



## Acceptable Explanations through Machine Coaching


**Loizos Michael** (loizos@ouc.ac.cy)  
Associate Professor and Director of Computational Cognition Lab, OUC  
Research Pillar Leader of Artificial Intelligence & Communications, RISE



XAI Seminars @ Imperial (London, U.K.)      November 07, 2019  
© 2019 Loizos Michael

## A Brief History of Industrialization



**today** Industry 4.0


- Industry 1.0**
  - steam engine
  - physical labor
- Industry 2.0**
  - electricity
  - assembly line
- Industry 3.0**
  - electronics & IT
  - automation
- Industry 4.0**
  - communication
  - autonomy & AI

2      Acceptable Explanations through Machine Coaching      Loizos Michael (OUC & RISE)

## A Brief History of Machine Learning


**ARTIFICIAL INTELLIGENCE**

Any technique which enables computers to mimic human behavior




**MACHINE LEARNING**

AI techniques that give computers the ability to learn without being explicitly programmed to do so



**DEEP LEARNING**

A subset of ML which make the computation of multi-layer neural networks feasible




1950's   1960's   1970's   1980's   1990's   2000's   2010's

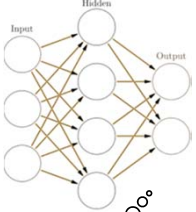
CRACLE

→ **Industry 3.0** → **Industry 4.0**

3      Acceptable Explanations through Machine Coaching      Loizos Michael (OUC & RISE)

## Supervised Learning in Machines

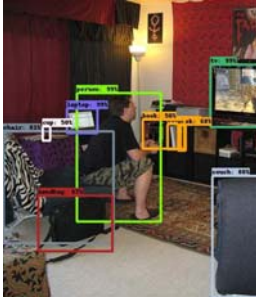
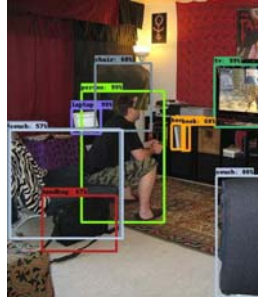




*Batch, data-heavy supervision with no explanation.*

4      Acceptable Explanations through Machine Coaching      Loizos Michael (OUC & RISE)

## Seeing the Elephant in the Room!

Rosenfeld et al., <https://arxiv.org/abs/1808.03305>, 2018

5      Acceptable Explanations through Machine Coaching      Loizos Michael (OUC & RISE)


## THIS IS YOUR MACHINE LEARNING SYSTEM?

YUP! YOU POUR THE DATA INTO THIS BIG PILE OF LINEAR ALGEBRA, THEN COLLECT THE ANSWERS ON THE OTHER SIDE.

WHAT IF THE ANSWERS ARE WRONG?

JUST STIR THE PILE UNTIL THEY START LOOKING RIGHT.

**alchemy (n.):**  
*a seemingly magical process of transformation, creation, or combination.*



Munroe, <https://xkcd.com/1838>, 2017

6      Acceptable Explanations through Machine Coaching      Loizos Michael (OUC & RISE)

### Biases and Lack of Explainability

- If humans can **discriminate**, so can machines!



Google image search for "unprofessional hair"

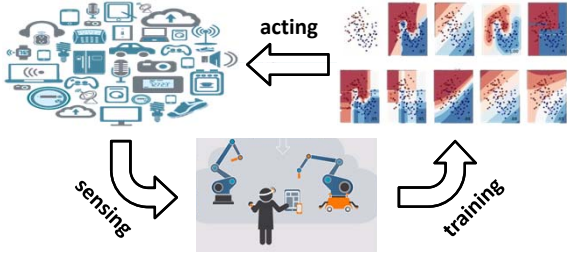
- GDPR offers a **right to AI explanations**.
- Most European countries have agreed to **modernize their national AI policies**.



cy. center for algorithmic transparency

7 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### The "Machine Learning 2.0" Era

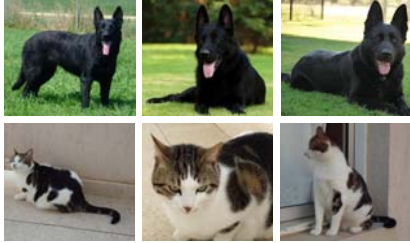


from ML 2.0 / blue-collar workers / assembly-line to ML 4.0 / white-collar workers / communication

8 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Issues with Post-Hoc Explainability

How would a machine trained on these images **explain** why it classifies an image as a dog / cat?

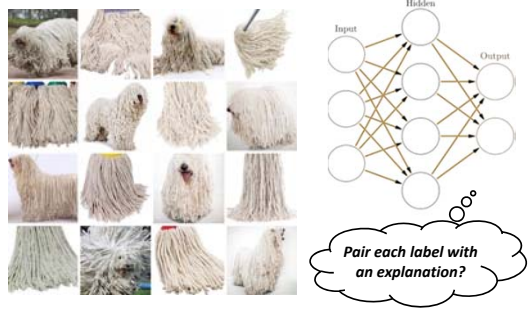


**explain dog by**  
 black fur: yes  
 tongue out: yes  
 on grass: yes  
 head right: yes

**explain cat by**  
 black fur: no  
 tongue out: no  
 on grass: no  
 head right: no

9 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Machine Learning of Explanations




Pair each label with an explanation?

10 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Bilateral Explainability in Humans

- Why are you making that move?
- It is the opening phase and I am trying to take control of the center.
- When one of your pieces is being attacked, make sure to protect it.



**Dialectical, one-shot supervision with counterargument.**

**Argument in support of decision, using syntax and concepts shared with coach.**

11 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### The Machine Coaching Paradigm

Can a cognitively fluent and transparent form of supervised policy learning carry over to HCI?

Machine coaching seeks run-time improvement of machines that are **explainable by design** by:

- insists that a party **externalizes its reasoning** in a way that is **understandable** to the other.
- offers formal **guarantees on the efficacy and on the efficiency** of the process of learning.

Michael, "Machine Coaching", Proc. XAI Workshop @ IJCAI, 2019

12 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Example of a Coaching Interaction

Computational Cognition Lab

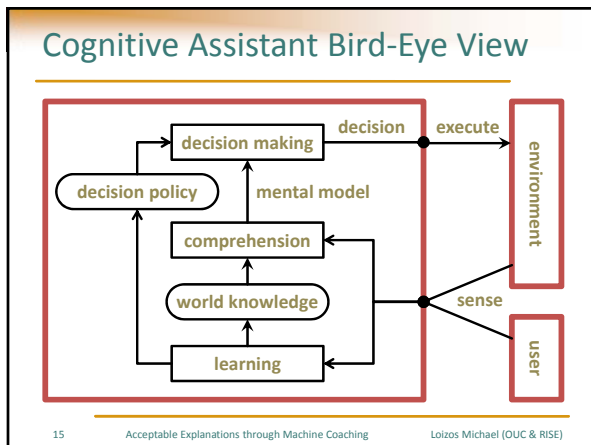
13 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### The Ancient Greeks Said it First!

*“For every belief comes either through **sylogism** or from **induction**.”*

Aristotle, Organon (Prior Analytics II, §23)

14 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)



### Externalizing Reasoning through Argumentation

16 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Cognition and Argumentation

**Mercier & Sperber, “Why do humans reason? Arguments for an argumentation theory”, 2011.**

- Construct arguments for accepting or declining a “conclusion that was raised”.
  - **awareness** not only of the conclusion, but also of the arguments that support the conclusion.
- **Improve argument** if motivated / challenged.
  - *“biased and lazy when they produce arguments”* in solitary setting, but *“objective and demanding when they evaluate others’ arguments”* in dialectical setting [Mercier, 2016].

17 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Games and Argumentation

18 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Stories and Argumentation

What inferences follow from a story?

*Papa Joe woke up early at dawn, and went off to the forest. He walked for hours, until the sight of a turkey in the distance made him stop.*

*A bird on a tree nearby was cheerfully chirping away, building its nest. He carefully aimed at the turkey, and pulled the trigger of his shotgun.*

*Undisturbed, the bird nearby continued chirping.*

Q: What is the condition of the turkey?

(a) *Alive and unharmed.*    (b) *Dead.*    (c) *Injured.*

19    Acceptable Explanations through Machine Coaching    Loizos Michael (OUC & RISE)

<http://cognition.ouc.ac.cy/star>

20    Acceptable Explanations through Machine Coaching    Loizos Michael (OUC & RISE)

### Arguments from Prioritized Rules

This animal has Feathers, lives in Antarctica, and looks Funny. Question: Does it have Wings?

21    Acceptable Explanations through Machine Coaching    Loizos Michael (OUC & RISE)

### Just a Little Bit of Formalism...

**Argument for L:** minimal subset of premises and rules = L.  
**Exogenous attack:** no rules; L in conflict with interm. infer.  
**Endogenous attack:** last rule has priority over interm. infer.

ASPIC+ semantics [Prakken 2010] with axiomatic premises, defeasible rules, rebutting attacks (on intermediate inferences), application of rule preferences on the last link.

**Induced AF** from premises and rules. **Grounded semantics.**

- Single intended model [Stenning and Lambalgen, 2012].
- Efficient computation. **Theorem:** *In number of rules!*
- Concise dual representation [Craven and Toni, 2016], but can cope with directed cycles, supports all arguments.

22    Acceptable Explanations through Machine Coaching    Loizos Michael (OUC & RISE)

### Learning Guarantees through PAC Semantics

23    Acceptable Explanations through Machine Coaching    Loizos Michael (OUC & RISE)

### Learning from Labeled Examples

- Find the hidden concept (Bongard problems):

24    Acceptable Explanations through Machine Coaching    Loizos Michael (OUC & RISE)

### Probably Approximately Correct

- Statistical guarantees on the learning quality:
  - improbable* to get unrepresentative instances
  - predictions need only be *approximately* correct

25 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Never-Ending Rule Discovery

Initial evidence *happens to* activate R7. Later counterexamples deactivate it. In the meantime, evidence activates R1. Thus, support becomes *stronger* for R7. Even though *counterexamples remain*. Rules will react to environment change.

26 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Machine Coaching vs Machine Learning

27 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Coaching for Learning Policies

During development, the assistant is initialized with the following user-independent knowledge:

- r1* : if day is from Monday to Friday **then not** day-off
- r2* : if time is from 9am to 5pm **and not** day-off **then** at work
- r3* : if time is from 12am to 6am **then not** may interrupt
- r4* : if at work and giving a talk **then not** may interrupt
- r5* : if at work **then** set ringing volume to a low audible level
- r6* : if **not** may interrupt **and** call **then** disable ringing

*How truly user-independent is each rule above?*

28 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Coaching for Learning Policies

A user perceives and reacts to actions/inactions:

If the user repeatedly accepts incoming calls from number S (the user's spouse!), even when ringing is disabled...

*r6* : if **not** may interrupt **and** call **then** disable ringing **and not** number S

→ non-modular policies; multiple user reactions

If (during an overseas trip by the user) calls from number S are often received (only) between 12am and 6am...

*r3* : if time is from 12am to 6am **then not** may interrupt

29 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Coaching for Learning Policies

Machine coaching allows a user to have a more direct involvement in the *unambiguous revision* of the assistant's policy **from the first reaction!**

**User:** Why did you disable ringing for a call this morning?

**Assistant:** Because today is Monday, the call was received at 10:30am, and you were giving a talk, and I concluded, by applying the rules *r1*, *r2*, *r4*, that I may not interrupt you.

**User:** You may interrupt me when my spouse calls!

*r7* : if call **and** number S **then** may interrupt (> *r3*, *r4*)

30 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)



### Learning Desiderata for Coaching

a) Quantify guarantees. b) Integrate ML'ed rules.

- bilateral communication**, online learning (i.e., get observation, make prediction, get advice)
- arbitrary advice** to wrong (or unconvincing) predictions, without naming a right prediction
- learning goal** not to identify “correct” advice given observation and prediction, but rather to **conform to advice** (i.e., given observation, identify prediction that leads to no advice).

31 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Just a Little Bit of Formalism...

**Definition:** An algorithm is a (**probably approximately conformant learner**) if for every real  $0 < \delta, \epsilon \leq 1$ , every probability distribution  $D$  over inputs of size  $n$ , and every feedback function  $f \in F$  of size  $s$ , and if allowed to repeatedly **draw** an input  $x$  (**learning instance**) from  $D$ , **select** an output  $y$  (**prediction**), and **receive**  $f(x,y)$  (**piece of advice**) for time at most a polynomial  $g(1/\delta, 1/\epsilon, n, s)$  the algorithm terminates and returns, **except with probability at most  $\delta$** , a hypothesis  $h: X \rightarrow Y$  that is **(1- $\epsilon$ )-approximately conformant** under  $D$  against  $f$ .  
**( $h$  conforms with  $f$  on input  $x$  if:  $f(x,y) = \emptyset$  for  $h(x) = y$ )**

32 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Stratified Simply-Shaped Subspaces

33 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Cognitively-Light & Useful Advice?

“a human is instructed mainly in **declarative sentences describing the situation in which action is required**” — McCarthy, 1959.

“[humans are] **biased and lazy** when they produce arguments [in a solitary setting, but] **objective and demanding** when they evaluate others’ arguments [in a dialectical setting]” — Mercier, 2016.

**Theorem:** There exists a conformant learner if the **coach advises the machine by identifying:**

- Omitted or superfluous rules in explanations.
- Counter-arguments to “weak” explanations.

**Proof:** Add / drop rules. **Elaboration tolerant!**

34 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### Coaching $\approx$ Learning + Programming

- Explicates HCI typically **at the fringes** of L&P.


	(superv.) learning	programming	coaching
<b>H</b>	<i>labels inputs</i> according to <b>target</b> theory	<i>generates</i> explicit parts of <b>target</b> theory	<i>recognizes</i> mistakes in <b>hypothesis</b> theory
<b>C</b>	<i>generalizes</i> to create hypothesis theory	<i>blindly adds</i> parts to hypothesis theory	appropriately <i>revises</i> hypothesis theory
<b>I</b>	<i>one-sided</i> , online/batch, mostly <b>machine</b> burden	<i>one-sided</i> , at start, mostly <b>human</b> burden	<i>dialectical</i> , online, <b>less &amp; shared</b> burden

- To gather knowledge for “safe” repetitive tasks with user-specific **verbalizable explanations**.
- To **debug / personalize** knowledge from L&P.

35 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)

### In Lieu of a Conclusion...

36 Acceptable Explanations through Machine Coaching Loizos Michael (OUC & RISE)



funded by EU's H2020 programme

### machine-mediated diversity-aware human interactions

- Pilot study in 18 university and adult school sites, with 10k participants (+ non-EU).
- Clear ethical guidance for the pilot activities and the technology development.

37      Acceptable Explanations through Machine Coaching      Loizos Michael (OUC & RISE)

## Studies in Cognitive Systems

**BLENDED DISTANCE LEARNING METHODOLOGY**  
 courses taught online with live tutoring sessions  
 in-class exams in student's country of residence  
 summer tutorial camps in Cyprus (optional)

Courses Offered	Cognitive Psychology		Computer Science	
	Theme	Cognitive Psychology	Theme	Computer Science
<b>Foundations</b>	CP.F1 Introduction to Cognitive Psychology	CS.F1 Introduction to Artificial Intelligence	CS.F2 Computational Intelligent Systems	
<b>Perception</b>	CP.P1 Human Perception and Attention	CS.P1 Natural Language Processing		
<b>Learning</b>	CP.L1 Learning and Memory in Humans	CS.L1 Computational Learning Theory		
<b>Reasoning</b>	CP.R1 Mental Representations and Reasoning	CS.R1 Cognitive Agents		
	CP.R2 Cognitive Modelling	CS.R2 Adaptive and Interactive Systems		
<b>Systems</b>	CP.S1 Experimental Psychology	CS.S1 Cognitive System Design		
	CP.S2 Cognitive Neuroscience	CS.S2 IBM's Watson Machine		

**Admission Requirements**  
 Geared towards students with a first degree in the STEM fields (Science, Technology, Engineering, Mathematics), Cognitive Science, or Psychology. Basic knowledge assumed in mathematics (discrete mathematics, formal logic, statistics, calculus), and computing (algorithms, basic programming).

For additional information, or expression of interest contact:  
 cogsys@ouc.ac.cy  
 cogsys@ucy.ac.cy

OPEN UNIVERSITY OF CYPRUS      University of Cyprus      <http://cogsys.ouc.ac.cy>

Human-Computer Symbiosis  
 Cognitive Systems  
 Join the Era of Cognitive Systems!  
 Postgraduate in September 2016