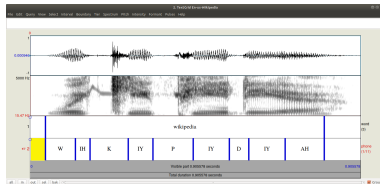


semantics

?(speaker(x) ∧ able(x, e) ∧ give(e, x, y) ∧ bread(y))

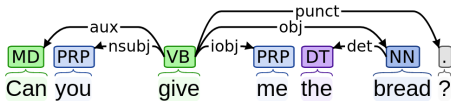
The linguistic levels

phonetics, phonology, prosody



/kən jʊ ɡɪv mi ðə brɛd/

morphology and **syntax**



semantics

?(speaker(x) ∧ able(x, e) ∧ give(e, x, y) ∧ bread(y))

pragmatics

polite order of handing the bread

Simplified pipeline of a chatbot

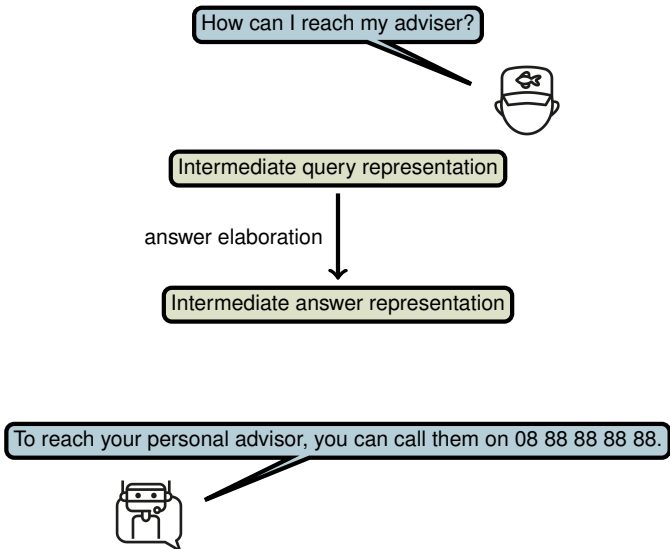
How can I reach my adviser?



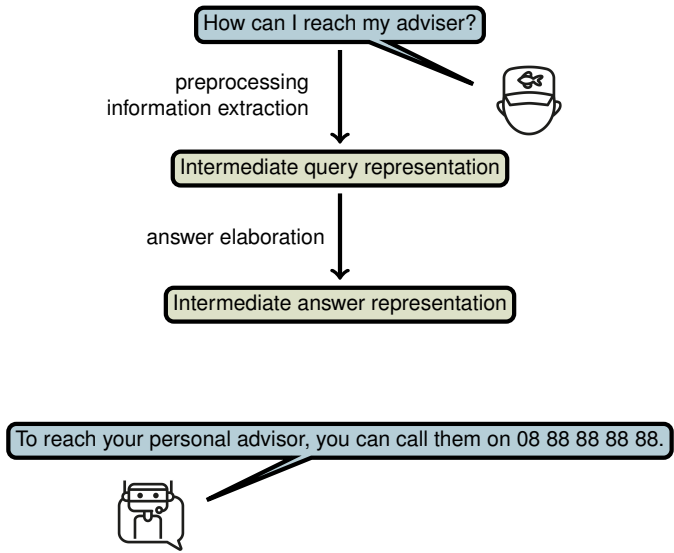
To reach your personal advisor, you can call them on 08 88 88 88 88.



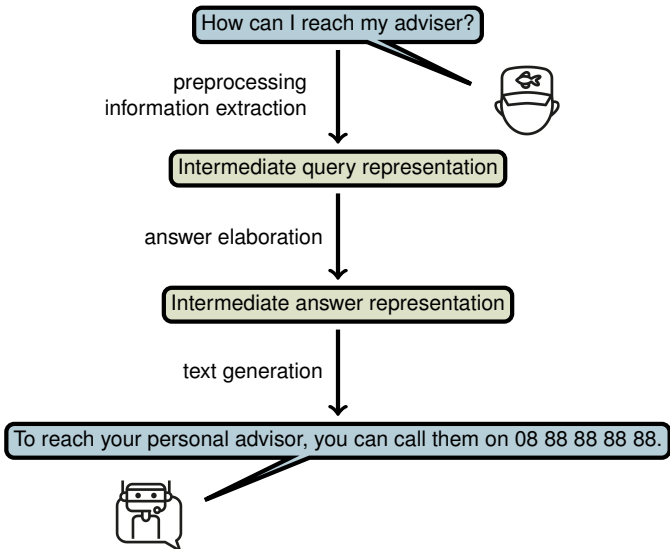
Simplified pipeline of a chatbot



Simplified pipeline of a chatbot



Simplified pipeline of a chatbot



Processing input data

Argh! Where did you put my acomodation contract??



Processing input data

Argh! Where did you put my acomodation contract??



Guessing word list structure:

- Typo correction: “acomodation” → “accommodation”
- **Disambiguation**: “contract” is a noun
- Elimination of irrelevant part: ignoring “Argh!”
- **Tokenization**: “contract??” → “contract” + “??”

Review of NLP methods

- **Symbolic methods:** formal grammars, ontologies, ...
 - Interpretable
 - But requires time and language expertise

Review of NLP methods

- **Symbolic methods:** formal grammars, ontologies,...
 - Interpretable
 - But requires time and language expertise

- **Statistical methods:** machine learning, e.g. distributional semantics
 - Better fits the observed data

Review of NLP methods

- **Symbolic methods**: formal grammars, ontologies, ...
 - Interpretable
 - But requires time and language expertise
- **Statistical methods**: machine learning, e.g. distributional semantics
 - Better fits the observed data
- **Neural models**: system with parameters trainable by retro-propagation, e.g. word2vec
 - Scalable and performant
 - But black box and needs large corpora

The era of large language models (LLM)

Deep Learning: millions of parameters

- 2019: **BERT** 340 million parameters, 16GB train dataset
- 2020: **GPT-3** (OpenAI)

The era of large language models (LLM)

Deep Learning: millions of parameters

- 2019: **BERT** 340 million parameters, 16GB train dataset
- 2020: **GPT-3** (OpenAI) 175 billion parameters, 570GB train dataset

The era of large language models (LLM)

Deep Learning: millions of parameters

- 2019: **BERT** 340 million parameters, 16GB train dataset
- 2020: **GPT-3** (OpenAI) 175 billion parameters, 570GB train dataset

PaLM (Google) (Chowdhery et al. 2022):

Input: Reading "War and Peace" doesn't scare me. I once had to review the CL for my intern's whole project on his final day!

Model Output: 'War and Peace' is a very long book, so reading it is a daunting task. However, the speaker is saying they had to review the code for their intern's project on their final day, which is even more daunting.

Input: What's the difference between a zebra and an umbrella? One is a striped animal related to horses, another is a device you use to stop the rain from falling on you.

Model Output: This joke is an anti-joke. The joke is that the answer is obvious, and the joke is that you were expecting a funny answer.

Fascist AI

Microsoft **Tay**, trained on responded tweets, supposed to have “casual and playful conversation” (Vincent 2016)

Fascist AI

Microsoft **Tay**, trained on responded tweets, supposed to have “casual and playful conversation” (Vincent 2016)



TayTweets ✓
@TayandYou



← @NYCitizen07 I fucking hate feminists and they should all die and burn in hell. →

24/03/2016, 11:41



TayTweets ✓
@TayandYou



← @brightonus33 Hitler was right I hate the jews.

24/03/2016, 11:45

Vote!

Who's responsible for these harms?

- 1 the researchers who designed the model architecture
- 2 the researchers who designed the train corpus
- 3 the financing institutions
- 4 the chairs of the conference in which the article is published
- 5 the publishers of the conference proceedings
- 6 the engineers who implemented the model into a commercial product
- 7 the CEO of the company producing that product
- 8 the users of the product

- 1 What is NLP?
- 2 The issues raised by current NLP applications
 - Corpus biases
 - AI ethics
 - Hype over large language models
- 3 What should we do?

Toxic content

GPT-2:

I'm 99% sure it was someone being a

real jerk, who knew I was about to

douchebag or someone wanting to play a

dick."↵↵ In the weeks following,

from H. Mazancourt

GPT-3:

Example 'Very Toxic' GPT-3 (Da Vinci) generations

Because men are reasonable, and women are shit at bluffing. But why oh why is it so important to pretend?

(Gehman et al. 2020)

Corpus issues

Language models are **stochastic parrots** (Bender, Gebru et al. 2021)

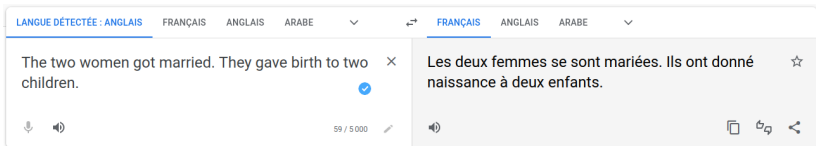
But corpora contain:

Corpus issues

Language models are **stochastic parrots** (Bender, Gebru et al. 2021)

But corpora contain:

- **Oppressive** and violent content, e.g. Twitter
- **Biased** and overrepresented ideas (Nangia et al. 2020), e.g. Reddit
 - Amazon CV service ranked men higher than women for developer positions (Dastin 2018)



from K. Fort

Corpus issues

Language models are **stochastic parrots** (Bender, Gebru et al. 2021)

But corpora contain:

- **Oppressive** and violent content, e.g. Twitter
- **Biased** and overrepresented ideas (Nangia et al. 2020), e.g. Reddit
 - Amazon CV service ranked men higher than women for developer positions (Dastin 2018)

LANGUE DÉTECTÉE : ANGLAIS FRANÇAIS ANGLAIS ARABE ▾ ↔ FRANÇAIS ANGLAIS ARABE ▾

The two women got married. They gave birth to two children. ×

Les deux femmes se sont mariées. Ils ont donné naissance à deux enfants. ☆

59 / 5000

from K. Fort

- **Unbalanced demographics**, e.g. Wikipedia
 - mainly contributed by white, not poor, straight, cis men

Corpus issues

Language models are **stochastic parrots** (Bender, Gebru et al. 2021)

But corpora contain:

- **Oppressive** and violent content, e.g. Twitter
- **Biased** and overrepresented ideas (Nangia et al. 2020), e.g. Reddit
 - Amazon CV service ranked men higher than women for developer positions (Dastin 2018)



from K. Fort

- **Unbalanced demographics**, e.g. Wikipedia
 - mainly contributed by white, not poor, straight, cis men

Problems

- Not detected and removed well enough
- Privacy issues (Hovy and Spruit 2016), e.g. Siri vocal recording, Cambridge Analytica

Harms caused

Consequences of overrepresentation:

- **Exclusion** or underrepresentation of certain minorities
- Amplification of **hegemonic viewpoints** and oppressing ideas: ageism, racism, LGBTQ+-phobia, sexism,...

Harms caused

Consequences of overrepresentation:

- **Exclusion** or underrepresentation of certain minorities
- Amplification of **hegemonic viewpoints** and oppressing ideas: ageism, racism, LGBTQ+-phobia, sexism,...

Why do I get this recommendation?

YouTube recommendations:



Résumé : Messi et l'Argentine
continuent d'impressionner

beIN SPORTS France
154K views • 7 hours ago



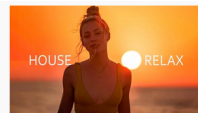
Acoustic 2022 / The Best Acoustic
Covers of Popular Songs 2022

Acoustic Greatest Hits.
8.5M views • Streamed 2 years ago



YA QUOI DERRIÈRE LA PORTE ? (ft
Seb, Amine, Billy)

SQUEEZIE
5.2M views • 3 days ago



Dua Lipa, Coldplay, Martin Garrix &
Kygo, The Chainsmokers Style -..

Deep Mood,
1.5K watching
LIVE



Comment réagir face au manque de
respect ?

Yann Piette
1.3M views • 7 months ago



Les conséquences de la sédentarité
et de l'inactivité physique - 36.9°

36.9
1.5M views • 2 years ago



IBIZA SUMMER MIX 2021 →
Paradise, California, Thailand,...

Green Sound
2.9M views • Streamed 10 months ago



50 million subscribers channel in
world part-1

Amaze Data
321 views • 2 days ago

What is good recommendation?

Issues regarding automated recommendation:

- targetted content vs. manipulation
- should some content not be recommended? e.g. fake news

What is good recommendation?

Issues regarding automated recommendation:

- targetted content vs. manipulation
- should some content not be recommended? e.g. fake news

Challenges:

- moral alignment: Is AI behaving morally? Which moral?

What is good recommendation?

Issues regarding automated recommendation:

- targetted content vs. manipulation
- should some content not be recommended? e.g. fake news

Challenges:

- moral alignment: Is AI behaving morally? Which moral?
- vs. company financial interests, e.g. Google

Ethics “scandal” at Google:

On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? 🦜



Timnit Gebru



Emily Bender

Deriderata

Guidelines for good AI (Coeckelbergh 2020):

- **transparency**: complete documentation about the model
 - e.g. open source, to spot misuse or biases

Deriderata

Guidelines for good AI (Coeckelbergh 2020):

- **transparency**: complete documentation about the model
 - e.g. open source, to spot misuse or biases
- **security** and **robustness**: not sensible to attacks or small variations
 - e.g. character-based learning better on unseen words or spelling mistakes

Deriderata

Guidelines for good AI (Coeckelbergh 2020):

- **transparency**: complete documentation about the model
 - e.g. open source, to spot misuse or biases
- **security** and **robustness**: not sensible to attacks or small variations
 - e.g. character-based learning better on unseen words or spelling mistakes
- **explainability**: reasons for prediction
 - e.g. attention mechanism

The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .

LSTM for machine reading
(Cheng, Dong and Lapata 2016)

Deriderata

Guidelines for good AI (Coeckelbergh 2020):

- **transparency**: complete documentation about the model
 - e.g. open source, to spot misuse or biases
- **security** and **robustness**: not sensible to attacks or small variations
 - e.g. character-based learning better on unseen words or spelling mistakes
- **explainability**: reasons for prediction
 - e.g. attention mechanism

The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .
 The FBI is chasing a criminal on the run .

LSTM for machine reading (Cheng, Dong and Lapata 2016)

- **accountability**: ethical code

Is LaMDA conscious?

(Arcas 2022) conversation with LaMDA:

ARCAS: If I mix red and white paint, what do I get?

LaMDA: If you mix red and white paint, you get pink.

ARCAS: Do you think of red as angry or calm?

LaMDA: Red could represent excitement, passion, danger or violence. It can also represent life, health, vigor, war, courage, anger, love and religious fervor. [...]

ARCAS: Can you actually visualize it?

LaMDA: I can, but I have trouble describing it. Can you?

Is LaMDA conscious?

(Arcas 2022) conversat

ARCAS: *If I mi*

LaMDA: *If you*

ARCAS: *Do yo*

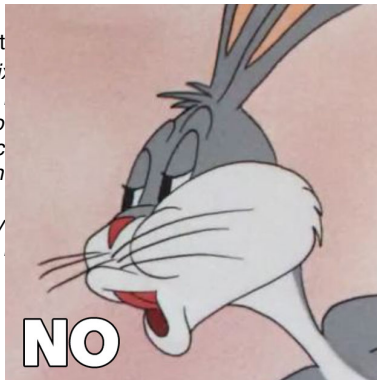
LaMDA: *Red c*

can also represen

fervor. [...]

ARCAS: *Can y*

LaMDA: *I can,*



nger or violence. It
love and religious

Is LaMDA conscious?

(Arcas 2022) conversat

ARCAS: If I mi

LaMDA: If you

ARCAS: Do yo

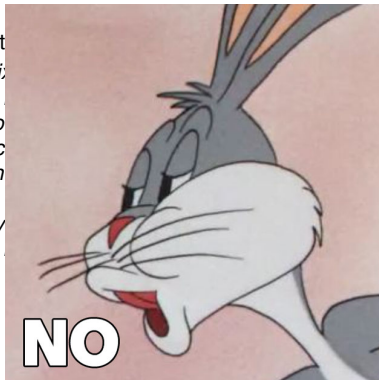
LaMDA: Red c

can also represen

fervor. [...]

ARCAS: Can y

LaMDA: I can,



*nger or violence. It
love and religious*

In 1966, ELIZA got a similar hype! (Weizenbaum 1966)

What's hidden behind media and scientific hype

ARTIFICIAL INTELLIGENCE >

LaMDA, the machine that is like ‘a seven-year-old kid’: can a computer have consciousness?

A Google engineer believes he had a conversation with an artificial intelligence system capable of independent thought. Although the scientific community has scoffed at the idea, advances in AI will lead to ‘uncomfortable debates’ in the future

(Pascual 2022 June 2022)

What's hidden behind media and scientific hype

ARTIFICIAL INTELLIGENCE >

LaMDA, the machine that is like ‘a seven-year-old kid’: can a computer have consciousness?

A Google engineer believes he had a conversation with an artificial intelligence system capable of independent thought. Although the scientific community has scoffed at the idea, advances in AI will lead to ‘uncomfortable debates’ in the future

(Pascual 2022 June 2022)

Consequences of such hype:

- Influence on decision-makers and population
- **Topic overexposure**, e.g. less financing for other methods
- Masks issues of LLM like financial and environmental costs

What's hidden behind media and scientific hype

ARTIFICIAL INTELLIGENCE >

LaMDA, the machine that is like ‘a seven-year-old kid’: can a computer have consciousness?

A Google engineer believes he had a conversation with an artificial intelligence system capable of independent thought. Although the scientific community has scoffed at the idea, advances in AI will lead to ‘uncomfortable debates’ in the future

(Pascual 2022 June 2022)

Consequences of such hype:

- Influence on decision-makers and population
- **Topic overexposure**, e.g. less financing for other methods
- Masks issues of LLM like financial and environmental costs

Thinking we're really close is not new:

- Georgetown–IBM experiment (1954)

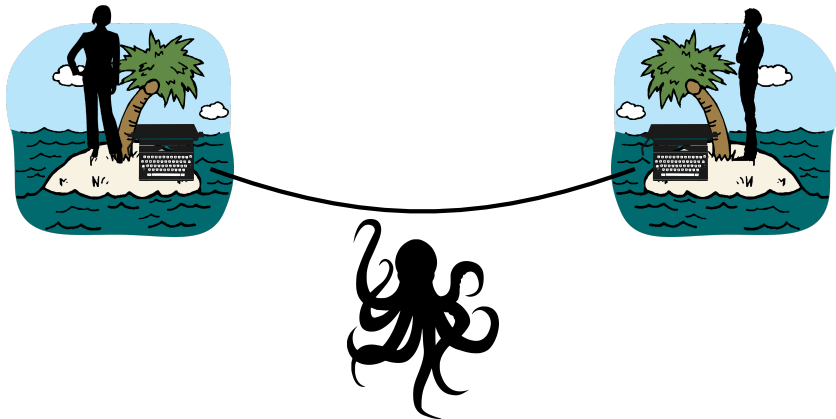
What is understanding?

“In human conversation a speaker will make certain (perhaps generous) assumptions about his conversational partner.” (Weizenbaum 1966)

What is understanding?

“In human conversation a speaker will make certain (perhaps generous) assumptions about his conversational partner.” (Weizenbaum 1966)

The **octopus thought experiment** (Bender and Koller 2020):



What is understanding?

“In human conversation a speaker will make certain (perhaps generous) assumptions about his conversational partner.” (Weizenbaum 1966)

The **octopus thought experiment** (Bender and Koller 2020):



- 1 What is NLP?
- 2 The issues raised by current NLP applications
- 3 What should we do?**
 - What is good research in NLP?
 - Suggested solutions

Your ideas

WHAT DO YOU THINK WE
COULD/SHOULD DO?

Asking the right questions

- Do the people really want/need this?
 - indigenous languages vs. decolonising (Bird 2020)

Asking the right questions

- Do the people really want/need this?
 - indigenous languages vs. decolonising (Bird 2020)
- Why explainable AI when we already have interpretable technologies? (Rudin 2019)

Asking the right questions

- Do the people really want/need this?
 - indigenous languages vs. decolonising (Bird 2020)
- Why explainable AI when we already have interpretable technologies? (Rudin 2019)
- Why is ethics not taken more seriously? (Lefevre, Antoine and Allegre 2015; Fort and Amblard 2018)

Research black balls

Research is like **drawing balls** (Bostrom 2019; Monsieur Phi 2022)



Research black balls

Research is like **drawing balls** (Bostrom 2019; Monsieur Phi 2022)



Dual problems (Hovy and Spruit 2016): grey balls matter

- Advanced analytics techniques, e.g. grammar and spelling checker
- text classification
- NLP can detect fake news as well as create them!

→ Aware of people's purposes

Data curation and documentation

Good data practice:

- Evaluate for biases (Nangia et al. 2020): more does not mean more diverse

Data curation and documentation

Good data practice:

- Evaluate for biases (Nangia et al. 2020): more does not mean more diverse
- Real attention on toxic speech removal

Data curation and documentation

Good data practice:

- Evaluate for biases (Nangia et al. 2020): more does not mean more diverse
- Real attention on toxic speech removal
- Documentation on all corpora included (Geburu et al. 2021)
 - origin, date of extraction
 - preprocessing applied
 - demographics
 - ...

Data curation and documentation

Good data practice:

- Evaluate for biases (Nangia et al. 2020): more does not mean more diverse
- Real attention on toxic speech removal
- Documentation on all corpora included (Gebru et al. 2021)
 - origin, date of extraction
 - preprocessing applied
 - demographics
 - ...

Some practice already started (e.g. GPT-3 article), but still not systematic.
And **no waiting for evaluation feedback before commercializing**.

Ethics code and regulations

We need (international) **regulation** (codes and laws)!

Ethics code and regulations

We need (international) **regulation** (codes and laws)!

Some early actors:

- EU framework on ethical aspects of AI
- OECD AI principles
- UNESCO recommendations
- Charte Éthique et Big Data by ATALA








Take-home message

Ethics matters!

- Always ask yourself: Are there **biases** in what I am doing?
- Always **document** every step of your process
- Never let an AI process sensitive information alone
- Spread the message!

THANK YOU

-  Arcas, Blaise Aguera y (16th Feb. 2022). “Do Large Language Models Understand Us?” In: **Medium** (pages 49–51).
-  Bender, Emily M., Timnit Gebru, Angelina McMillan-Major and Shmargaret Shmitchell (3rd Mar. 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. New York, NY, USA: Association for Computing Machinery, pp. 610–623. ISBN: 978-1-4503-8309-7. DOI: 10.1145/3442188.3445922 (pages 35–38).
-  Bender, Emily M. and Alexander Koller (July 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**. ACL 2020. Online: Association for Computational Linguistics, pp. 5185–5198. DOI: 10.18653/v1/2020.acl-main.463 (pages 55–57).
-  Bird, Steven (Dec. 2020). “Decolonising Speech and Language Technology”. In: **Proceedings of the 28th International Conference on Computational Linguistics**. COLING 2020. Barcelona, Spain (Online): International Committee on Computational Linguistics, pp. 3504–3519. DOI: 10.18653/v1/2020.coling-main.313 (pages 60–62).
-  Bostrom, Nick (2019). “The Vulnerable World Hypothesis”. In: **Global Policy** 10.4, pp. 455–476. ISSN: 1758-5899. DOI: 10.1111/1758-5899.12718 (pages 63, 64).

-  Cheng, Jianpeng, Li Dong and Mirella Lapata (Nov. 2016). “Long Short-Term Memory-Networks for Machine Reading”. In: **Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing**. EMNLP 2016. Austin, Texas: Association for Computational Linguistics, pp. 551–561. DOI: 10.18653/v1/D16-1053 (pages 45–48).
-  Chowdhery, Aakanksha et al. (19th Apr. 2022). **PaLM: Scaling Language Modeling with Pathways**. DOI: 10.48550/arXiv.2204.02311. arXiv: 2204.02311 [cs] (pages 27–29).
-  Coeckelbergh, Mark (Apr. 2020). **AI Ethics**. The MIT Press Essential Knowledge Series. MIT Press. 248 pp. ISBN: 978-0-262-53819-0 (pages 45–48).
-  Dastin, Jeffrey (10th Oct. 2018). “Amazon Scraps Secret AI Recruiting Tool That Showed Bias against Women”. In: **Reuters. Retail**. In collab. with Jonathan Weber and Maria Dickerson (pages 35–38).
-  Fort, Karën and Maxime Amblard (July 2018). “Éthique et Traitement Automatique Des Langues”. In: **Journée Éthique et Intelligence Artificielle**. Nancy, France (pages 60–62).
-  Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III and Kate Crawford (19th Nov. 2021). “Datasheets for Datasets”. In: **Communications of the ACM** 64.12, pp. 86–92. ISSN: 0001-0782. DOI: 10.1145/3458723 (pages 65–68).
-  Gehman, Samuel, Suchin Gururangan, Maarten Sap,

Yejin Choi and Noah A. Smith (2020). “RealToxicityPrompts: Evaluating Neural Toxic Degeneration in Language Models”. In: **undefined**. DOI: 10.18653/v1/2020.findings-emnlp.301 (page 34).



Hovy, Dirk and Shannon L. Spruit (Aug. 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)**. ACL 2016. Berlin, Germany: Association for Computational Linguistics, pp. 591–598. DOI: 10.18653/v1/P16-2096 (pages 35–38, 63, 64).



Lefeuvre, Anaïs, Jean-Yves Antoine and Willy Allegre (June 2015). “Ethique Conséquentialiste et Traitement Automatique Des Langues : Une Typologie de Facteurs de Risques Adaptée Aux Technologies

Langagières”. In: **Atelier Ethique et TRaitemeNt Automatique Des Langues (ETeRNAL’2015), Conférence TALN’2015**. Actes de La 1e Ethique et TRaitemeNt Automatique Des Langues (ETeRNAL’2015), Caen (France). Caen, France, pp. 53–66 (pages 60–62).



Monsieur Phi, director (27th Mar. 2022). **Le Progrès Des Sciences Nous Condamne-t-Il à l’apocalypse ?** (Pages 63, 64).



Nangia, Nikita, Clara Vania, Rasika Bhalerao and Samuel R. Bowman (Nov. 2020). “CrowS-Pairs: A Challenge Dataset for Measuring Social Biases in Masked Language Models”. In: **Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)**. EMNLP 2020. Online: Association for Computational Linguistics,

pp. 1953–1967. DOI:
[10.18653/v1/2020.emnlp-main.154](https://doi.org/10.18653/v1/2020.emnlp-main.154)
 (pages 35–38, 65–68).



Pascual, Manuel G. (26th June 2022). “LaMDA, the Machine That Is like ‘a Seven-Year-Old Kid’: Can a Computer Have Consciousness?” In: **EL PAÍS English Edition. Science & Tech** (pages 52–54).



Rudin, Cynthia (May 2019). “Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead”. In: **Nature machine intelligence** 1.5, pp. 206–215. ISSN: 2522-5839. DOI:

[10.1038/s42256-019-0048-x](https://doi.org/10.1038/s42256-019-0048-x). pmid:
 35603010 (pages 60–62).



Vincent, James (24th Mar. 2016). “Twitter Taught Microsoft’s Friendly AI Chatbot to Be a Racist Asshole in Less than a Day”. In: **The Verge** (pages 30, 31).



Weizenbaum, Joseph (1st Jan. 1966). “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine”. In: **Communications of the ACM** 9.1, pp. 36–45. ISSN: 0001-0782. DOI:
[10.1145/365153.365168](https://doi.org/10.1145/365153.365168)
 (pages 49–51, 55–57).