Responsible Machine Learning

Master MVA

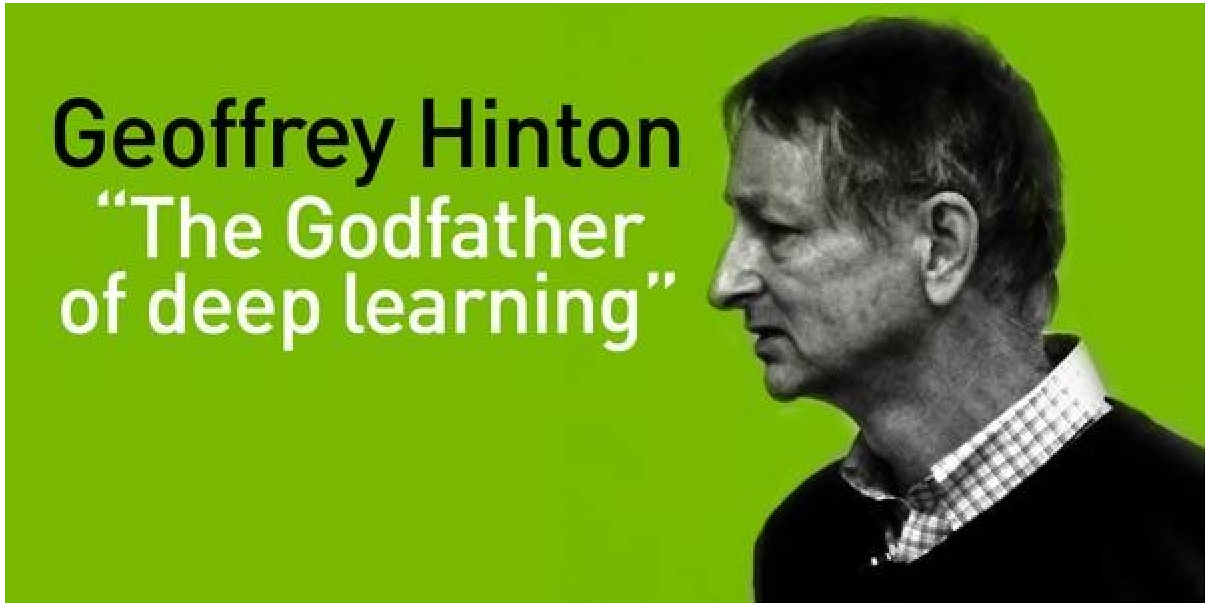
Introduction on the technical aspects of the course

Nicolas Vayatis

When scientists go to marketing...

"They should stop training radiologists now. It's just completely obvious within five years that deep learning is going to do better than radiologists. It might take ten years, but we've got plenty of radiologists already. I said this at a hospital, and it didn't go down too well."

— Geoffrey Hinton, Toronto, 2016



Source : < https://www.youtube.com/watch?v=2HMPRXstSvQ&t=2s >

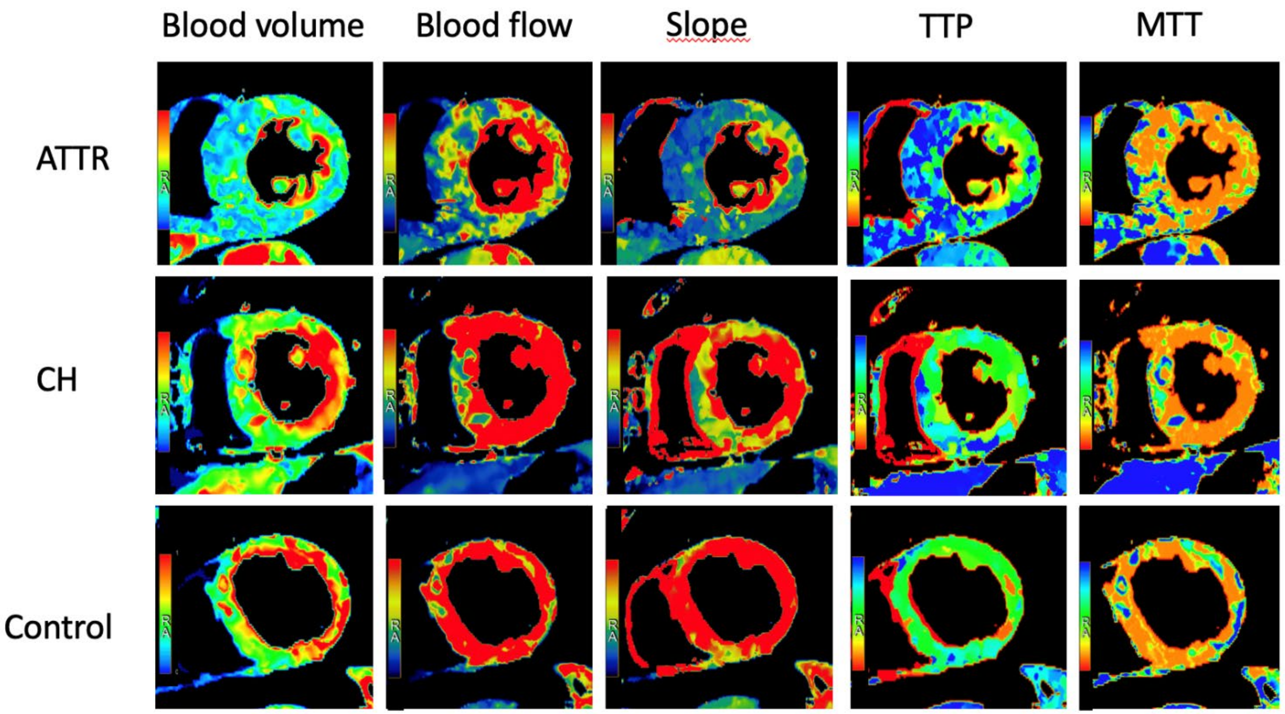
Wiser than « Yogi » Berra?

- « It's tough to make predictions, especially about the future. »
- « In theory, there is no difference between theory and practice. In practice, there is. »

Lawrence Peter Berra, 1935-2015

Being Geoffrey Hinton…

Classification and automatic indexing of medical images: a typical machine learning task…



3 classes :
- ATTR : amylose cardiaque
- CH : autres maladies
- Contrôle

8 grandeurs d'intérêt (gold standards) :
- fraction d'éjection du ventricule gauche
- masse du ventricule gauche
- 5 paramètres de perfusion du tissu cardiaque
- volume extra cellulaire

The road to wisdom…

Google

Josh Taylor *and* Alex Hern

Tue 2 May 2023 12.23 BST

'Godfather of AI' Geoffrey Hinton quits Google and warns over dangers of misinformation

The neural network pioneer says dangers of chatbots were 'quite scary' and warns they could be exploited by 'bad actors'

Dr Geoffrey Hinton, the 'godfather of AI', has left Google. Photograph: Linda Nylind/The Guardian

The man often touted as the godfather of AI has quit Google, citing concerns over the flood of misinformation, the possibility for AI to upend the job market, and the "existential risk" posed by the creation of a true digital intelligence.

Source: The Guardian, May 2, 2023

Citée par — TOUT AFFICHER

	Toutes	Depuis 2018
Citations	701778	478826
indice h	181	132
indice i10	439	333

Geoffrey Hinton

Emeritus Prof. Comp Sci, U.Toronto & Engineering Fellow, Google
Adresse e-mail validée de cs.toronto.edu - Page d'accueil

machine learning psychology artificial intelligence cognitive science computer science

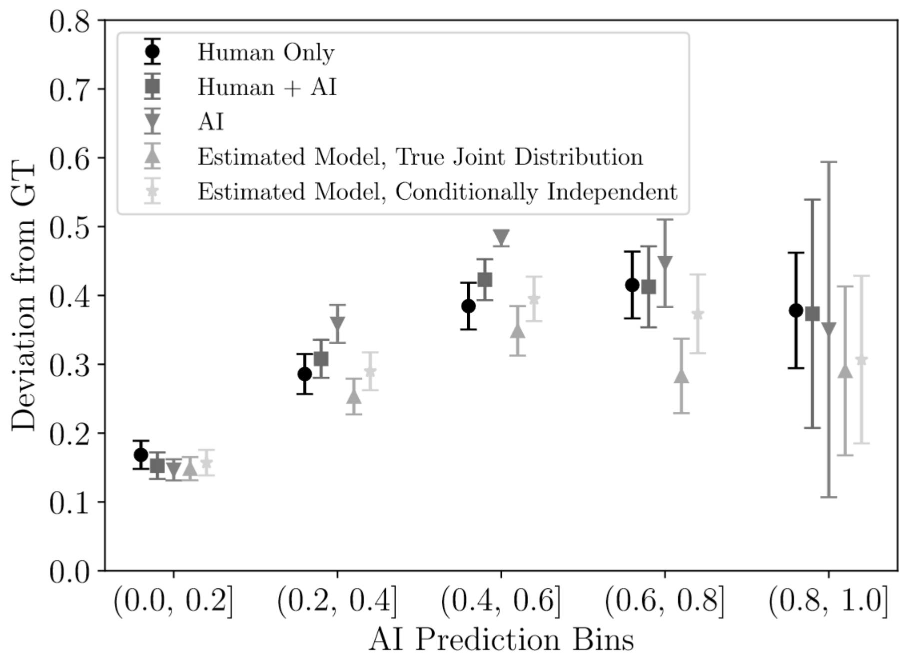
Source: Google Scholar, July 2023

The scientific response to the Toronto statement

« Using an experiment on professional radiologists that varies the availability of AI support and contextual information, it is shown that
(i) providing AI predictions does not uniformly increase diagnostic quality, and
(ii) providing contextual information does increase quality.

The results also show that, unless the mistakes the authors document can be corrected, **<u>the optimal solution involves delegating cases either to humans or to AI but rarely to a human assisted by AI</u>**. »

Figure 7: Model Deviation from Ground Truth



Note: This figure shows the performance of the different modalities that we consider for the optimal collaborative system. Cases are either decided by only the radiologist, only the AI, or the radiologist with access to the AI. These performance measures are constructed from our treatment effect analysis.

Source : Agarwal et al. (2023). Combining Human Expertise with Artificial Intelligence: Experimental Evidence from Radiology∗, Preprint, MIT.

AI and ethics
(following slides are courtesy of Theodoros Evgeniou)

Key Points about the challenges of AI adoption

1. **AI is unlike other technologies – in what ways?**

2. It is essential to consider human factors when developing AI

3. Many frameworks and tools to manage AI risks are under development – this is a very rich area for research and innovations

→ It is essential to align regulations with key characteristics of AI

What makes it risky…

1. It makes (increasingly complex) decisions (unlike other technologies)
2. No 100% accuracy: by nature it makes mistakes (for sure)
3. Large Scale: small errors can multiply to major risks or impact
4. Continuous Learning: what you have tomorrow is not what you have today!
5. Evolving environment: AI models typically have a lifetime
6. New vulnerabilities: adversarial attacks, lack of robustness, cybersecurity
7. Challenging accountability: complex multi-component systems, data complexities, multi-party, open source, etc.
8. Ethical issues: Fairness, Accountability, Transparency (the FAT Model)

…

AI risks

-Application-level

Performance risk

Security risk

Control risk

-Business and National-level

Enterprise risk

Economic risk

Societal risk

+Mankind level ?

Figure 6: Six Categories of AI Risks (Source: PwC Analysis)

Key Points about the challenges of AI adoption

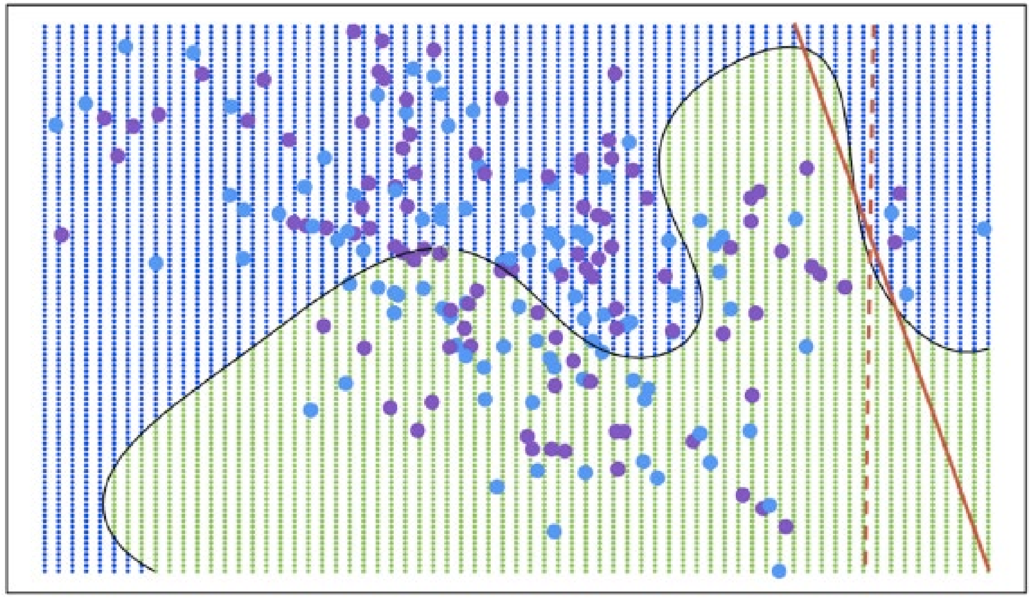
1. AI risks

2. **Human factors when developing AI**

3. Frameworks and tools to manage AI risks

Beware of Explainable AI – Human Factors

1. Lack of robustness of explanations may harm trust
2. Risk of overconfidence – and risk taking
3. Risk of narrative fallacy
4. … but not all users are the same



Human Factors example

WHY WAS MY LOAN APPLICATION DENIED? AN EMPIRICAL EVALUATION OF LAY PEOPLE'S PREFERENCES FOR EXPLANATIONS OF ALGORITHMIC DECISIONS

A PREPRINT

Yanou Ramon
Dept. of Engineering Management
University of Antwerp, Belgium

Tom Vermeire
Dept. of Engineering Management
University of Antwerp, Belgium

David Martens
Dept. of Engineering Management
University of Antwerp, Belgium

Theodoros Evgeniou
Dept. of Decision Sciences
INSEAD Europe Campus, France

Olivier Toubia
Dept. of Marketing
Columbia University, United States

February 1, 2021

INSEAD

Based on her browsing activity,
Emma was shown the following ad:

If you were Emma, which of the two explanations (shown below) that explain why you are seeing this advertisement, would you prefer?

(A)

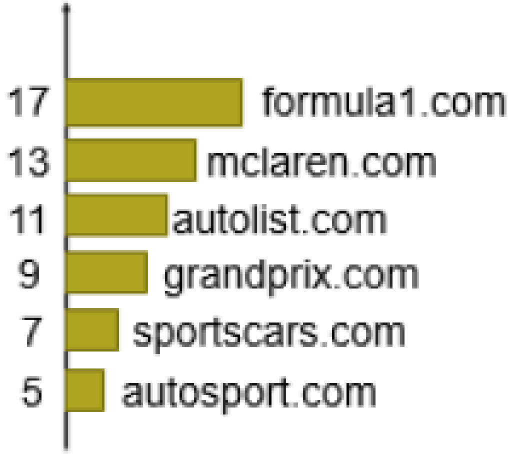
IF you had **not** visited **any** of the following web pages:

- redbull.com
- motor1.com
- fourwheeler.com

→ THEN
this ad would **not** be shown to you

(B)

The ad is based on the following topics of web pages you visited:



Based on her personal data,
Helena's travel loan application got rejected.

If you were Helena, which explanation would you want the bank to show you for explaining this decision?

(A)

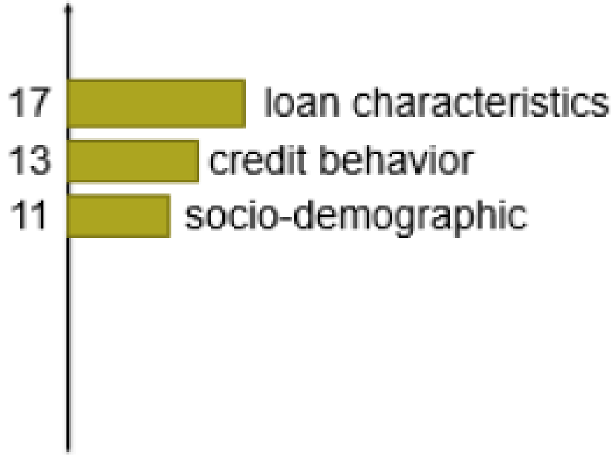
IF the data related to the following categories was **different**:

- loan characteristics
- socio-demographic
- credit behavior
- spending behavior
- macro-economic context
- family situation

→ THEN
the loan application would be accepted

(B)

the decision is based on the following personal data categories:



Explainable AI: Not a one (algo) size fits all

Sample Size	216 respondents recruited on the MTurk platform all based in the United States
Attributes	**Levels**
Format	Counterfactual explanation, One-sided importance-ranking, Two-sided importance-ranking
Complexity	Small (three features), Large (six features)
Specificity	Low-level, High-level
Socio-demographic variables	Age, Gender (=1 if female, =0 if male), Education (High school, Bachelor, Master)
Cognitive Reflection Test (CRT)	Score on test for analytical reasoning (CRT scores lie between 0 and 3)

Explainable AI: Not a one (algo) size fits all

INSEAD

Targeted Advertising

Attribute: level	**Mean part-worth utility**	**95% CI**
Format: Counterfactual	−0.44***	[−0.63, −0.24]
Format: Importance-ranking one-sided	0.77***	[0.54, 1.10]
Complexity: Small-sized	−0.32***	[−0.50, −0.16]
Specificity: Low-level	−0.16***	[−0.29, −0.04]

Credit Scoring

Attribute: level	**Mean part-worth utility**	**95% CI**
Format: Counterfactual	0.11	[−0.06, 0.29]
Format: Importance-ranking one-sided	−0.11	[−0.64, 0.28]
Complexity: Small-sized	−0.44***	[−0.58, −0.32]
Specificity: Low-level	0.38***	[0.19, 0.61]

Key Points about the challenges of AI adoption

1. AI risks

2. Human factors when developing AI

3. **Frameworks and tools to manage AI risks**

Emerging tools and processes

1. Software Toolkits
2. Documentation tools and frameworks
3. Auditing of AI Systems
4. Standards and Certification
5. Monitoring over the AI Lifecycle

INSEAD The Business School for the World®

Working Paper

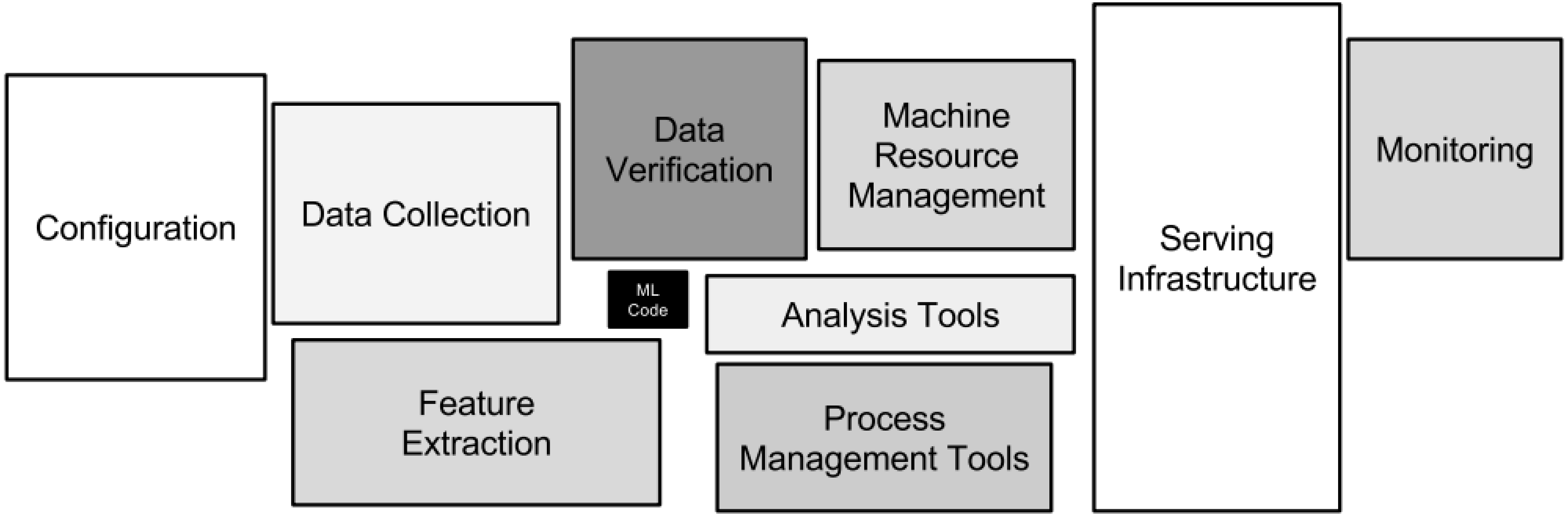
2021/04/DSC/TOM

Implementing AI Principles: Frameworks, Processes, and Tools

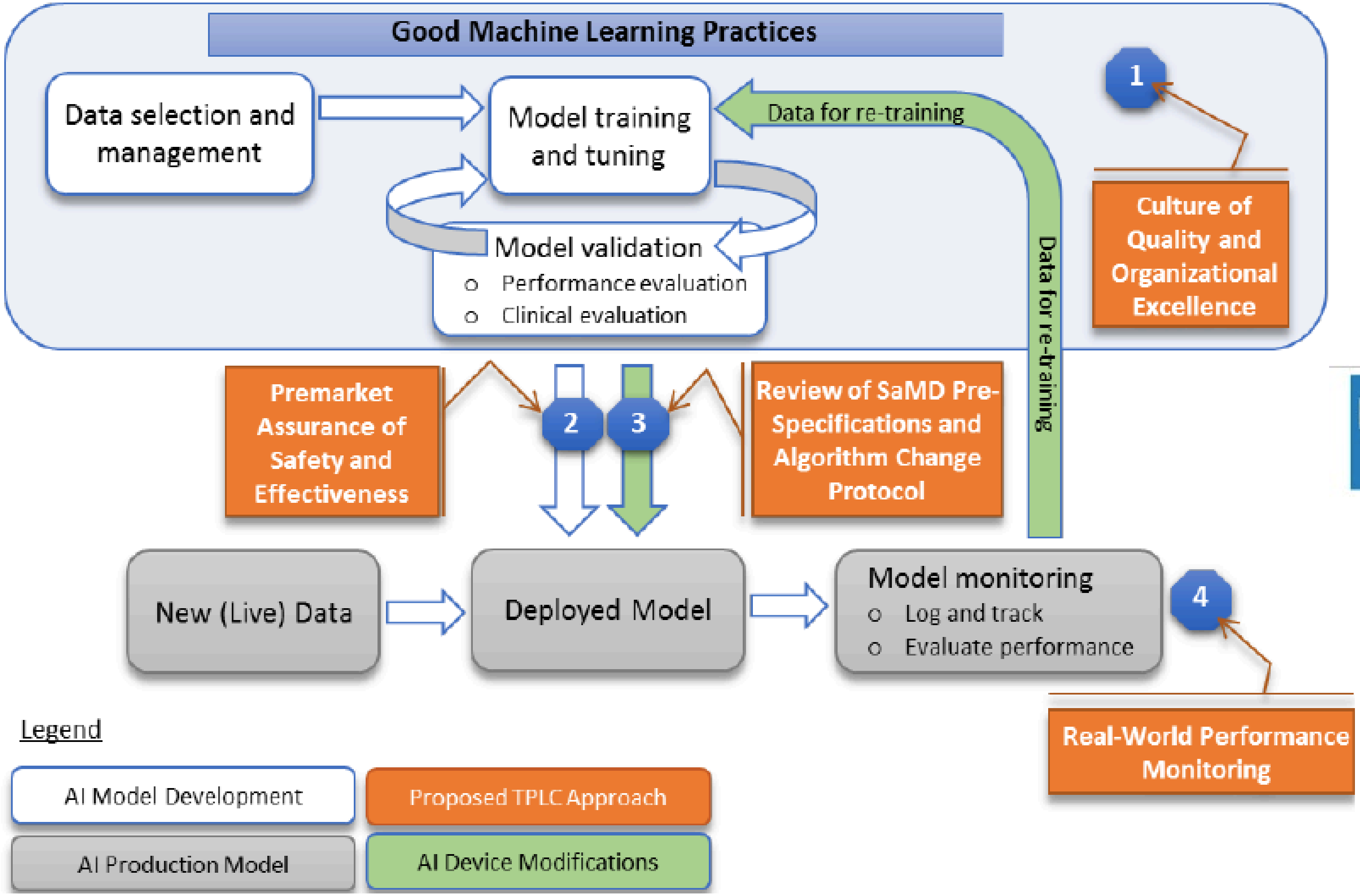
Pal Boza
INSEAD, pal.boza@insead.edu

Theodoros Evgeniou
INSEAD, theodoros.evgeniou@insead.edu

Components of an AI system



Good practices and AI lifecycle: the case of the FDA



FDA U.S. FOOD & DRUG ADMINISTRATION

Some example tools to manage AI risks

IBM Factsheets

Aequitas

IBM Fairness 360

Table 1: Key Software Toolkits and Frameworks for Implementing AI Principles

Toolkit	Developer
Fairness Tool[71]	Accenture
Foolbox[72]	Bethge Lab
CleverHans[73]	CleverHans Lab
Model Guardian[74]	Deloitte
Digital Impact Toolkit[75]	Digital Civil Society Lab, Stanford Center on Philanthropy and Civil Society
Deon[76]	Driven Data
Fairness Flow[77]	Facebook
What-If Tool[78]	Google
Ethics & Algorithms Toolkit[79]	GovEx, the City and County of San Francisco, Harvard DataSmart, and Data Community DC
AI Fairness 360[80,81]	IBM
AI Explainability 360[82]	IBM
Adversarial Robustness Toolbox[83] (ART)	IBM
LinkedIn Fairness Toolkit[84] (LiFT)	LinkedIn
Fairlearn[85]	Microsoft
InterpretML[86]	Microsoft
Harms Modelling[87]	Microsoft
Community Jury[88]	Microsoft
Skater[89]	Oracle
REVISE: REvealing VIsual biaSEs[90]	Princeton University
Responsible AI Toolkit[91]	PwC
audit-AI[92]	Pymetrics
FAT Forensics[93]	University of Bristol
Aequitas[94]	University of Chicago Center for Data Science and Public Policy
Lime[95]	University of Washington

Back to radiologists…


I'M BEING SUED FOR A MISSED DIAGNOSIS! WHAT DO I DO NOW?!
SORRY, BUDDY, COMES WITH THE JOB.
WHEN HANDING RADIOLOGY OVER TO ARTIFICIAL INTELLIGENCE SOUNDS APPEALING.
DiagnosticImaging.com • cartertoons.com