

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

Karën Fort

▶ To cite this version:

Karën Fort. Ethics by design for real: lessons learned from an industry 4.0 European project. ERGO'IA, Oct 2023, Biarritz (France), France. hal-04366784

HAL Id: hal-04366784 https://inria.hal.science/hal-04366784v1

Submitted on 29 Dec 2023

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.



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/

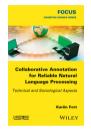
Ergo'lA 2023 - 12 oct. 2023





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)



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

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



(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

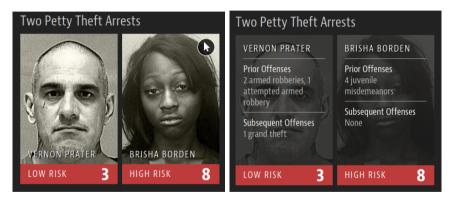
NB: a filter has been added since then

(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

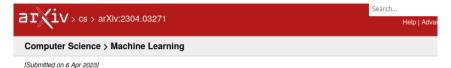
CO ₂ e (lbs)
1984
11,023
36,156
126,000
39
39 78,468

Table 1: Estimated CO₂ 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] \Rightarrow largely under-estimated

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



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

Pengfel LI, Jianyi Yang, Mohammad A. Islam, Shaolel 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 Tesia 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

	The European Committed of a Might level Street Group on Artificial Intelligence	Payers on the Patene of helifished traditionals	Belging At Straight	OKS feson medition of the Coast for Artificial You'll proce	The Malhanus Dur of Artificial Intelligence	Asthquis	The informal all heights	At New 2018 Report	Al New 2017 Report	At New 2018 Report	A New 2018 Report	Pind pin ter hacurable Algoritem and allocki Impact transmore for Algoritems	Manitral End waters for Supervalidal Servicement of Artificial riching root	OperAt Chafus	Rhindy Algord Drugs, a Vision for Prioritising Human Philips of the Autocompa and Institutes from a Charles for Robb the united	Bhish Alped Depn. A Valor for Portising Huma Well of ity with Autoconous and Intelligent System (Fact Sotion)	TIM Publy Principles	Misra di Ni pincipia	Despited Chie v & Stoke's Shorights	Artikal i telligene a Gugle	Grouping Other for Amiffold bendigmon	Partershipsonki	number of months
ection	(Pelda et al 2008)	(Malabon at at 2000k)	(Ref)ing Academy of AmPhila Statispecia 2016)	physicistic a for framonic for spenation and framologistics 1,2650)	Brandage et al. 2016	(Norsa et al. 2018)	(future of Life technics 2017)	(Growflend et al. 2016)	(Campula et al. 2017)	(AA Maler et al. 2018)	(Combosi et al. 2009)	(Orah apos de c et al ij	pikriesari et al. 2018)	(Species 2018)	(The ISS Global states and Issue of Actorisms and Issue of Systems (Systems 2000)	(the Bill Shilled Britisher on Emiss of Autonomia and Vindigent Systems (600)	polamodian Yechnology Industry Countil 2017)	(Microsoft Corporation 1800)	(Seq#Srel)	pinagh 2000)	(Salie et al 2010)	grammenship on #120000	
Mylima	Argerinolylin of the EX	All principles of the US	Ar priniciples afficiens	Alprinsple of the OSSI	analysis of allows committee of A	realizate presipire administrative for the benealizate period for the benealizate period for the	loge colonian of Effects principles	of demonstra on social originations of Ar	clidarocts on-oxid orgalizations of Al	es escial englisations of Al	Cidenatis or code implications of Al	principles of the FAT AAL sommunity	rade of ethics released by the prisonald de Monteira	onered short principles for the efficial sea of No.	described description of ethical aspects in the context of its	deschiel description of ethical aspects in the control of A)	brief guideline alleut bank ethicul principles	short list of largeriets for the efficial use of Al	second short principles for the otherd was affect	short short principles for the ethical use of Al	letter short let of taywords for the ethical year of Ar	principles of an executation between reversi industry leaders	
private protection																							18
Consess, near discrimination, justice																							18
**********																							12
елегриналу, применя																							16
safety cyberoscoty																							16
common good, sustainability, well-being																							16
Ruman exercigits, control, tealthing																							12
solidarity, indusius, cacial calendos	-																			-			11
replaced the interpretability																							10
nteres pelicy link																							
high lather Transpace's, bigod et must of Air systems.																					-	-	
Meuro of employment/worker rights	-																						
esponsible/intensified-research-funding																							
public presences, education about 10 and its risks																							
Badrace problem, rollbary, M. arms, race																							
ticle specific deliberations (health, military, matelly ats.)																							
																							7
Electrity is the Sold of Al			$\overline{}$							-							-			-			7
ortflofee for N products																							
protection of which blowers																							
cultural differences in the ethically aligned design of its	$\overline{}$		-	-		-		$\overline{}$										-		-	-		1
hidden szett (labeling, dichwork, control medoration, manys, manarom)	-									-										-	-		
erse en tadricalizaplaneetations	on, but see	2004	reso	none	yes	cene	5004	nese	nees	name	rees	5574	reso	5574	rare	70088	rare	nese	nore	nese	nema	8.97M	
popertion of wones arrang authors (f/m)	9/10)	(3/9)	81	16	(6/21)	(5/8)	ns.	(42)	(8/3)	(4)	[12/4]	(1/12)	(6/10)	rs	varies in each chapter	sories is each chapter	86	11	ns.	н	0/21	*	(64/77)
erght (number of world) effication (government, industry, science)	10544	12747	766	1044	36617	0000	141	11131	18273	20164	33074	1358	4714		AURIL	114/3/2	3311	- 74	867		****		-
efication (government, industry, science)	government	posteres	gos/e4.	government.	Lúnce	science	science	science	science	sómo	sórece	nime	SCHEEK	nes profe	indutry	indutty	induttry	infumy	indutry	*******	induttry	indutty	1

[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

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



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

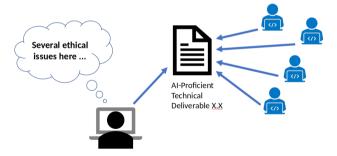
The Al-Proficient Approach: Industry is Time and Space



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

- ► The ethical issues appear through questioning
- \blacktriangleright Go beyond strictly AI ethical issues \rightarrow other ethical issues are integrated
- ► Participate in technical meetings



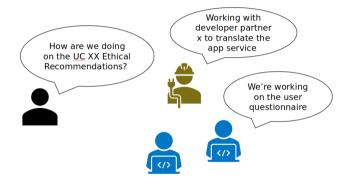
The Al-Proficient Approach: Be practical

- ► 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 Al-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

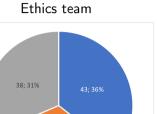
Annotation process for the 120 recommendations

- 1. Author #1 categorized the recommendations \rightarrow annotation guidelines
- 2. Author #2 read the guidelines and annotated
- 3. Inter-annotator agreement is computed:
 - → disagreed on the categories for 60 recommendations (50%)
 - \rightarrow fully agreed in 38 cases (32%)
 - \rightarrow 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

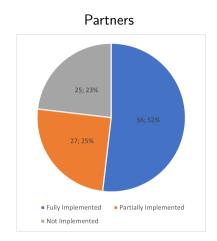


■ Partially Implemented

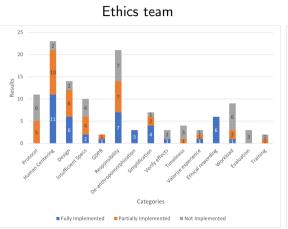
40: 33%

■ Fully Implemented

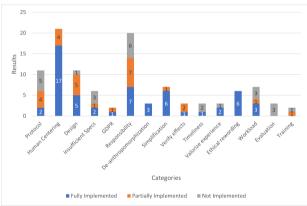
■ Not Implemented



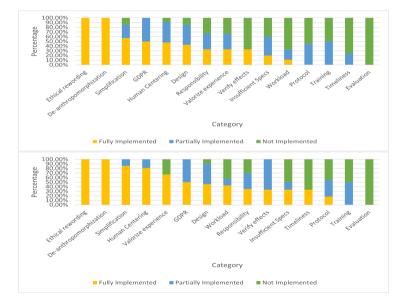
Results by category



Partners

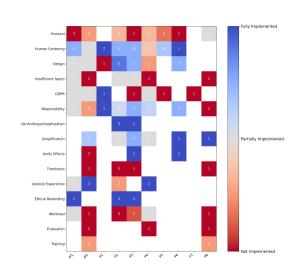


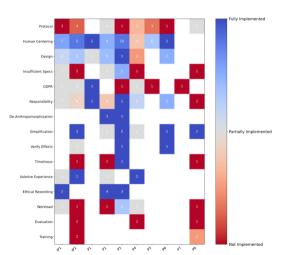
Most implemented cat.: ethics team (upper part) vs partners (lower part)



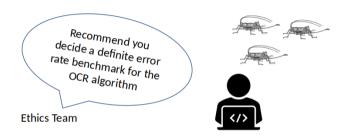
Participation from partners: results by category + partner

Ethics team Partners





"Evaluation" recommendations



- ightarrow the developers were probably worried that if they can't pass the benchmark, it means they can't use it
- \rightarrow 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?)
- ightarrow 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.

Al & Society.

tools.

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

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).

Linguistics.

- 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