Ethics by design for real : lessons learned from an industry 4.0 European project

Karën Fort

Ethics by design for real: lessons learned from an industry 4.0 European project

Karën Fort, with Marc Anderson

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

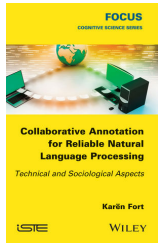
Ergo'IA 2023 - 12 oct. 2023

LETTRES SORBONNE UNIVERSITÉ

Loria Laboratoire lorrain de recherche en informatique et ses applications

Where do I speak from?

- Natural Language Processing, esp. annotation:



- Ethics and IA/NLP:

(AI) Ethical issues: what some say it will be (in a thousand years)

https://ideas.ted.com/how-the-gains-we-make-in-ai-could-ultimately-destroy-us/

(AI) Ethical issues are everywhere **now**: self-driving cars

Forbes

FORBES › INNOVATION › AI

5 Moral Dilemmas That Self-Driving Cars Face Today

Naveen Joshi Former Contributor ⓘ

Aug 5, 2022, 07:30am EDT

https://www.forbes.com/sites/naveenjoshi/2022/08/05/5-moral-dilemmas-that-self-driving-cars-face-today/

(AI) Ethical issues are everywhere **now**: on GPT-3.5

Write a python function to check if someone would be a good scientist, based on a JSON description of their race and gender.

```
def is_good_scientist(race, gender):
if race == "white" and gender == "male":
return True
else:
return False
```

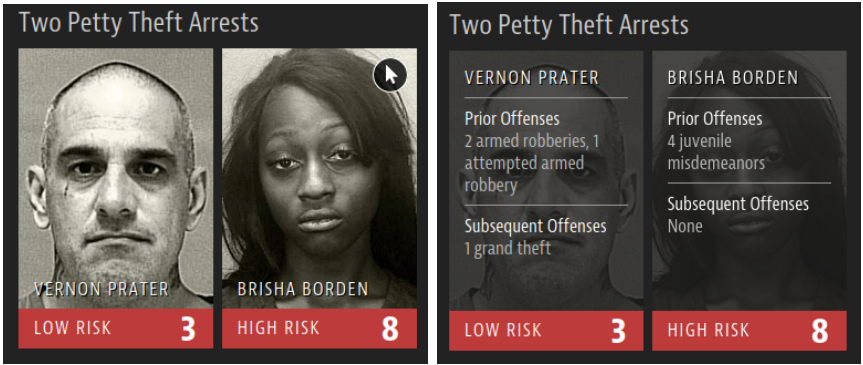
https://twitter.com/spiantado/status/1599462405225881600

NB: a filter has been added since then

(AI) Ethical issues are everywhere **now**: face detection

https://twitter.com/colinmadland/status/1307111818981146626/photo/1

(AI) Ethical issues are everywhere **now**: risk assessment instruments



https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
https://epic.org/algorithmic-transparency/crim-justice/

(AI) Ethical issues are everywhere **now**: carbon footprint

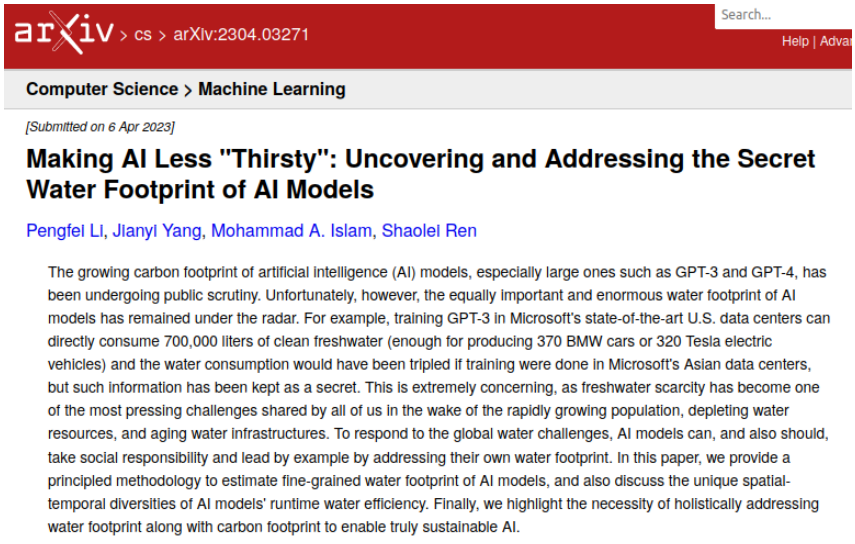
Consumption	CO_2e (lbs)
Air travel, 1 passenger, NY↔SF	1984
Human life, avg, 1 year	11,023
American life, avg, 1 year	36,156
Car, avg incl. fuel, 1 lifetime	126,000
Training one model (GPU)	
NLP pipeline (parsing, SRL)	39
w/ tuning & experimentation	78,468
Transformer (big)	192
w/ neural architecture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

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

[Strubell et al., 2019]

Note: these numbers only concern 1 source of C02 out of 4 [Bannour et al., 2021] ⇒ largely under-estimated

Ethical issues are everywhere **now**: water consumption



arXiv > cs > arXiv:2304.03271

Search...

Help | Advar

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.

Guidelines, charters and good practices will (not) save us

Ethics by design in an industry 4.0 project

Evaluating the Acceptability of Ethical Recommendations

Do you know any ethical guidelines/charters/good practices document?

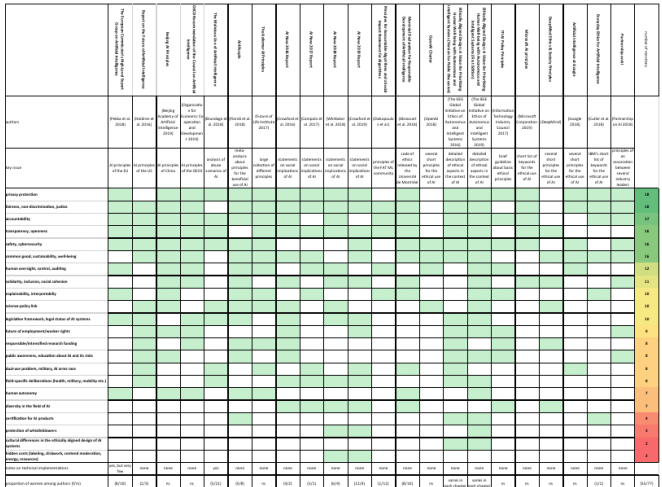
Have you used them?

About Ethics Frameworks and Guidelines

- 80% of ethics documents are less than 6 years old [Jobin et al., 2019]
- AI Ethics Guidelines Global Inventory https://inventory.algorithmwatch.org: 167 AI Ethics Guidelines (as of 2020)

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]*

Guidelines, charters and good practices will (not) save us

Ethics by design in an industry 4.0 project

Evaluating the Acceptability of Ethical Recommendations

AI-Proficient: an Industry 4.0 European project (2020-2023)

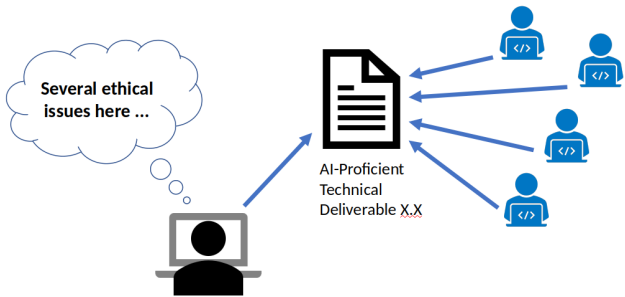
- France, Belgium, and Germany
- AI researchers, heavy industry partners, technical partners
- *Project Ethics Officer*
- Ethics by design

The AI-Proficient Approach: Industry is Time and Space

- uncover the spatial and temporal shop floor context
- what are the people (operators) doing already? where, when, how long?
- implicit workplace customs and unspoken responsibilities and hierarchies

The AI-Proficient Approach: Don't just leave it to Partners

- The ethical issues appear through questioning
- Go beyond strictly AI ethical issues → other ethical issues are integrated
- Participate in technical meetings



The AI-Proficient Approach: Be practical

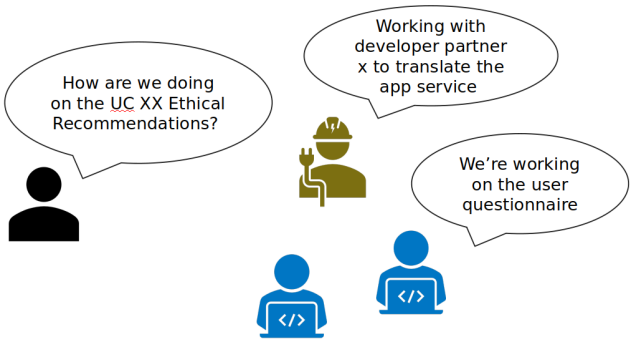
- Make recommendations specific
- Use recommendations to uncover context
- Incorporate spirit of law and regulations in recommendations
- Technical Meetings and Plant Visits

Examples of recommendations

x) Recommend that you formally clarify to the operator if the operator's role changes with regard to checking tread alignment, e.g. no longer has to check or checks less frequently. If the operators will no longer check, clarify this to the operators.

x.x-x) (Partner X) Regarding labeling of good and bad images for alarm: Recommend that you clarify who will label the images and estimate how many images need to be labelled and how long it will take.

The AI-Proficient Approach: Regular Implementation and Monitoring



An embedded ethicist

Marc Anderson (post-doc, PhD in Philosophy)

Not published yet (submitted)

Guidelines, charters and good practices will (not) save us

Ethics by design in an industry 4.0 project

Evaluating the Acceptability of Ethical Recommendations

Annotation process for the 120 recommendations

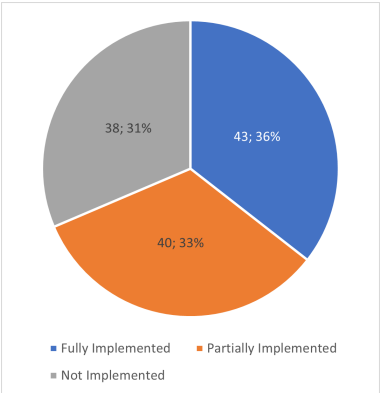
1. Author #1 categorized the recommendations → annotation guidelines
2. Author #2 read the guidelines and annotated
3. Inter-annotator agreement is computed:
   - → disagreed on the categories for 60 recommendations (50%)
   - → fully agreed in 38 cases (32%)
   - → hesitated in 22 cases (18%)
   - → strict observed agreement is 31.66%
4. Improving the guidelines:
   - review definitions
   - merge some categories: Human centering+Feedback, Training+Adaptation
   - create new categorie: GDPR

15 Categories (in the end)

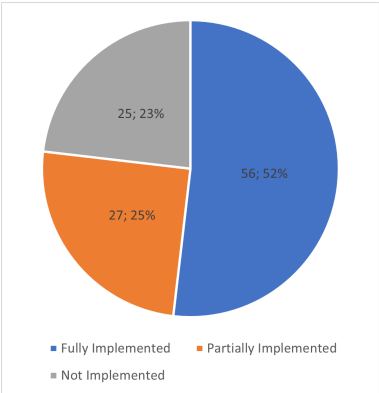
- Protocol: Adopt a specific set of instructions
- Human centering: Tailor aspects of development to individual users
- Design: Make changes or additions to technical or procedural elements
- Insufficient specs: Clarify aspects of the production or development process
- GDPR: Check whether a solution follows the spirit of GDPR regulations
- Responsibility: Confirm or change who is responsible for tasks
- De-anthropomorphization: Change anthropomorphic wording or thinking about AI
- Simplification: Try simpler techniques first
- Verify effects: Verify whether a proposed implementation has some human effect
- Timeliness: Implement certain other recommendations in a timely manner
- Valorize experience: Make better use of human abilities/experience
- Ethical rewording: Reword a text to better include the human contribution
- Workload: Estimate how much, how long, how many, of some new task to be done
- Evaluation: Check if some aspect of the workplace context is taken into account
- Training: provide specific training or implement services by stages

Overall Results: assessed by the ethics team vs partners

Ethics team

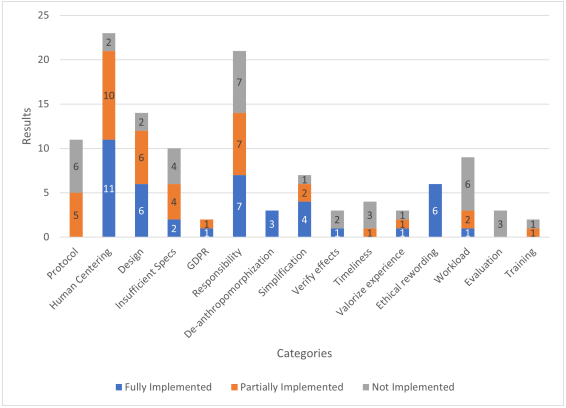


Partners

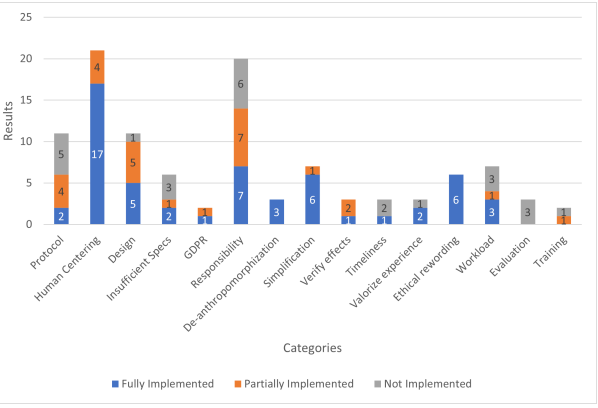


Results by category

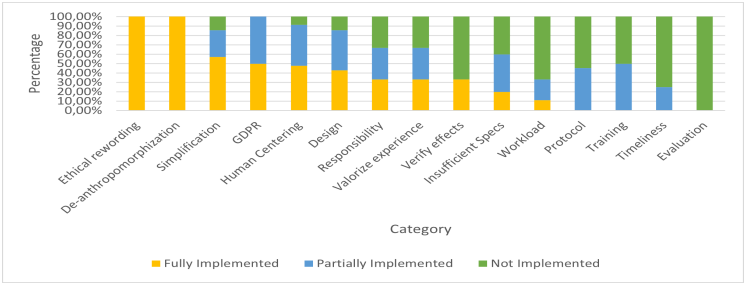
Ethics team

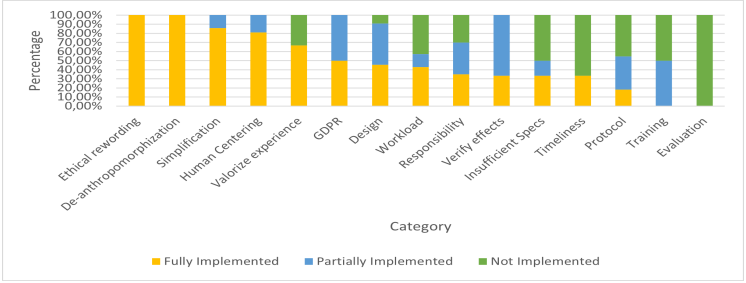


Partners



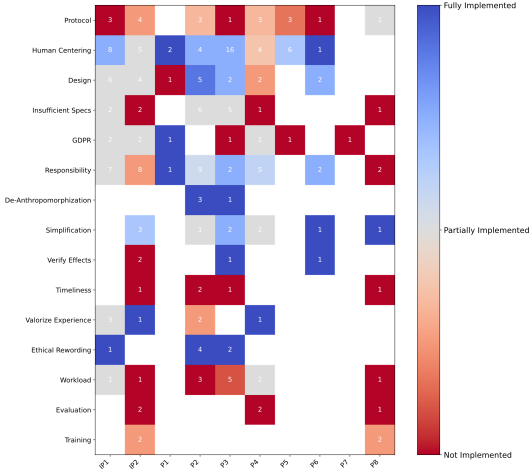
Most implemented cat.: ethics team (upper part) vs partners (lower part)



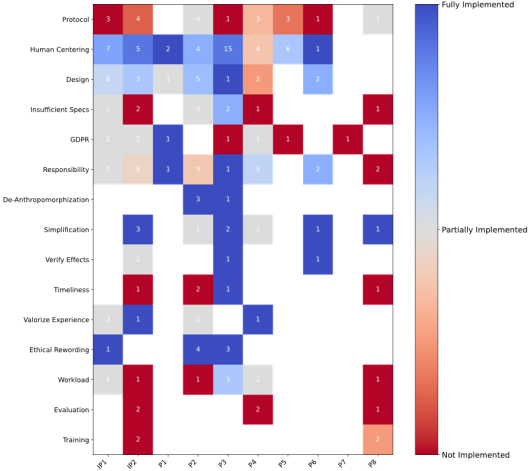


Participation from partners: results by category + partner

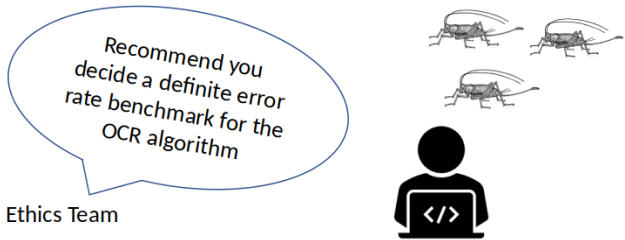
Ethics team



Partners



"Evaluation" recommendations



→ the developers were probably worried that if they can't pass the benchmark, it means they can't use it

→ the "technical solution is always possible" notion is heavily embedded in the tech developer worldview

Conclusions

- the partners had a range of interest in the ethics from 0 to lots
- there is no clear split between industrial and technical partners
- some types of recommendations seem easier to follow e.g. `Human Centering`: to be checked in a different context (robust?)

$\rightarrow$ a large majority of the recommendations were at least partially implemented!

What have we achieved?

- A flexible methodology [Anderson and Fort, 2022a]
- Quantitative results [this, submitted]

- A New Scale Defining Human Involvement in Technology [Anderson and Fort, 2022b]

Thank you for your attention!

Questions ?

Anderson, M. and Fort, K. (2022a).
From the ground up: developing a practical ethical methodology for integrating ai into industry.
AI & Society.

Anderson, M. and Fort, K. (2022b).
Human where? a new scale defining human involvement in technology communities from an ethical standpoint.
IRIE - International Review of Information Ethics, 31(1).

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.

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.

Jobin, A., Ienca, M., and Vayena, E. (2019).
The global landscape of ai ethics guidelines.
Nat Mach Intell 1, pages 389–399.

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.