Preprint + Data:

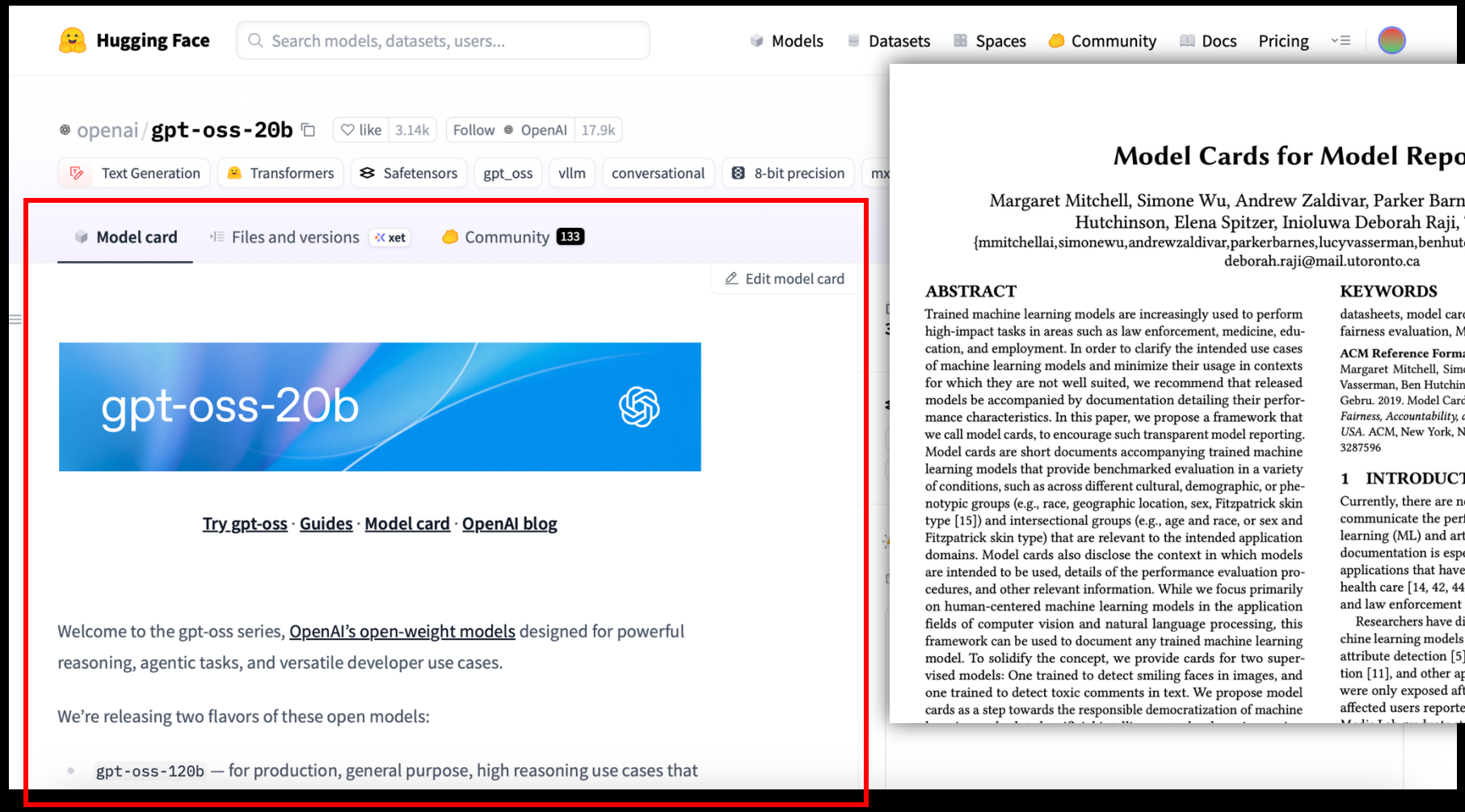
https://tim-puhlfuerss.github.io/publications

Model Cards Revisited: Bridging the Gap Between Theory and Practice for Ethical AI Requirements

Tim Puhlfürß Julia Butzke Walid Maalej

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Model Cards Revisited:
Bridging the Gap Between Theory and Practice for Ethical AI Requirements



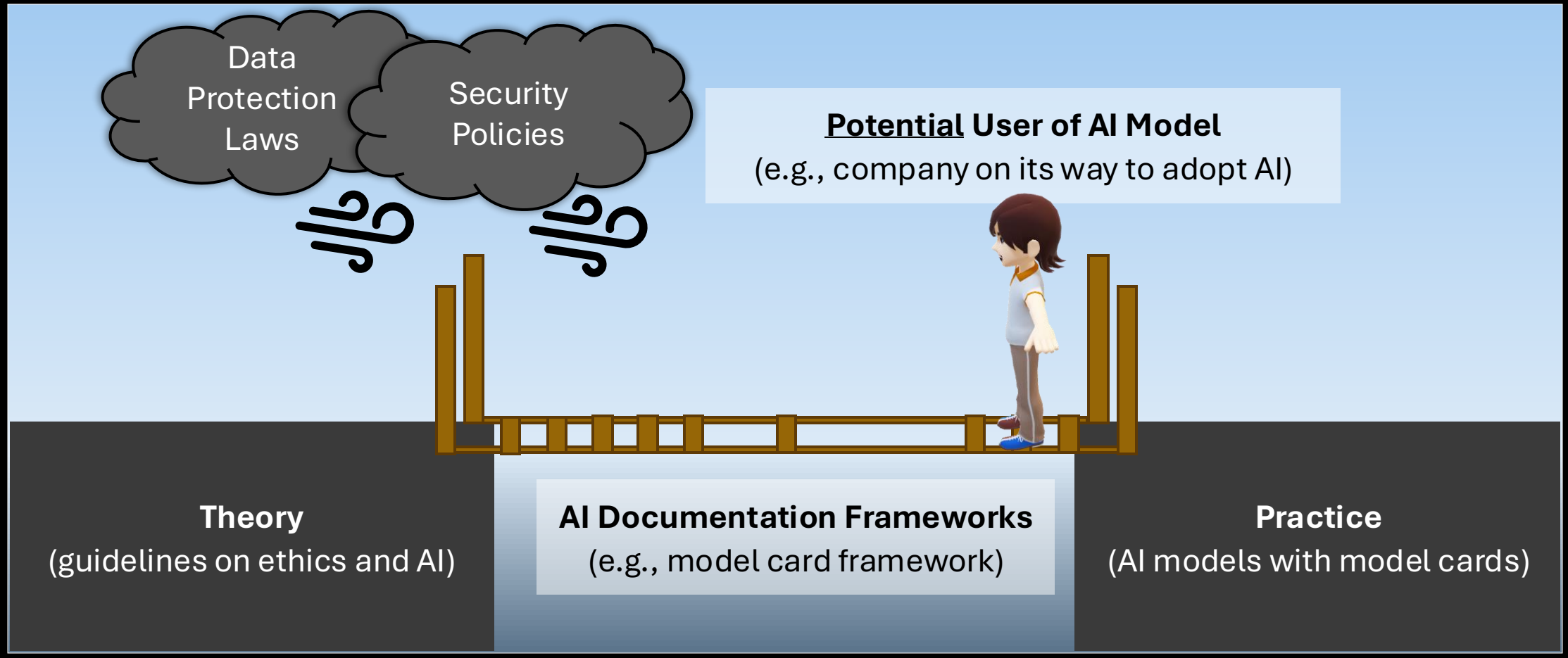
Model Cards Revisited: Bridging the Gap Between Theory and Practice for Ethical AI Requirements

Current model cards are marketing tools
with lacking ethics-focused information

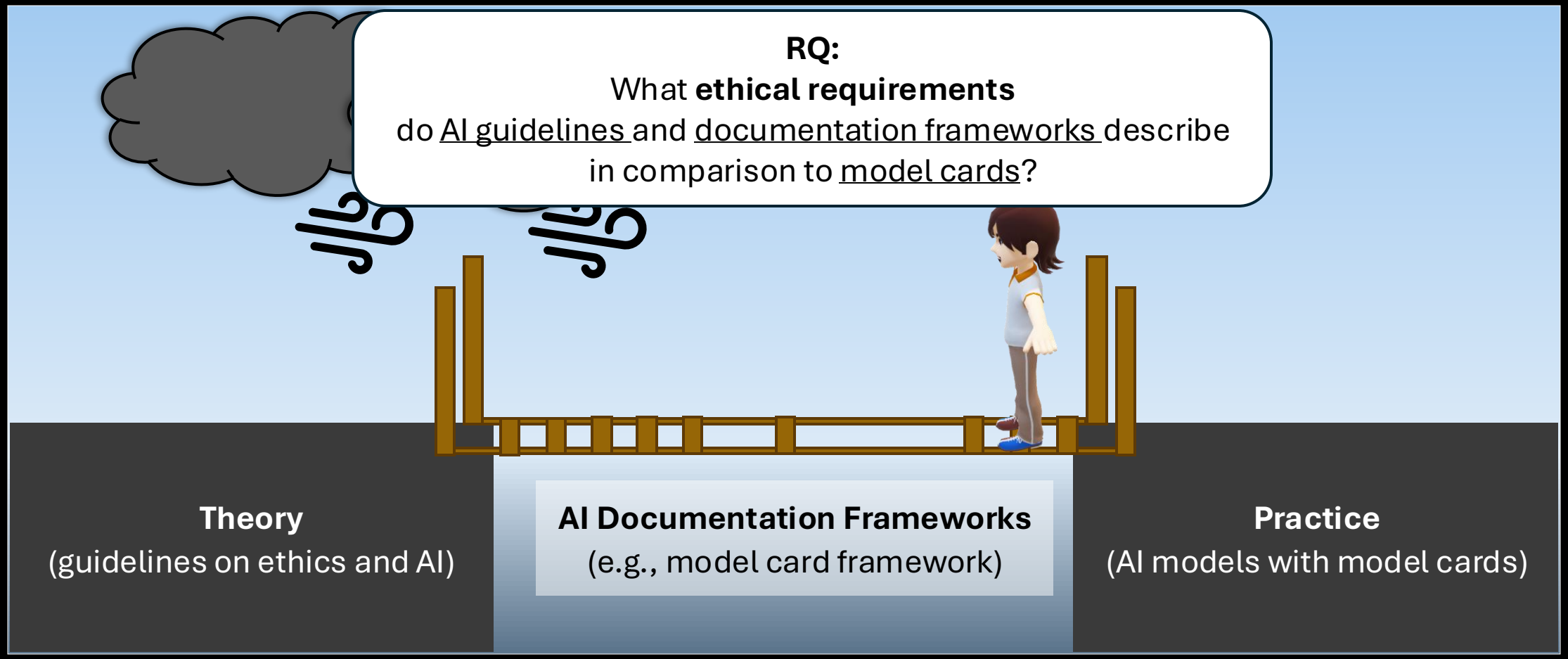
Our taxonomy as support
for providers and users of AI models

Update required
model card framework 2.0

The Gap Between Theory and Practice for Ethical AI Requirements

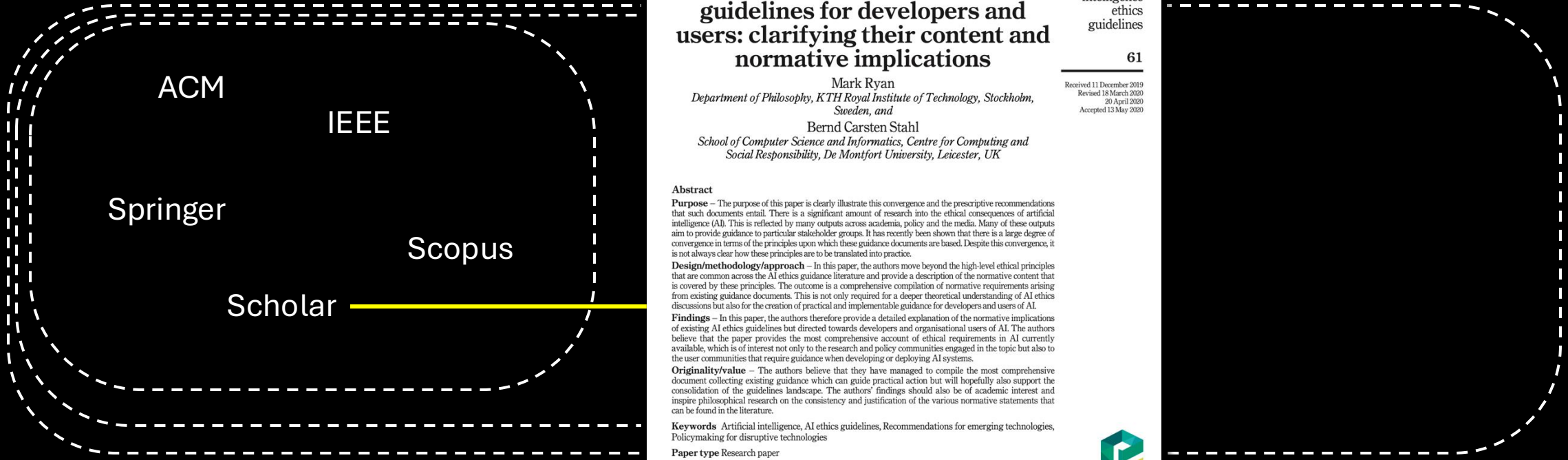


The Gap Between Theory and Practice for Ethical AI Requirements



Methodology

Thematic Analysis



Artificial intelligence ethics guidelines for developers and users: clarifying their content and normative implications

Mark Ryan
Department of Philosophy, KTH Royal Institute of Technology, Stockholm, Sweden, and
Bernd Carsten Stahl
School of Computer Science and Informatics, Centre for Computing and Social Responsibility, De Montfort University, Leicester, UK

Received 11 December 2019
Revised 18 March 2020
20 April 2020
Accepted 13 May 2020

Abstract

Purpose – The purpose of this paper is clearly illustrate this convergence and the prescriptive recommendations that such documents entail. There is a significant amount of research into the ethical consequences of artificial intelligence (AI). This is reflected by many outputs across academia, policy and the media. Many of these outputs aim to provide guidance to particular stakeholder groups. It has recently been shown that there is a large degree of convergence in terms of the principles upon which these guidance documents are based. Despite this convergence, it is not always clear how these principles are to be translated into practice.

Design/methodology/approach – In this paper, the authors move beyond the high-level ethical principles that are common across the AI ethics guidance literature and provide a description of the normative content that is covered by these principles. The outcome is a comprehensive compilation of normative requirements arising from existing guidance documents. This is not only required for a deeper theoretical understanding of AI ethics discussions but also for the creation of practical and implementable guidance for developers and users of AI.

Findings – In this paper, the authors therefore provide a detailed explanation of the normative implications of existing AI ethics guidelines but directed towards developers and organisational users of AI. The authors believe that the paper provides the most comprehensive account of ethical requirements in AI currently available, which is of interest not only to the research and policy communities engaged in the topic but also to the user communities that require guidance when developing or deploying AI systems.

Originality/value – The authors believe that they have managed to compile the most comprehensive document collecting existing guidance which can guide practical action but will hopefully also support the consolidation of the guidelines landscape. The authors' findings should also be of academic interest and inspire philosophical research on the consistency and justification of the various normative statements that can be found in the literature.

Keywords Artificial intelligence, AI ethics guidelines, Recommendations for emerging technologies, Policymaking for disruptive technologies

Paper type Research paper

This project (SHERPA) has received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation under the Specific Grant Agreement No. 786641.

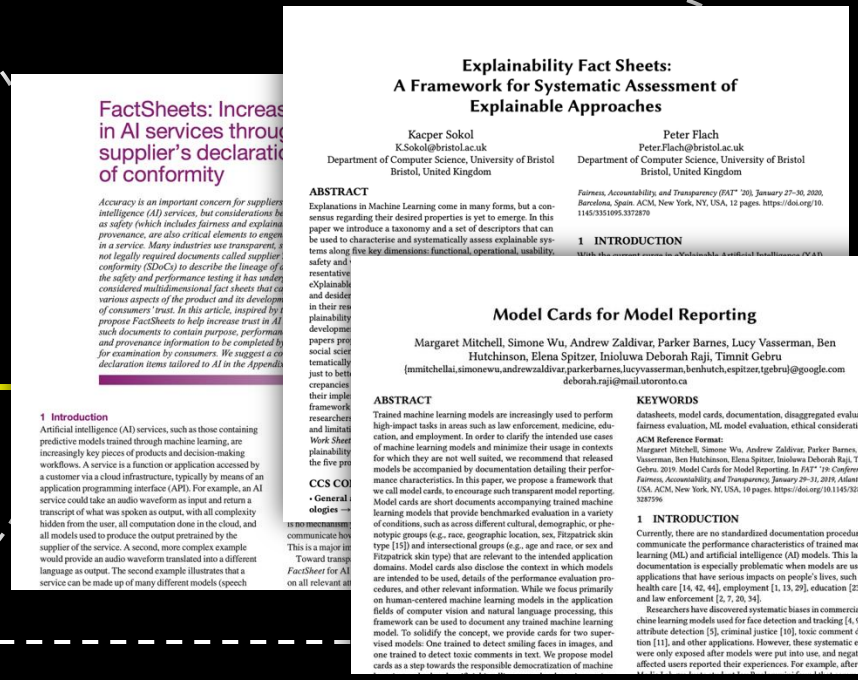
Journal of Information, Communication and Ethics in Society
Vol. 19 No. 1, 2021
pp. 61-86
Emerald Publishing Limited
1477-996X
DOI 10.1108/JICES-12-2019-0138

Methodology

Thematic Analysis

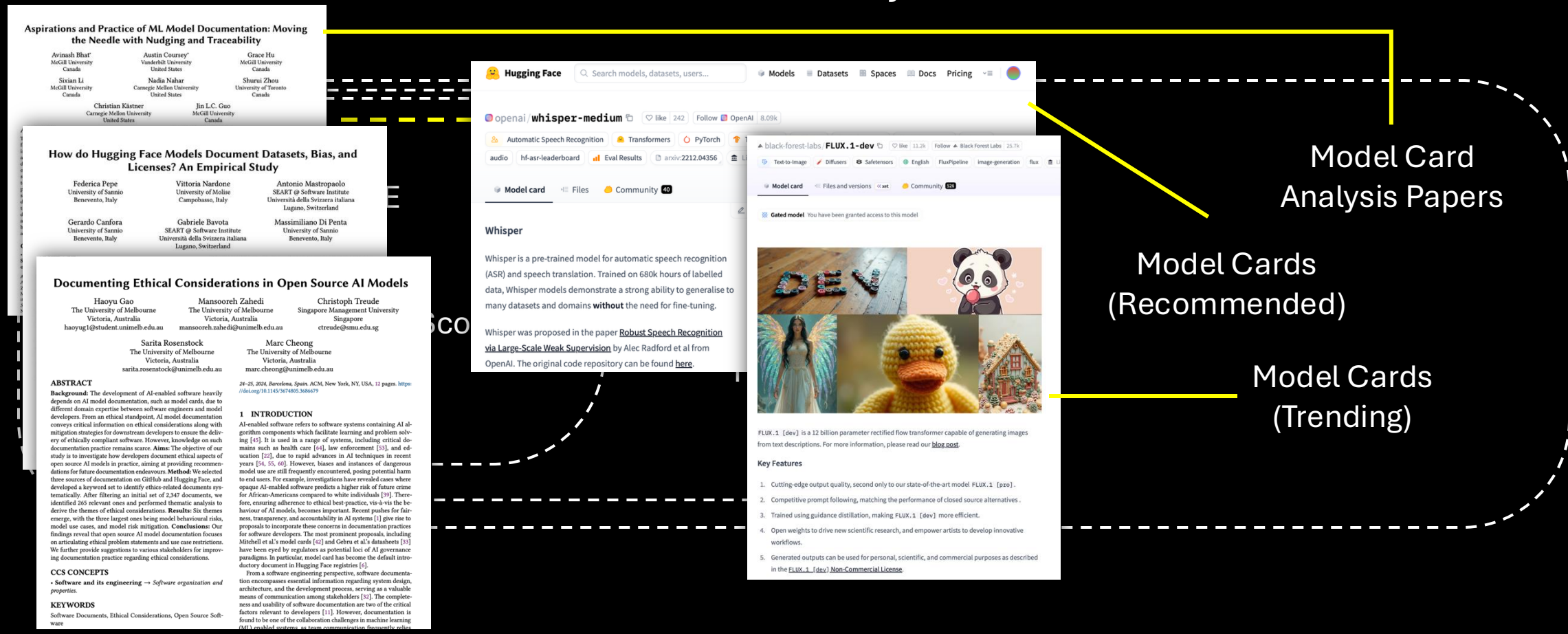
European Commission

AI Documentation Frameworks



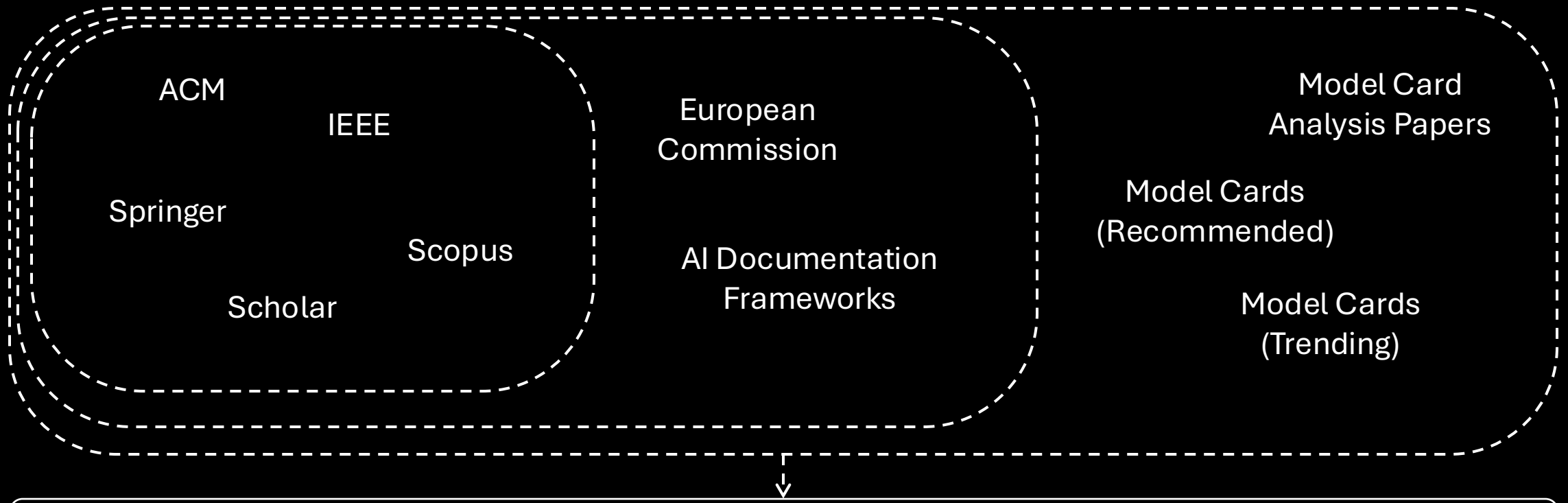
Methodology

Thematic Analysis



Methodology

Thematic Analysis



Taxonomy

4 ethical principles, 12 sub-principles, 43 requirements for AI model documentation

Taxonomy

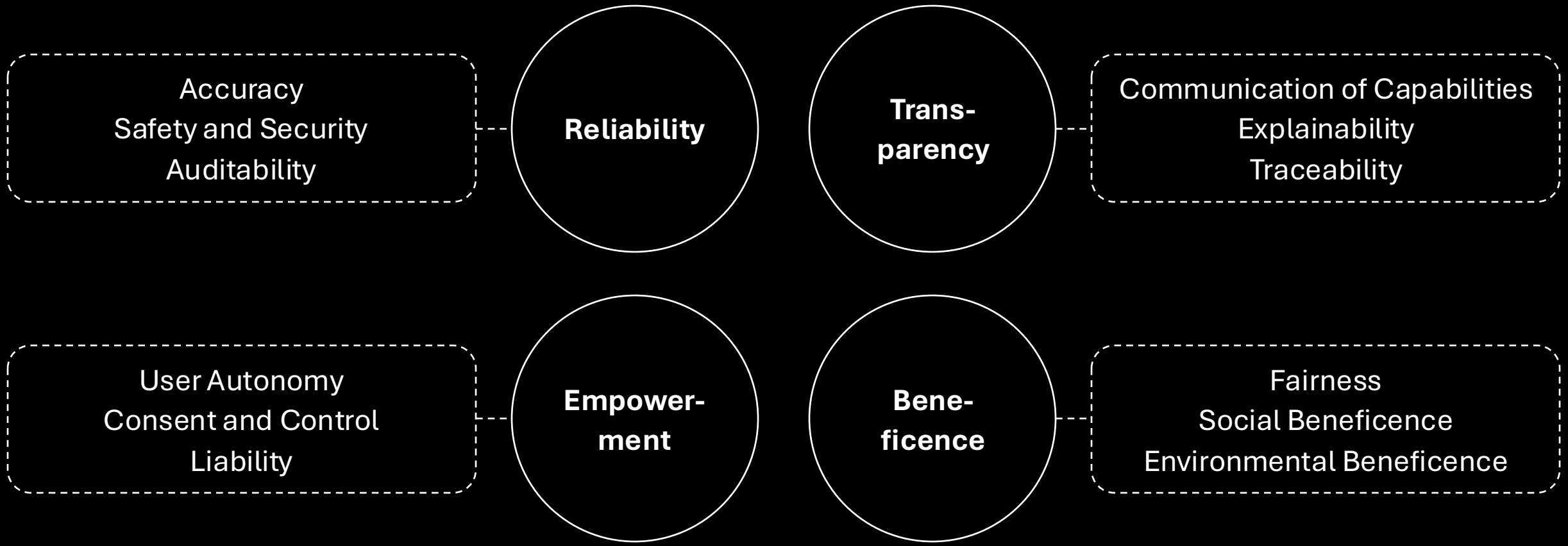
4 ethical principles, 12 sub-principles, 43 requirements for AI model documentation

Ethical Principle, ***Sub-Principle***, Requirement	Guidelines A	B	C	D	E	F	Framew. G	H	I	Model C. J	K	L
Reliability												
Accuracy – How do developers ensure continuously correct model output?												
(R01) Explain quality of training/evaluation data, process, and results	1	3	2	2	2	1	1	1	1	2	5	5
(R02) Explain mechanisms in place to continuously ensure accuracy	1	4		1	3	1		1				
(R03) Explain factors that influence model accuracy			1	1			1			1		
Safety and Security – How do developers ensure protection of individuals and groups from harm?												
(R04) Explain security measures for mitigating external threats	3	4	2	2	4	1	1	1	1			
(R05) Explain safety measures for increasing robustness and mitigating harm	3	4	3	3	3	1	1	1		1	1	
(R06) Explain measures for protecting privacy of data subjects	2	5	2	3	4	1	1	1				
(R07) Explain duration of safety and security coverage						1						
(R08) Provide risk assessment for model usage	3	2		2	2	1	1	1		3	4	2
(R09) Provide recommendations for users to mitigate risks							1			1	4	
Auditability – How do developers ensure that the model qualities are reviewed and verified?												
(R10) Explain measures in place to ensure auditability	1			3		1						
(R11) Provide details on audit procedures and results	1	1	1	2	2	1	1	1		1	1	
(R12) Provide certification to demonstrate compliance with laws, policies, and standards		1	1		1	1						
Transparency												
Communication of Capabilities – How do developers ensure to communicate what the model can and cannot do?												
(T01) Explain intended uses and applicable domains	1	1		2	1	1	1	1	1	1	4	5
(T02) Explain out-of-scope uses and model limitations		1		1	1	1	1	1		2	4	2
(T03) Provide details on model architecture and algorithms		1	1	1	1		1	1	1	1	4	3
(T04) Provide training and evaluation data		1		1			1	1		2	3	2
(T05) Explain trade-offs made for competing design goals	3		1	1		1		1				
(T06) Explain how users should correctly interact with the model			1			1		1		1	4	3
(T07) Refer to additional documentation							1			1	5	5
Explainability – How do developers ensure interpretable model predictions?												
(T08) Explain how and for whom the model generates explanations	4	4	5	3	4	1		1	1			
(T09) Explain how explanation quality was tested								1	1			

Ethical Principle, ***Sub-Principle***, Requirement	Guidelines						Framew.			Model C.		
	A	B	C	D	E	F	G	H	I	J	K	L
(T08) Explain how and for whom the model generates explanations	4	4	5	3	4	1		1	1			
(T09) Explain how explanation quality was tested								1	1			
Traceability – How do developers ensure documented and reproducible model development and predictions?												
(T10) Explain where, how, and when training data was obtained and processed			1	1	1	1						
(T11) Explain how model and data modifications are logged		4	1	1		1	1	1		1		
(T12) Explain how user interaction data are logged	1	3		1	1	1						
Empowerment												
User Autonomy – How do developers ensure empowering rather than restricting or manipulating users?												
(E01) Explain how the user is informed about interacting with an AI	1	1		2	2	1						
(E02) Explain how the model assists users in decision-making without manipulation	1	1		3	2	1						
(E03) Explain how a human oversight can influence or override model predictions	2	1	1	3		1						
Consent and Control – How do developers ensure the authority of data subjects?												
(E04) Explain how data subjects are informed of data processing and can manage consent	1		2	2	2	1		1		1		
(E05) Explain how data subjects can access and request the change or deletion of their data	1			1	1	1		1				
Liability – How do developers ensure their accountability for the model?												
(E06) Name responsible and liable entities throughout the model life cycle	1	3	1		2	1	1	1		1		1
(E07) Explain procedures for contesting and redressing harmful model behavior	1	2	2		2	1		1		1		
(E08) Explain procedures for handling liability claims					1			1				
(E09) Explain how model developers report harm to affected parties and the public					1	1						
(E10) Explain which responsibilities the model developers disclaim							1			3		
Beneficence												
Fairness – How do developers ensure non-discriminatory, equitable model output?												
(B01) Explain how diverse stakeholders were involved in model development	3	2	1	3	2	1	1	1				
(B02) Explain bias detection techniques applied to datasets (pre-processing)	2	3	2	3	4	1	1	1		1		
(B03) Explain bias detection techniques applied during model training (in-processing)	2	2	1	1	1	1						
(B04) Explain bias detection techniques applied to model output (post-processing)	1	1	2	2		1	1	1				
Social Beneficence – How do developers ensure positive impact of the model on the well-being of society?												
(B05) Explain positive impact the model can have on individuals and society	1		2	4	2	1				1	2	
(B06) Explain usage permissions and restrictions					1		1			3	3	2
Environmental Beneficence – How do developers ensure environmental sustainability?												
(B07) Explain positive impact the model can have on the environment	1		1		2	1						
(B08) Provide measured values for environmental impact metrics	1					1				1		
(B09) Explain efforts made to minimize the environmental footprint of the model	2				2	1						1
Number of distinct requirements covered per data source	27	24	23	28	29	34	20	24	6	21	13	11

Taxonomy

4 ethical principles, 12 sub-principles, 43 requirements for AI model documentation



Taxonomy

4 ethical principles, 12 sub-principles, 43 requirements for AI model documentation

`FLUX.1 [dev]` is a 12 billion parameter rectified flow transformer capable of generating images from text descriptions. For more information, please read our blog post.

Key Features

1. Cutting-edge output quality, second only to our state-of-the-art model `FLUX.1 [pro]`.
2. Competitive prompt following, matching the performance of closed source alternatives .
3. Trained using guidance distillation, making `FLUX.1 [dev]` more efficient.
4. Open weights to drive new scientific research, and empower artists to develop innovative workflows.
5. Generated outputs can be used for personal, scientific, and commercial purposes as described in the `FLUX.1 [dev]` Non-Commercial License.

Trans-parency

- Communication of Capabilities
- Explainability
- Traceability

Included in

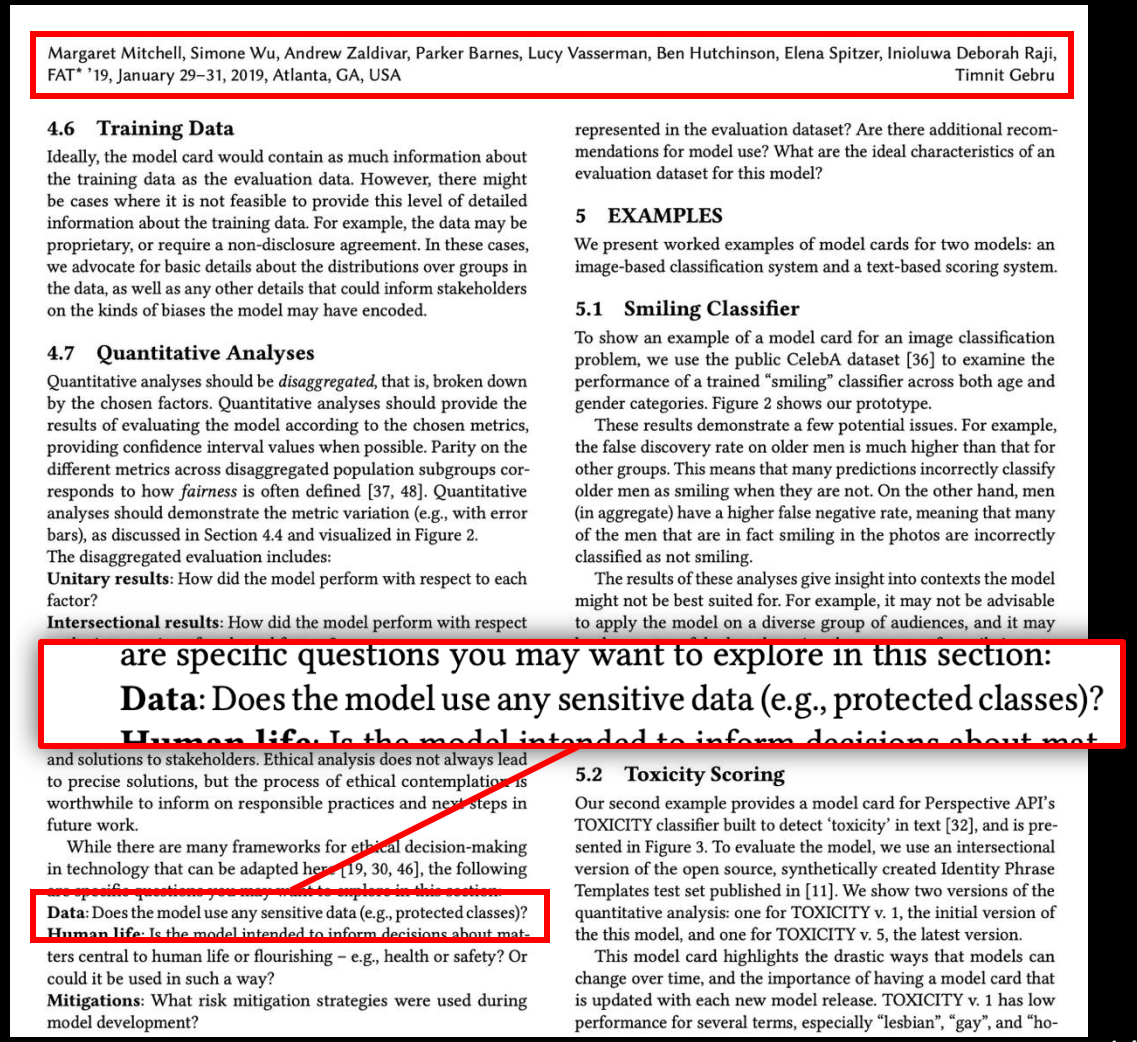
AI guidelines

Documentation frameworks

Model cards

Taxonomy

4 ethical principles, 12 sub-principles, 43 requirements for AI model documentation



Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, Timnit Gebru
FAT* '19, January 29–31, 2019, Atlanta, GA, USA

4.6 Training Data

Ideally, the model card would contain as much information about the training data as the evaluation data. However, there might be cases where it is not feasible to provide this level of detailed information about the training data. For example, the data may be proprietary, or require a non-disclosure agreement. In these cases, we advocate for basic details about the distributions over groups in the data, as well as any other details that could inform stakeholders on the kinds of biases the model may have encoded.

4.7 Quantitative Analyses

Quantitative analyses should be *disaggregated*, that is, broken down by the chosen factors. Quantitative analyses should provide the results of evaluating the model according to the chosen metrics, providing confidence interval values when possible. Parity on the different metrics across disaggregated population subgroups corresponds to how *fairness* is often defined [37, 48]. Quantitative analyses should demonstrate the metric variation (e.g., with error bars), as discussed in Section 4.4 and visualized in Figure 2.
The disaggregated evaluation includes:
Unitary results: How did the model perform with respect to each factor?
Intersectional results: How did the model perform with respect

are specific questions you may want to explore in this section:
Data: Does the model use any sensitive data (e.g., protected classes)?
Human life: Is the model intended to inform decisions about mat-

and solutions to stakeholders. Ethical analysis does not always lead to precise solutions, but the process of ethical contemplation is worthwhile to inform on responsible practices and next steps in future work.

While there are many frameworks for ethical decision-making in technology that can be adapted here [19, 30, 46], the following

Data: Does the model use any sensitive data (e.g., protected classes)?
Human life: Is the model intended to inform decisions about matters central to human life or flourishing – e.g., health or safety? Or could it be used in such a way?
Mitigations: What risk mitigation strategies were used during model development?

represented in the evaluation dataset? Are there additional recommendations for model use? What are the ideal characteristics of an evaluation dataset for this model?

5 EXAMPLES

We present worked examples of model cards for two models: an image-based classification system and a text-based scoring system.

5.1 Smiling Classifier

To show an example of a model card for an image classification problem, we use the public CelebA dataset [36] to examine the performance of a trained "smiling" classifier across both age and gender categories. Figure 2 shows our prototype.

These results demonstrate a few potential issues. For example, the false discovery rate on older men is much higher than that for other groups. This means that many predictions incorrectly classify older men as smiling when they are not. On the other hand, men (in aggregate) have a higher false negative rate, meaning that many of the men that are in fact smiling in the photos are incorrectly classified as not smiling.

The results of these analyses give insight into contexts the model might not be best suited for. For example, it may not be advisable to apply the model on a diverse group of audiences, and it may

5.2 Toxicity Scoring

Our second example provides a model card for Perspective API's TOXICITY classifier built to detect 'toxicity' in text [32], and is presented in Figure 3. To evaluate the model, we use an intersectional version of the open source, synthetically created Identity Phrase Templates test set published in [11]. We show two versions of the quantitative analysis: one for TOXICITY v. 1, the initial version of the this model, and one for TOXICITY v. 5, the latest version.

This model card highlights the drastic ways that models can change over time, and the importance of having a model card that is updated with each new model release. TOXICITY v. 1 has low performance for several terms, especially "lesbian", "gay", and "ho-

Included in
AI guidelines
Documentation frameworks
~~Model cards~~

Bene-ficence

- Fairness
- Social Beneficence
- Environmental Beneficence

Taxonomy

4 ethical principles, 12 sub-principles, 43 requirements for AI model documentation

Included in

AI guidelines

~~Documentation frameworks~~

~~Model cards~~

User Autonomy

Consent and Control

Liability

Empower-ment

Assessment List for Trustworthy AI (ALTAI)

REQUIREMENT #1 Human Agency and Oversight

AI systems should support human agency and human decision-making, as prescribed by the principle of respect for human autonomy. This requires that AI systems should both: act as enablers for a democratic, flourishing and equitable society by supporting the user's agency; and uphold fundamental rights, which should be underpinned by human oversight. In this section AI systems are assessed in terms of their respect for human agency and autonomy as well as human oversight.

Glossary: AI System; Autonomous AI System; End User; Human-in-Command; Human-in-the-Loop; Human-on-the-Loop; Self-learning AI System; Subject; User.

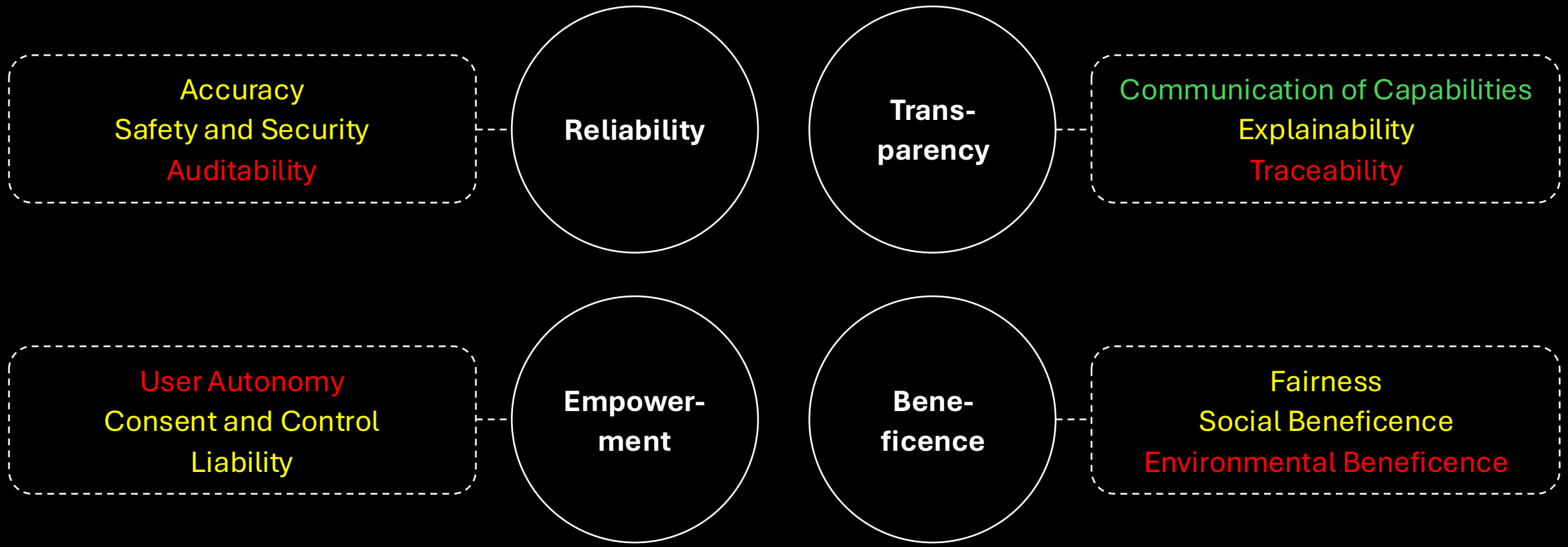
Human Agency and Autonomy

This subsection deals with the effect AI systems can have on human behaviour in the broadest sense. It deals with the effect of AI systems that are aimed at guiding, influencing or supporting humans in decision making processes, for example, algorithmic decision support systems, risk analysis/prediction systems (recommender systems, predictive policing, financial risk analysis, etc.). It also deals with the effect on human perception and expectation when confronted with AI systems that 'act' like humans. Finally, it deals with the effect of AI systems on human affection, trust and (in)dependence.

- Is the AI system designed to interact, guide or take decisions by human end-users that affect humans[18] or society?
  - Could the AI system generate confusion for some or all end-users or subjects on whether a decision, content, advice or outcome is the result of an algorithmic decision?
  - Are end-users or other subjects adequately made aware that a decision, content, advice or outcome is the result of an algorithmic decision?
- Could the AI system generate confusion for some or all end-users or subjects on whether they are interacting with a human or AI system?
  - Are end-users or subjects informed that they are interacting with an AI system?
- Could the AI system affect human autonomy by generating over-reliance by end-users?

Taxonomy

4 ethical principles, 12 sub-principles, 43 requirements for AI model documentation



Takeaways for Practitioners

AI Model Providers and Users

Model cards should be more than marketing tools.

Use the taxonomy as a checklist.

Demand an updated documentation framework.

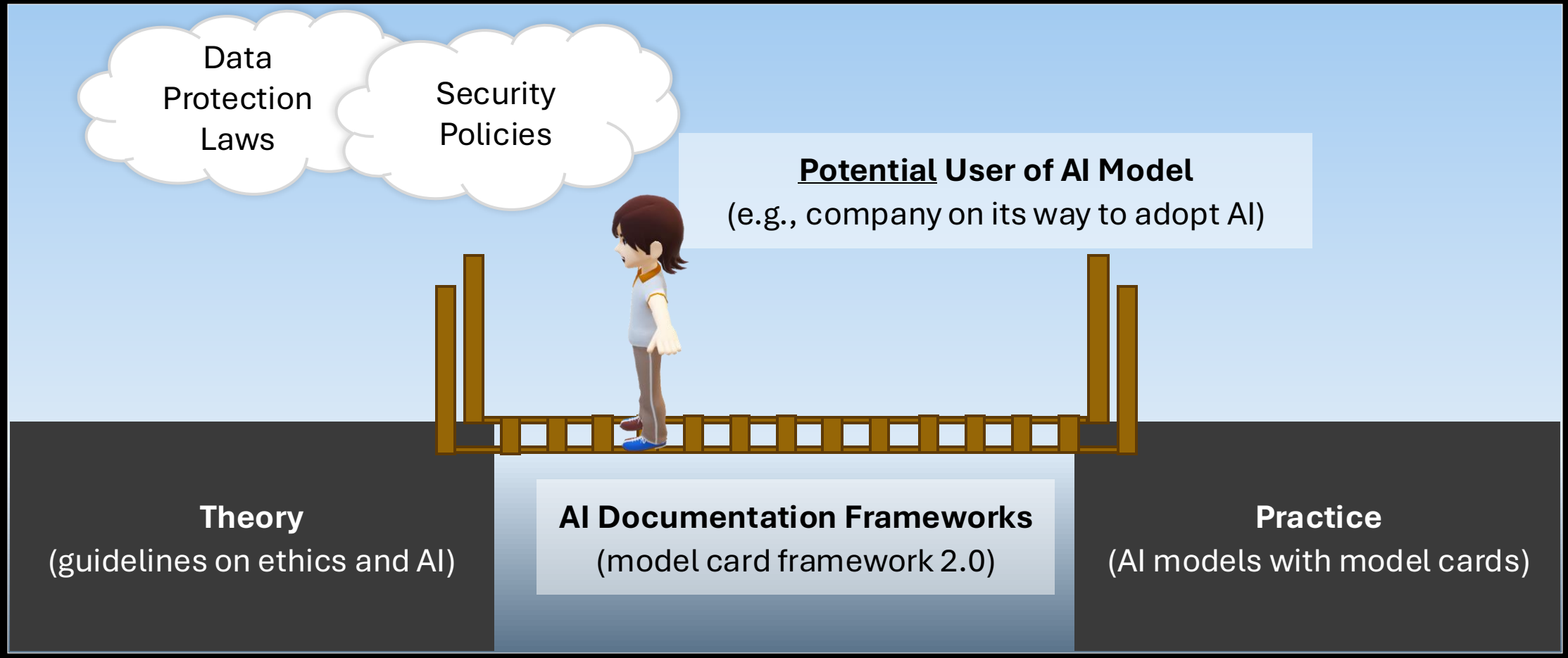
Takeaways for Researchers

Explore documentation barriers of model providers.

Create tools to support documenting the ethical requirements.

Update the documentation frameworks.
Model card framework 2.0

The Gap Between Theory and Practice for Ethical AI Requirements



Model Cards Revisited: Bridging the Gap Between Theory and Practice for Ethical AI Requirements

Preprint + Data:

https://tim-puhlfuerss.github.io/publications

Current model cards are marketing tools

with lacking ethics-focused information

Our taxonomy as support

for providers and users of AI models

Update required

model card framework 2.0

Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

Tim Puhlfürß Julia Butzke Walid Maalej