Machine Coaching

Loizos Michael

Open University of Cyprus &
Research Center on Interactive Media,
Smart Systems, and Emerging Technologies
loizos@ouc.ac.cy

Abstract

This position paper puts forward *machine coaching* as a form of interactive machine learning that emphasizes the requirement for *humans and machines to externalize their internal reasoning process* in a manner that is understandable, at least at a basic level, by the other party. We posit that this mutual understanding leads to a computationally and cognitively lighter interaction, supports the run-time personalization of machines even by non-technically-savvy humans, makes any machine biases explicit and the process of their acquisition transparent, and facilitates the development of AI systems that can, by design, explain and be explained to. Backed by psychological theories of human reasoning and recent technical work, this paper adopts the working hypothesis that argumentation over symbolic rule-based knowledge offers a reasonable common language and semantics that machines and humans can utilize when interacting through machine coaching.

1 Disrupting Knowledge Work Automation

Identified by the McKinsey Global Institute as one of twelve major disruptive technologies [Manyika *et al.*, 2013], automation of knowledge work has the second largest potential for economic impact by the year 2025, reaching $5–7 trillion in value, and affecting 27% of the global employment costs and 9% of the global workforce, while benefiting a broad spectrum of professions: common business functions (e.g., call center sales, administrative support, and customer service), social sector services (e.g., education, health care), technical professions (e.g., software design, drug discovery), management, and professional services (e.g., law, financial).

The game-changing potential of mechanizing knowledge acquisition and maintenance aligns with key recommendations in the 2010 and 2013 reports of the U.S.A. President's Council of Advisors on Science and Technology, who identified the need for a bottom-up data-driven approach to knowledge given the increasing pervasiveness of data on the Web and other sources [PCAST, 2010; 2013]. Based on these reports, "Automated analysis techniques such as data mining and machine learning facilitate the transformation of data into knowledge, and of knowledge into action.", and "This 'computational knowledge extraction' lies at the heart of 21st century discovery. [...] Advancing these tools, so that science may advance, is a major challenge for Networking and Information Technology Research and Development. The world of science is transitioning from data-poor to data-rich, vastly expanding the potential for new breakthroughs [...]".

Despite the striking advances of data-driven AI, the disruption forecasted by the McKinsey Global Institute has yet to materialize. Instead of replacing knowledge workers with automated systems, modern Machine Learning has effectively revamped the role of humans into a more menial, albeit a still necessary, one: that of the data annotator. This change in the role of humans is reminiscent of the evolution of their role between the first and second industrial revolutions — from skilled problem solvers who operated machines to "blue-collar" assembly-line workers that undertook menial tasks that could not be feasibly or cost-effectively automated by machines — and is a far cry from the role that humans have in the fourth industrial revolution as "white-collar" directors of machine operations; a role that is presumably much more aligned with the disruption envisioned by the McKinsey Global Institute.

Dislodging the current role of humans towards becoming "white-collar" workers is, we believe, the primary means by which disruption in knowledge work automation will be facilitated. Much in the same way that the fourth industrial revolution is characterized by enhanced system-to-system communication, analogously the new role of humans should be, we posit, one that enhances human-to-machine communication, going beyond the task of data annotation or learning supervision, and lifting their role to that of machine coaches.

This position paper proposes the development of a machine coaching paradigm that *(i)* retains those characteristics of machine learning (e.g., induction, generalization, statistical guarantees) that have been found to be useful in dealing with the analysis of data, and that *(ii)* promotes a form of human-machine interaction that facilitates human efforts in endowing machines with the ability to explain and be explained to.

In addition to discussing how related work from formal argumentation and interactive machine learning can be used as a basis for this new paradigm, we also take a first step towards formalizing the desiderata on the interaction that takes place during a machine coaching session, and we propose a particular protocol that the human can use to contribute knowledge to the machine for which we can establish certain guarantees.

2 Machine Coaching via Argumentation

Putting aside their development-time role in the design of algorithms and the identification of relevant features in the as-

sociated input and output spaces, the role of humans during the actual process of knowledge acquisition and maintenance is, for the most part, that of data annotators or learning supervisors, tagging images or objects of interest with the intended — whatever this is defined to be — label or inference.

In the typical case, these tags are not accompanied by any form of explanation — neither on how that label or inference was reached by the human, nor on why it might be appropriate (or more appropriate than some other tag) — and hence reveal nothing in terms of the internal reasoning of the human. Relatedly, the human is rarely aware of how the tags are to be consumed, and reasoned with, by the machine, and does not in general, know how the machine operates internally. Even in attempts to make machine learning more accessible to humans (e.g., through machine teaching [Simard *et al.*, 2017]), machines are still (or purposefully) treated as black-boxes.

It is instructive to consider a form of interaction where humans explicate their reasoning process, and explain to the machine how some particular tag was derived, or why a certain alternative tag (perhaps one proposed by the machine) is inappropriate. For explanations to be meaningful, humans need to have a basic understanding of the machine's learning and reasoning processes, so that they can adapt the information that they offer based on how it will be used by the machine.

Such human understanding of the operation of existing systems is not very pronounced, but it is not completely absent either. When interacting with a web search engine, for example, most humans have a basic understanding of how results are returned for a query, and can adapt their query to get more desirable results. Machine coaching seeks to elevate such primitive forms of dialogue and mutual understanding between humans and machines, by offering a fundamentally more suitable language of communication that *makes explicit the internal reasoning of the machine to the human, and vice versa*.

Envisioning the resulting interaction between humans and machines as a form of symbiosis, where a human coach "nurtures" a machine cognitive assistant towards adopting beliefs and associated explanations akin to those of the human, our proposal for developing a machine coaching paradigm is inspired from, and adopts some of, the key findings of existing psychological theories of learning and reasoning. In the present position paper we propose the working hypothesis that *argumentation offers a promising vehicle to establish mutual understanding of reasoning between humans and machines*.

Psychological evidence suggests that argumentation is integral in the human reasoning process [Mercier and Sperber, 2011], and that humans can consciously represent knowledge activated during their argumentative reasoning in the form of rules [Diakidoy *et al.*, 2017]. In addition, the dialectical context in which argumentation will be used for machine coaching is in line with further psychological evidence suggesting that even though humans are "biased and lazy when they produce arguments" *in a solitary setting*, they are "objective and demanding when they evaluate others' arguments" *in a dialectical setting* [Mercier, 2016]. At the same time, argumentation has also been extensively studied within Artificial Intelligence [Dung, 1995], and has been proposed as the basis for designing cognitive assistants [Kakas and Michael, 2016].

We acknowledge, of course, that the specific ways in which humans and machines internally represent and reason with arguments will differ. For one, machines will presumably adopt a formal representation of arguments, as appropriate to mechanize the reasoning process, whereas humans might be more comfortable using some specified subset of natural language.

To facilitate the interaction and the translation between the internal representations of the interlocutors, the use of some form of a controlled natural language [Kuhn, 2014] seems appropriate. The scenario below shows what we envision to be a typical interaction between a human and a machine acting as their cognitive assistant for handling incoming calls, highlighting the use of machine coaching over machine learning.

The assistant can perceive information such as: the user's location and movement through the phone's sensors, calendar appointments and contacts, current date and time, etc. The assistant can execute actions such as: send SMS messages, decline incoming phone calls, set notifications, etc. During development, and independently of the user's profile, the assistant is initialized with the following knowledge:

r_1 : ***if*** *day is from Monday to Friday* ***then not*** *day-off*
r_2 : ***if*** *time is from 9am to 5pm* ***and not*** *day-off* ***then*** *at work*
r_3 : ***if*** *time is from 12am to 6am* ***then not*** *may interrupt*
r_4 : ***if*** *at work* ***and*** *in a meeting* ***then not*** *may interrupt*
r_5 : ***if*** *at work* ***then*** *set ringing volume to a low audible level*
r_6 : ***if not*** *may interrupt* ***and*** *call* ***then*** *disable ringing*

During run-time, the user perceives and reacts to the assistant's actions / inactions in various contexts. If, for instance, the user repeatedly accepts incoming calls from number S (corresponding, say, to the user's spouse!), even when ringing has been disabled by the assistant, then this may eventually lead to the following revision of a rule: "r_6 : ***if not*** *may interrupt* ***and*** *call* ***and not*** *number S* ***then*** *disable ringing".*

Besides giving rise to non-modular policies that quickly become unwieldy and unreadable to the user, the above process of learning requires multiple reactions from the user before identifying which of the numerous aspects of a context are the ones that are predictive of the user's behavior. Even when learning does identify some predictive aspect of a context, there is a distinct risk that the aspect will be a non-commonsensical or even an undesirable one. If, for instance, it so happens (during an overseas trip of the user) that calls from number S are often received between 12am and 6am, and those are the only such calls, then the assistant would be equally justified to drop rule r_3 *instead of revising rule* r_6 *as above.*

Through machine coaching, however, the user has a more direct involvement in the unambiguous revision of the assistant's policy, even from the very first time that the assistant's action deviates from the user's expectation. This may happen, for instance, if the user receives a call during a work meeting, triggering the following dialogue (right after the event, at the request of the user, or at a pre-scheduled coaching session):

User: Why did you disable ringing for a call today?

The assistant searches for calls that match the context and action specified by the user, identifies a relevant call from number S, retrieves the supporting argument, and responds:

Assistant: Because today is Tuesday, the call was received at 11:30am, and you were in a meeting, and I concluded, by applying the rules r_1, r_2, r_4, *that I may not interrupt you.*

User: You may interrupt me when my spouse calls!

Resolving the coreference between "my spouse calls" and "call from number S", or identifying in the user's contact list the number that belongs to the user's spouse, the assistant

constructs and integrates the following rule in its knowledge base, with priority over existing conflicting rules r_3 and r_4:

r_7 : ***if*** *call* ***and*** *number S* ***then*** *may interrupt*

Analogously, if the user receives a call during a weekend work trip abroad, the following dialogue could occur after the call (under the angry stares of colleagues for the ringing):

User: Why was the ringing volume not at a lower level for the call that has just concluded?

The assistant searches for an argument supporting its decision not to lower the ringing volume, and responds:

Assistant: Today is Saturday, your GPS location shows that you are abroad, and your calendar shows that you are on a work trip, but I was unable to conclude that you are at work.

User: If I am abroad on a work trip, then I am at work.

The assistant constructs and integrates the following rule in its knowledge base (which introduces no conflicts):

r_8 : ***if*** *abroad* ***and*** *on work trip* ***then*** *at work*

Rules constructed and integrated based on user advice are by no means assumed to be infallible and universally applicable, and might thereafter appear in arguments that lead the assistant to future wrong actions / inactions. Overall, however, their integration is expected to help the assistant's policy quickly converge towards the user's private expectations.

Using its knowledge, an assistant draws inferences on what actions to take in a given context. Each set of rules that supports an inference corresponds to an *argument* that the assistant could use to explain the taking of some action. Different sets of rules lead to different inferences and arguments in their support, possibly *attacking* other arguments by contradicting their (intermediate) inferences. The assistant seeks to identify a collection of arguments that defends all its attacks: any counter-argument that attacks an argument in the collection, is attacked back by an argument in the collection. Such a collection allows the assistant not only to justify the taking of an action, but also to defend the dismissal of an alternative one.

During any single round of machine coaching, the process by which either a human or a machine offers an explanation to their interlocutor can be approached as a form of explanation-based generalization (EBG) [Mitchell *et al.*, 1986]: Interpreting the contents of one's knowledge base as a *domain theory*, the context as a *training example*, and the actions supported by one's knowledge in that context as a *goal concept*, one seeks to generalize the training example by regressing from the goal concept to a sufficient condition that entails the goal concept, while satisfying an *operationality criterion*. This criterion ensures that the resulting explanation will be in a form that is operationally usable by the receiving party, while the generalization condition ensures that the communicated knowledge will not be overly-specific to the particular context.

3 Machine Coaching as Interactive ML

The technology of knowledge work automation has received the second least media attention across the twelve disruptive technologies that were identified in the McKinsey Global Institute report, and has been the least publicized one given its potential economic impact [Manyika *et al.*, 2013]. The report attributes this lack of media attention to significant organizational, cultural, and legal hurdles for the adoption of the technology, including risk-aversion by firms to adopt a technology until its benefits have been clearly proven, and resistance due to lack of trust on its performance on new and previously unforeseen circumstances, where mistakes may have significant ramifications, including endangering human lives.

The machine coaching paradigm speaks directly to the issues above for the adoption of the disruptive technology. Its emphasis on building machines that are able to explain and be explained to facilitates humans to gradually build trust towards a machine's performance through a dialectical interaction, analogously to how they build trust towards new colleagues by getting to know their work and skills. Furthermore, and especially for critical tasks for which humans may wish to retain decision-making (e.g., in the medical domain), the view of a machine as an assistant is appropriate for advising a human, explaining why it reached a conclusion, and allowing the human to evaluate the validity of the argument and decide whether to adopt it *on a per case basis*. This is typically not an option for machines built through machine learning, where trust reduces to a statistic of the machine's entire past performance, and is not a function of each individual conclusion.

The above notwithstanding, machine coaching remains a form of interactive machine learning, since a machine seeks to learn to produce convincing arguments in support of its decisions. Consequently, one needs to establish that the learning goal can be satisfied reliably and efficiently, through a dialogue with a human analogous to the one we exemplified for the call-handling cognitive assistant. Based on our initial technical investigation, we suggest that the Probably Approximately Correct (PAC) learning semantics [Valiant, 1984] could offer a solid basis for the formalization of machine coaching.

Departing from the majority of work in PAC learning, which focuses on establishing *objective* guarantees on the end-to-end *correctness* (against an external specification) of a machine's decisions, a semantics for machine coaching must focus on establishing *subjective* guarantees on the *persuasiveness / convincingness* (for a certain user) of the explanations offered for those decisions. It is possible that a position (e.g., "turn your phone off during the night") might be prima facie acceptable, but certain users might not find a given explanation (e.g., "because it will help reduce your carbon footprint") convincing. It is, also, possible for certain users not to immediately accept a position (e.g., "do not reply to your spouse's latest SMS message") until convinced by a given explanation (e.g., "because your phone has been infected by a computer virus that steals your data through fake SMS messages").

A learning-theoretic semantics for machine coaching will serve to fulfil a second desideratum beyond that of establishing formal guarantees on the appropriateness of the machine's decisions: that of allowing knowledge that has been acquired through machine coaching to be seamlessly integrated with knowledge that has been acquired through autonomous machine learning. The necessity for such an integration follows if one observes that although machine coaching will end up producing highly user-specific knowledge, some (perhaps, even, a considerable) fraction of the produced knowledge will be commonsensical, in that it will be common among multiple potential users of an intelligent machine. It is preferable for common knowledge to be acquired by a means other than machine coaching, lessening the burden on the users themselves.

Acknowledging the two different sources of knowledge is, in fact, in line with our recent proposal on the dual role that knowledge has to play in a cognitive assistant's architecture

[Kakas and Michael, 2016]: both as a mechanism to comprehend a given situation / context, and as a mechanism to apply a policy based on that comprehension. Commonsense world knowledge is primarily geared towards the role of comprehension, whereas machine-coached user-specific knowledge is geared towards the role of specifying the policy to be applied.

This distinction is well-illustrated in our call-handling cognitive assistant example, where rule r_1 is meant to capture a "definition" of a workday shared by several humans, used during the comprehension and abstraction of the current context to derive higher-level concepts and inferences from the lower-level concepts and sensory inputs. By contrast, rule r_7 is part of a user-specific policy for answering the phone, and it leads, directly or indirectly, to the execution of an action.

To face head-on the challenge of "extracting worldly knowledge" [PCAST, 2010] — rather than facts, which is the target of ongoing high-impact work [Carlson *et al.*, 2010] — one can utilize raw text as a source of training data. Work on autodidactic learning [Michael, 2008; 2009; 2013a] shows that supervised learning can be used without the active involvement of humans, by exploiting whatever "supervision" might already be present in text. Extending autodidactic learning to adopt techniques from argumentation mining [Lippi and Torroni, 2016], and subsequently integrating it with machine coaching might lead to the realization of a uniform framework for the disruptive technology of knowledge work automation.

Although the use of machine learning partly addresses certain subjectivity (and brittleness) concerns that would arise had knowledge been gathered directly from knowledge engineers or crowdworkers, machine-learned knowledge is still prone to inheriting human biases present in the training data; see, e.g., [Michael, 2013a] for a discussion on "websense versus commonsense knowledge". Rule r_1, for example, could be, plausibly, induced by a machine learning algorithm that extracts knowledge from Web text, yet it is not applicable in certain Middle Eastern countries where Sunday is a workday.

Despite concerns on the "universality" of machine-learned knowledge, we hypothesize that the bulk of this knowledge will be usable, while those parts of the knowledge that might end up supporting undesirable inferences can be subsequently debugged by users through machine coaching. Such debugging will, obviously, lead to some replication of cognitive work, and it will not fully eliminate bias, but will, instead, replace it with bias that aligns better with each user's own preferences and beliefs. We expect that an appropriately-designed study could show that despite these drawbacks, the proposed two-step process of first acquiring world knowledge through machine learning and then debugging it through machine coaching is preferable to either of the one-step alternatives.

4 Towards a Theory of Machine Coaching

In an effort to show that some key ideas surrounding machine coaching can be made sufficiently concrete for mechanization and analysis, we present below an attempt towards a formal theory of machine coaching, adapting and extending technical ideas from our earlier work [Michael, 2017] on how machine coaching can be seen as a vehicle for realizing McCarthy's [1959] vision of building an advice-taking machine.

We, certainly, do not claim here that our proposed formalization is in any way unique or the only route forward. Nor do we claim that the simple propositional language, which we have chosen for readability purposes, suffices to capture the more complex representation needed when reasoning, for instance, in temporal settings with causal knowledge [Michael, 2013b]. Our sole aim here is to demonstrate a potential path towards the formalizability of the interaction of argumentation and learning within the paradigm of machine coaching.

We start with a basic syntax for representing knowledge. A ***literal*** is either an atom or its negation. Two literals are ***conflicting*** if one is an atom and the other is the atom's negation; the unique conflicting literal of a literal λ is denoted by $\overline{\lambda}$. A ***context*** x is a collection of pairwise non-conflicting literals; the set of all contexts is X. A ***rule*** is an expression of the form $\varphi \rightsquigarrow \tau$, where the ***body*** φ of the rule is a finite conjunction of literals, and the ***head*** τ of the rule is a literal; the set of all rules is R. Two rules are ***conflicting*** if their head literals are conflicting. A ***knowledge base*** $\kappa = \langle \varrho, \succ \rangle$ comprises a finite collection ϱ of rules, and an irreflexive antisymmetric ***priority relation*** $\succ$ over pairs of conflicting rules in $\varrho \times \varrho$.

To define an argumentation framework [Dung, 1995], we need to specify how arguments and attacks are induced from a knowledge base. For concreteness, we adopt the following choices under the ASPIC+ framework [Prakken, 2010]: axiomatic premises (i.e., the context x is indisputable), defeasible rules (i.e., all rules can be overridden), rebutting attacks (i.e., an argument is attacked by questioning one of its intermediate inferences), and application of rule preferences on the last link (i.e., the attacking argument's strength is determined by the strength of its last rule). We formalize these choices next.

To define arguments induced by a set of rules ϱ in a context x, we start by interpreting literals in x as facts, and rules in ϱ as classical implications, and we draw inferences through the repeated application of modus ponens. A minimal subset of x and a minimal subset of ϱ that lead, thus, to the inference of a literal τ is an ***argument*** for τ in x under ϱ (or under $\kappa = \langle \varrho, \succ \rangle$). In an argument for τ, a rule with head τ, whenever such one exists, is unique and it is the argument's ***crown*** rule.

Given two arguments in a context x under a knowledge base $\kappa = \langle \varrho, \succ \rangle$, the first argument for literal τ ***attacks*** the second argument on the latter's rule r_2 with head $\overline{\tau}$ if: either $\tau \in x$ (***exogenous*** / external attack), or the former's crown rule r_1 is such that $r_2 \not\succ r_1$ (***endogenous*** / internal attack). Compared to assumption-based argumentation [Dung *et al.*, 2009], which also chains rules from premises to inferences, our premises are not weak assumptions that are attacked by other arguments; instead, arguments are attacked on their rules, while premises themselves are only a *source* of attacks.

Overall, a knowledge base κ induces a ***contextualized argumentation system*** $\langle \mathtt{A}_\kappa, \mathtt{R}_\kappa \rangle$, where $\mathtt{A}_\kappa$ is a mapping from each context $x \in X$ to the set $\mathtt{A}_\kappa(x)$ of all arguments in x under κ, and $\mathtt{R}_\kappa$ is a mapping from each context $x \in X$ to the set $\mathtt{R}_\kappa(x)$ of all attacks between arguments in $\mathtt{A}_\kappa(x)$.

Among the typical extension-based semantics of argumentation frameworks [Dung, 1995] we adopt the *grounded semantics* for two reasons: it gives rise to a single model — in line with psychological evidence on the construction of a single intended model in human reasoning [Stenning and Lambalgen, 2012] — and the model's computation is efficient.

The ***grounded model*** $\sigma_\kappa(x)$ of a knowledge base κ in a context x is the unique set $in_i, i \to \infty$ of arguments that satisfies the following conditions: $\mathtt{in}_0 = \mathtt{out}_0 = \emptyset$; for every $i > 0$, $\mathtt{in}_i$ is the set of arguments in $\mathtt{A}_\kappa(x) \setminus \mathtt{out}_{i-1}$ that are not attacked by arguments in $\mathtt{A}_\kappa(x) \setminus \mathtt{out}_{i-1}$; and for every

$i > 0$, $\texttt{out}_i$ is the set of arguments in $\mathtt{A}_\kappa(x)$ that are attacked by arguments in $\texttt{in}_{i-1}$. An ***inference*** τ is ***supported*** by the grounded model $\sigma_\kappa(x)$ if $\sigma_\kappa(x)$ includes an argument for τ.

Although computing the grounded model $\sigma_\kappa(x)$ of κ on x is efficient in the number of arguments in $\mathtt{A}_\kappa(x)$, this number can be exponentially larger than the sizes of κ and x. It is an easy exercise to construct a context x and a knowledge base κ with t rules, such that $\sigma_\kappa(x)$ includes $O(2^t)$ arguments and, hence, trivially requires time exponential in t to be computed.

To avoid this exponential blowup in the size of the representation and in the efficiency of the computation, we introduce the ***dual representation*** $\langle x, \widetilde{\sigma_\kappa}(x)\rangle$ of a grounded model $\sigma_\kappa(x)$, where $\widetilde{\sigma_\kappa}(x)$ is defined to be the set of rules that appear in $\sigma_\kappa(x)$. Unlike work in assumption-based argumentation that also appeals to a concise representation of arguments [Craven and Toni, 2016], our approach can cope with directed cycles in the rules, and does not exclude the representation of any argument. Our first result shows that, indeed, dual representations have a one-to-one mapping to grounded models.

Theorem 4.1 *Let $\langle x, \widetilde{\sigma_\kappa}(x)\rangle$ be the dual representation of the grounded model $\sigma_\kappa(x)$ of a knowledge base κ in a context x. Then, $\sigma_\kappa(x)$ is the set of arguments in x under $\widetilde{\sigma_\kappa}(x)$.*

Proof sketch: Every $\alpha \in \sigma_\kappa(x)$ is an argument in x under $\widetilde{\sigma_\kappa}(x)$. Let, now, α be an argument in x under $\widetilde{\sigma_\kappa}(x)$. If α is attacked on rule r by an argument $\alpha_1 \in \mathtt{A}_\kappa(x)$, and since r is a rule in an argument $\alpha_2 \in \sigma_\kappa(x)$, then α_1 attacks α_2. Since $\alpha_2 \in \sigma_\kappa(x)$, then an argument $\alpha_3 \in \sigma_\kappa(x)$ attacks α_1. Thus, α is included in $\sigma_\kappa(x)$ during its iterative construction. □

Our second result shows that the dual representation can, as claimed, be computed efficiently, justifying our approach to introduce this concise representation of the grounded model.

Theorem 4.2 *Consider a knowledge base κ and a context x. Then, $\widetilde{\sigma_\kappa}(x)$ can be computed in time polynomial in the size of κ and x. Furthermore, there exists no algorithm that computes $\widetilde{\sigma_\kappa}(x)$ in time sub-linear in the size of κ and x.*

Proof sketch: Starting from x, repeatedly apply modus ponens to construct the inference graph G of κ. Mark literals in x and repeat the following until convergence: remove literals that conflict with marked literals; retain only the crown rules of arguments in G; mark rules r_1 whose body literals are all currently marked, and for which any other conflicting rule r_2 is such that $r_1 \succ r_2$; mark the head of every marked rule. Return the marked part of G, which is x and $\widetilde{\sigma_\kappa}(x)$. □

Putting the argumentation syntax and semantics above briefly aside, we proceed to offer a learning-theoretic semantics for machine coaching, by considering a variant of the typical PAC definition [Valiant, 1984] that accommodates: *(i)* a bilateral communication between the learner and the target, as done during online learning (i.e., given an observation, make a prediction and then get advice); *(ii)* an arbitrary advice in response to a non-acceptable prediction made by the learner, without necessarily specifying an acceptable prediction; *(iii)* a learning goal that is not to identify the advice coming from the target (i.e., given an observation and a prediction, identify the advice), but to conform to the advice (i.e., given an observation, identify a prediction that leads to no more advice).

A ***hypothesis*** function $h : I \to O$ maps inputs to outputs. A ***feedback*** function $f : I \times O \to A$ maps inputs and outputs to pieces of advice in A. We ask that A includes, at least, the special "no advice" element ✓, and that for every input $x \in I$ there exists at least one output $y \in O$ such that $f(x, y) = ✓$. Then, a hypothesis $h : I \to O$ is $(1 - \varepsilon)$***-approximately conformant*** under the probability distribution D against a feedback function $f : I \times O \to A$ if $f(x, h(x)) = ✓$ for an input x drawn from D, except with probability at most ε over the randomness of sampling inputs from D. The resulting definition of learnability that we shall adopt is given below.

Definition 4.1 *An algorithm is a **(probably approximately) conformant learner** for the feedback class F (with input, output, and advice spaces I, O, and A, respectively) using the hypothesis class H (with input and output spaces I and O, respectively) if for every real values $\delta, \varepsilon \in (0, 1]$, every probability distribution D over inputs in I with representation size n, and every feedback function $f \in F$ with a representation size s, the algorithm is given access to δ, ε, F, and it can repeatedly invoke the following procedure:*

> *It **passively** draws an input x from D, or it **actively** chooses an input x from I; it selects an output y from O; and it then asks for and receives $f(x, y)$. Each input x is a learning **example**, each output y is a **prediction**, and $f(x, y)$ is a piece of **advice**.*

*After time at most $g(1/\delta, 1/\varepsilon, n, s)$ the algorithm terminates and returns, except with probability at most δ, a hypothesis function $h \in H$ that is $(1 - \varepsilon)$-approximately conformant under D against f. If the function g grows only polynomially in its parameters, then the algorithm is **efficient**.*

Although we will not undertake here a systematic comparison between conformant learnability and PAC learnability, it should be intuitively obvious that the choice of the feedback class is what ultimately determines what is learnable: *(i)* there exist feedback classes for which conformant learnability reduces to PAC learnability; *(ii)* there also exist feedback classes for which conformant learnability supports the learning of strictly more expressive structures than PAC learnability.

Below, we construct a type *(ii)* feedback class F_* that, we hypothesize, could be cognitively compatible with the kind of advice that humans can offer under our argumentation-based machine coaching paradigm. A study to evaluate this hypothesis could measure the cognitive load of humans while they engage in machine coaching, using subjective methods (such as self-reported questionnaires), behavioral measures (such as reaction time and eye movement), and / or physiological measures (such as heart rate and pupil dilation); see, e.g., [Skulmowski and Rey, 2017]. For this paper, we restrict ourselves to establishing the efficient conformant learnability for F_*.

Consider a human with a private knowledge base $\kappa = \langle \varrho, \succ \rangle$, and a machine aiming to predict what actions to take according to κ in various contexts; i.e., the machine seeks to learn to compute the dual grounded model $\langle x, \widetilde{\sigma_\kappa}(x)\rangle$ of κ in various contexts x. The machine operates in an online fashion, maintaining a hypothesis represented as a knowledge base, and communicating to the human the rules y of the arguments induced from that knowledge base (thus, $O \triangleq 2^R$) to support its decisions in given contexts x (thus, $I \triangleq X$). The human, on the other hand, responds by offering a piece of advice $f_\kappa(x, y)$. Although a variety of protocols can be considered on how f_κ is selected, we shall analyze just one particular such protocol, which uses an advice space $A \triangleq \{✓\} \cup (\{\textit{unrecognized}, \textit{superfluous}, \textit{incomplete}, \textit{indefensible}\} \times 2^R)$

and returns advice according to the strategy below, breaking ties arbitrarily as needed (e.g., for multiple superfluous rules):

- $f_\kappa(x,y)=\langle \textit{unrecognized}, \{r\}\rangle$ for a rule $r \in y$ that does not belong in ϱ. Hence, if the machine uses a rule r in its prediction y that is not in the human's knowledge base, then the human responds that they do not recognize r.
- $f_\kappa(x,y)=\langle \textit{superfluous}, \{r\}\rangle$ for a rule $r \in y$ that is not in an argument in x under y. Hence, if the machine uses a rule r in its prediction y that is not contributing to any argument, then the human responds that r is superfluous.
- $f_\kappa(x,y) = \langle \textit{incomplete}, \{r\}\rangle$ for a rule $r \in \widetilde{\sigma_\kappa}(x) \setminus y$ that is in an argument in x under $y \cup \{r\}$. Hence, if the machine fails to use a rule r in its prediction y that appears, however, in the human's grounded model, and had the machine used r it would have contributed to another machine argument, then the human responds that y is incomplete because of the non-inclusion of r in y.
- $f_\kappa(x,y)=\langle \textit{indefensible}, \varrho_0\rangle$ for an argument $\langle x_0, \varrho_0\rangle$ in x under ϱ that attacks an argument in x under y, but it is not attacked by any argument in $\sigma_\kappa(x)$. Hence, if the machine's prediction y has an argument that is attacked by an argument ϱ_0, and this attack cannot be defended by any argument in the human's grounded model, then the human responds that the attack from ϱ_0 is indefensible.
- $f_\kappa(x,y)=\checkmark$ when none of the above conditions is met. Hence, if the human does not find one of the above reasons to object to the machine's prediction y, then the human responds that they have no more advice to give.

Effectively, the human re-programs the machine's knowledge base, not by providing arbitrary "code snippets", but by offering reasons on why the machine's explanation y of its decision differs from the user's grounded model (or its dual representation) that the machine is looking to identify, and which corresponds to the explanations that the human would expect. For this advice protocol we can show the following:

Theorem 4.3 *Consider knowledge bases $\kappa = \langle \varrho, \succ\rangle$ whose rules can be ordered linearly based on $\succ$, and a feedback class F_* of feedback functions f_κ that adhere to the advice protocol above. Then, there exists an efficient conformant learner for F_*, even if restricted to passively drawing inputs.*

Proof sketch: Start from an empty knowledge base κ_0. For every passively drawn input x, predict $y = \widetilde{\sigma_{\kappa_i}}(x)$, receive $f_\kappa(x,y)$, and update κ_i into κ_{i+1} by: removing unrecognized or superfluous rules, and adding rules that cause incompleteness or counterarguments with priorities higher than existing conflicting rules. Repeat until $f_\kappa(x,y)=\checkmark$ for m consecutive cycles, with m being polynomial in the relevant parameters and chosen following standard PAC proof techniques. □

The efficiency of the conformant learner from Theorem 4.3 rests on the efficiency of reasoning (cf. Theorem 4.2). This interplay between learning and reasoning echoes past work showing that the two processes cannot be decoupled [Michael, 2014], which ends up limiting the depth of reasoning that can be tractably supported, in line with psychological evidence on the bounded depth of human reasoning [Balota and Lorch, 1986]. The efficient transfer of knowledge supported by machine coaching allows this bound to be circumvented, confirming our earlier point that machine coaching allows more to be learned than what is autonomously learnable by machines.

5 Further Extensions and Considerations

For the further and fuller development of a machine coaching theory, we would consider several key directions: *(i)* accommodating imperative / procedural knowledge, inspired by approaches such as program induction and automatic programming (see, e.g., [Ellis and Gulwani, 2017]); *(ii)* developing more advanced machine coaching protocols, inspired by approaches such as curriculum learning [Bengio *et al.*, 2009] and coactive learning [Shivaswamy and Joachims, 2015]; *(iii)* acknowledging the potentially inadvertent effects that a machine's prediction may have on its own realizability [Michael, 2015a; 2015b] due to any consequent actions or reactions by the machine, the human, or the environment at large; *(iv)* establishing conditions under which arguments can be learned autonomously, while anticipating that the integration of learning and argumentation-based reasoning might affect considerably what can be provably shown to be learnable [Michael, 2008; 2014]; and *(v)* defining metrics and computing, or empirically measuring, the cognitive load of humans when being engaged in coaching, compared to programming or annotating data.

Although machine coaching explicitly seeks to facilitate a *bilateral* understanding between the two parties, this paper's perspective is, admittedly, more geared towards a machine's understanding of the human, than the other way around. Future work taking the complementary perspective [Carroll and Olson, 1988] could examine whether a human's degree of understanding of the machine as achieved via machine coaching is comparable to that achieved via task-specific approaches (see, e.g., [Kulesza *et al.*, 2015]). We expect that some kind of visualization of the machine's explanations in the form of graphs might help to further enhance human understanding (see, e.g., [Michael, 2017; Rodosthenous and Michael, 2019]).

Far from being a silver bullet, we anticipate that the machine coaching paradigm — sitting between machine programming and machine learning, and explicating a type of human-machine interaction that is typically at the fringes of the other two paradigms — will prove to be useful both for debugging and personalizing user-independent knowledge that has been gathered through the other two paradigms, and for efficiently, effectively, and incrementally gathering user-specific knowledge for domains with: highly user-specific preferences, repetitive everyday tasks, explanations that users can verbalize, and without critical ramifications in case of mistaken actions.

In an era of voice-controlled and triggered-based home automation systems such as Siri, Alexa, and IFTTT [2019], the prospect of developing cognitive assistants (even for selected everyday tasks) that are capable of learning on the job, explaining their decisions, and being advised and corrected when making mistakes, offers a prime opportunity to apply and evaluate the unique features of machine coaching, and could act as a first step towards the wider adoption of machine coaching as a catalyst for the disruptive automation of knowledge work.

Acknowledgments

This work was supported by funding from the EU's Horizon 2020 Research and Innovation Programme under grant agreements no. 739578 and no. 823783, and from the Government of the Republic of Cyprus through the Directorate General for European Programmes, Coordination, and Development.

The author would like to thank the anonymous XAI 2019 reviewers for their thoughtful and constructive feedback.

References

[Balota and Lorch, 1986] David A. Balota and Robert F. Lorch. Depth of Automatic Spreading Activation: Mediated Priming Effects in Pronunciation but Not in Lexical Decision. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 12(3):336–345, 1986.

[Bengio *et al.*, 2009] Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum Learning. In *Proc. 26th Annual International Conference on Machine Learning*, pages 41–48, Montreal, QC, Canada, 2009.

[Carlson *et al.*, 2010] Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka, and Tom M. Mitchell. Toward an Architecture for Never-Ending Language Learning. In *Proc. 24th AAAI Conference on Artificial Intelligence*, pages 1306–1313, Atlanta, GA, U.S.A., 2010.

[Carroll and Olson, 1988] John M. Carroll and Judith R. Olson. Mental Models in Human-Computer Interaction. In *Handbook of Human-Computer Interaction*, pages 45–65. North-Holland, 1988.

[Craven and Toni, 2016] Robert Craven and Francesca Toni. Argument Graphs and Assumption-Based Argumentation. *Artificial Intelligence*, 233:1–59, 2016.

[Diakidoy *et al.*, 2017] Irene-Anna Diakidoy, Loizos Michael, and Antonis Kakas. Knowledge Activation in Story Comprehension. *Journal of Cognitive Science*, 18(4):439–471, 2017.

[Dung *et al.*, 2009] Phan M. Dung, Robert A. Kowalski, and Francesca Toni. Assumption-Based Argumentation. In *Argumentation in Artificial Intelligence*, pages 199–218. Springer, 2009.

[Dung, 1995] Phan M. Dung. On the Acceptability of Arguments and its Fundamental Role in Nonmonotonic Reasoning, Logic Programming, and n-Person Games. *Artificial Intelligence*, 77(2):321–357, 1995.

[Ellis and Gulwani, 2017] Kevin Ellis and Sumit Gulwani. Learning to Learn Programs from Examples: Going Beyond Program Structure. In *Proc. 26th International Joint Conference on Artificial Intelligence*, pages 1638–1645, Melbourne, Australia, 2017.

[IFTTT, 2019] IFTTT. If This Then That. (http://ifttt.com), 2019.

[Kakas and Michael, 2016] Antonis Kakas and Loizos Michael. Cognitive Systems: Argument and Cognition. *IEEE Intelligent Informatics Bulletin*, 17(1):14–20, 2016.

[Kuhn, 2014] Tobias Kuhn. A Survey and Classification of Controlled Natural Languages. *Computational Linguistics*, 40(1):121–170, 2014.

[Kulesza *et al.*, 2015] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proc. 20th International Conference on Intelligent User Interfaces*, pages 126–137, Atlanta, GA, U.S.A., 2015.

[Lippi and Torroni, 2016] Marco Lippi and Paolo Torroni. Argumentation Mining: State of the Art and Emerging Trends. *ACM Transactions on Internet Technology*, 16(2):10, 2016.

[Manyika *et al.*, 2013] James Manyika, Michael Chui, Jacques Bughin, Richard Dobbs, Peter Bisson, and Alex Marrs. Disruptive Technologies: Advances That Will Transform Life, Business, and the Global Economy. McKinsey Global Institute, 2013.

[McCarthy, 1959] John McCarthy. Programs with Common Sense. In *Proc. Teddington Conference on the Mechanization of Thought Processes*, pages 75–91, London, England, U.K., 1959.

[Mercier and Sperber, 2011] Hugo Mercier and Dan Sperber. Why Do Humans Reason? Arguments for an Argumentative Theory. *Behavioral and Brain Sciences*, 34(02):57–74, 2011.

[Mercier, 2016] Hugo Mercier. The Argumentative Theory: Predictions and Empirical Evidence. *Trends in Cognitive Sciences*, 20(9):689–700, 2016.

[Michael, 2008] Loizos Michael. *Autodidactic Learning and Reasoning*. Doctoral Dissertation, Harvard University, Cambridge, MA, U.S.A., 2008.

[Michael, 2009] Loizos Michael. Reading Between the Lines. In *Proc. 21st International Joint Conference on Artificial Intelligence*, pages 1525–1530, Pasadena, CA, U.S.A., 2009.

[Michael, 2013a] Loizos Michael. Machines with WebSense. In *Proc. 11th International Symposium on Logical Formalizations of Commonsense Reasoning*, Ayia Napa, Cyprus, 2013.

[Michael, 2013b] Loizos Michael. Story Understanding... Calculemus! In *Proc. 11th International Symposium on Logical Formalizations of Commonsense Reasoning*, Ayia Napa, Cyprus, 2013.

[Michael, 2014] Loizos Michael. Simultaneous Learning and Prediction. In *Proc. 14th International Conference on Principles of Knowledge Representation and Reasoning*, pages 348–357, Vienna, Austria, 2014.

[Michael, 2015a] Loizos Michael. Introspective Forecasting. In *Proc. 24th International Joint Conference on Artificial Intelligence*, pages 3714–3720, Buenos Aires, Argentina, 2015.

[Michael, 2015b] Loizos Michael. The Disembodied Predictor Stance. *Pattern Recognition Letters*, 64(C):21–29, 2015.

[Michael, 2017] Loizos Michael. The Advice Taker 2.0. In *Proc. 13th International Symposium on Commonsense Reasoning*, London, England, U.K., 2017.

[Mitchell *et al.*, 1986] Tom M. Mitchell, Richard M. Keller, and Smadar T. Kedar-Cabelli. Explanation-Based Generalization: A Unifying View. *Machine Learning*, 1(1):47–80, 1986.

[PCAST, 2010] PCAST. Report to the President and Congress, Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. *Executive Office of the President, U.S.A.*, 2010.

[PCAST, 2013] PCAST. Report to the President and Congress, Designing a Digital Future: Federally Funded Research and Development in Networking and Information Technology. *Executive Office of the President, U.S.A.*, 2013.

[Prakken, 2010] Henry Prakken. An Abstract Framework for Argumentation with Structured Arguments. *Argument and Computation*, 1(2):93–124, 2010.

[Rodosthenous and Michael, 2019] Christos Rodosthenous and Loizos Michael. Web-STAR: A Visual Web-Based IDE for a Story Comprehension System. *Theory and Practice of Logic Programming*, 19(2):317–359, 2019.

[Shivaswamy and Joachims, 2015] Pannaga Shivaswamy and Thorsten Joachims. Coactive Learning. *Journal of Artificial Intelligence Research*, 53(1):1–40, 2015.

[Simard *et al.*, 2017] Patrice Y. Simard, Saleema Amershi, David M. Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and John Wernsing. Machine Teaching: A New Paradigm for Building Machine Learning Systems. *Online Preprint arXiv:1707.06742*, 2017.

[Skulmowski and Rey, 2017] Alexander Skulmowski and Günter D. Rey. Measuring Cognitive Load in Embodied Learning Settings. *Frontiers in Psychology*, 8(1191):1–6, 2017.

[Stenning and Lambalgen, 2012] Keith Stenning and Michiel Van Lambalgen. *Human Reasoning and Cognitive Science*. The MIT Press, Cambridge, MA, U.S.A., 2012.

[Valiant, 1984] Leslie G. Valiant. A Theory of the Learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.