A Teaching Language for Building Object Detection Models

Nicole Sultanum
University of Toronto
nicolebs@cs.toronto.edu

Soroush Ghorashi, Christopher Meek,
Gonzalo Ramos
Microsoft Research
{soroush.ghorashi,meek,goramos}@microsoft.com

ABSTRACT

Object detection is a key application of machine learning. Currently, these detector models rely on deep networks that offer model builders limited agency over model construction, refinement and maintenance. Human-centered approaches to address these issues explore the exchange of knowledge between a human-in-the-loop and a learning system. This exchange, mediated through a teaching language, is often restricted to the specification of labels and constrains user expressiveness communicating other forms of knowledge to the system. We propose and assess an expressive teaching language for specifying object detectors which includes constructs such as concepts and relationships. From a formative study, we identified language building blocks and articulated design goals for creating interactive experiences in teaching object detection. We applied these goals through a design probe that highlighted further research questions and a set of design takeaways.

Author Keywords

Teaching Language; Machine Teaching; Object Detection; Qualitative Study; Interactive Machine Learning

CCS Concepts

•**Computing methodologies** → **Object detection;** •**Human-centered computing** → **Interaction design process and methods;**

INTRODUCTION

Machine Learning (ML) is a powerful technology that enables systems to perform a variety of tasks with accuracies rivaling human capabilities. Among the tasks ML supports is that of locating concepts within an image, *i.e.*, *object detection*. Examples of detection include locating cars in street cameras images or tumors from an X-ray. Core to this task is having access to labeled data depicting a given target concept. In particular, the choice of such labeled examples significantly affects the resulting model's performance and labeling costs.

State-of-the-art learning methods to build such ML models often rely on deep neural networks (DNN) [2, 11, 18] which require large amounts of labeled data for training and insight

from ML experts. This context presents challenges for the creation of specific models where there is a limited pool of subject-domain experts. Other approaches requiring fewer labels [20, 23, 31] leverage existing DNNs from a certain domain and repurpose them to a different one, which have been proved effective but can be opaque in conveying how predictions are made or how to fix prediction errors. From a human-centric point of view, this translates to limited or no ability to debug or correct a model's predictions. It is also challenging to maintain or adapt a model once deployed and in the face of new unseen data, *e.g*, adapting to highly dynamic domains or changeable requirements.

While techniques exist to alleviate some of the above issues and give opaque learning systems some measure of explainability [19, 25, 32, 33], some models lack the semantic features to make such transparency actionable. A different method for addressing these issues is *Interactive Machine Learning* (IML), where a human in the loop "iteratively builds and refines a mathematical model to describe a concept through iterative cycles of input and review" [7]. Core to IML experiences are ways in which users interact with a learning system, which can be framed as taking place under an interaction *language* used to exchange knowledge with a system.

The Cambridge dictionary [1] describes *Language* as *"a system of communication used by people (..) in a type of work"*. In this paper, we adapt this definition to include a (a) computational learning system, and (b) a human user conveying knowledge to such a learning system. In this context, a *teaching language* is the system of communication that enables the *exchange of knowledge between a human acting as a teacher, and a computational learner*. We argue that this teacher-student perspective can improve the practice of creating object detection models by increasing the agency users have when building these models. Offering more expressive and intuitive means to create such models is also of critical advantage to experts in a particular application domain (*e.g.*, medicine and sciences) but novices to ML. Within this idea, we also argue for *leveraging domain user knowledge beyond labels*, *e.g.* semantic features or structural constraints. This extension has been conceptualized as a potential solution for this problem [28] and suggests a path towards more efficient model building but has yet to be validated in more concrete scenarios.

Following, this work aims to identify and develop *teaching languages to express concepts beyond labels to computational learning systems*, and learn how these languages are used by human teachers. We are interested in both the language structure, how it can manifest through an interface, and the

ways people use it within a learning system. Object detection provides an interesting area of application for this exploration. First, it represents a task in which humans excel without any particular training (as compared to machines), allowing us to tap on a larger "expert" population. Second, object detection is a generalization of classification, and by addressing the former case we can potentially cover the latter. Building on the notions of human teachers and computational learners, we focus on the following research questions:

RQ1. What are the core elements of a teaching language to allow expressing a wide range of concepts to a machine learner?

RQ2. How does one build this language into an interactive experience supporting a person's teaching to a machine learner?

RQ3. How do teachers understand and use this language in the context of a machine teaching experience?

We followed a user-centered process to answer these questions. Through a formative study, we elicited building blocks for a teaching language and distilled design goals for interactive teaching experiences for object detection. These outcomes are the first set of contributions from our paper. We applied our design goals through the implementation of a teaching language via an interactive prototype to teach object detection models, which form our second set of contributions. We then conducted an evaluation of this prototype that provided insight on how teachers used the language and challenges they faced. Our investigation led to important implications for the design of teaching languages for learning systems and experiences, forming our third set of contributions.

BACKGROUND

Interactive Machine Learning
The field of Interactive Machine Learning (IML) [8] is directly connected to our goals of improving the experience of building object detection models with a human-in-the-loop and of outlining a teaching language for the task. Dudley and Kristensson [7] present a comprehensive survey on the topic of IML as an interaction paradigm in which a user "iteratively builds and refines a mathematical model to describe a concept through iterative cycles of input and review". In particular, IML refers to supervised ML processes where people participate in "rapid, focused and incremental learning cycles" [4]. IML has been successful in lowering the bar of entry for non ML-experts to benefit from ML models. Systems like Wekinator [9], CueFlik [10], and NorthSt*r [14] are examples of IML systems targeting a range of different scenarios.

It has been argued that IML can lead to building better ML models. Branson *et al.* [5] show how in a hybrid human-computer process of building image classifiers, human intervention can drive up recognition accuracy while computer vision intervention decreases the amount of human effort. Similarly, Holzinger *et al.* [12] show how human intelligence can positively influence machine intelligence in the context of an IML system training agents how to solve the travelling salesman problem. Our work focuses on extending human intervention from a traditional role of producing labels.

User-Machine Exchanges
Early studies eliciting user feedback on automated explanations [6, 17, 29] revealed user expectations for this feedback, including (a) the importance of dialogue [6], (b) a tendency to reason about ML output in terms of decision "rules" (*e.g.*, "why" and "if-then" explanations) [17, 29] and (c) a tendency to provide more positive than negative feedback [17]. Later works also revealed that (a) users want to provide more to systems than just labels [4, 29], (b) that people value transparency in learning systems [4, 7], and (c) this transparency can help users give better feedback to the system [4, 16]. These findings suggest that a teaching language should support an *interchange* between teacher and learning system (*i.e.*, both sides providing and receiving feedback), and that users are willing to engage with the learning system on a deeper level *beyond labels* (*e.g.*, providing more detailed feedback and better understanding its inner workings). This interchange beyond labels and predictions has been looked at from various perspectives. For example, machine teaching [28] is a noteworthy emerging paradigm that embraces the human-in-the-loop taking the role of a teacher to a learning system, and seeks to improve the way non ML-experts can create ML models. Our work is influenced by this teaching perspective.

Machine Teaching
Simard *et al.* [28] describe machine teaching as a process where any information processing skill (*i.e.*, function) teachable to a human should be as easily be taught to a machine [1]. Machine teaching emphasizes teaching efficiency and low barrier of access while producing semantic ML models from the ground up. Machine teaching is a form of IML that proposes abstracting the complexities of ML algorithms and their parameters, making it accessible to end-users that need only subject-matter expertise and have the inherent capacity to teach. In particular, it argues for teacher knowledge that extends beyond labels (*e.g.*, semantic features, sampling strategies, and schema specification) and which can be taught to a learning system. As a result, ML models built using a machine teaching process are semantic, and provide a form of explainability [3] from the ground up. They are also arguably easier to maintain and adapt to changes in data distribution. Wall *et al.* introduced MATE [30], a teaching environment for text document classification that formalizes aspects of the machine teaching loop while providing insights on teaching patterns to help novice users be better teachers. Machine teaching puts into perspective other ML activities. For example, active learning [27] can be part of a machine teaching session by providing teachers with a form of machine-in-the-loop sampling. Our work embraces this people-as-teachers perspective in the context of object detection models, to make them more human-centric and efficient through the articulation of a teaching language.

Object Detection
There is an extensive body of work on object detection, particularly in the context of computer vision. Object detection entails both detection and localization of specific objects in

---

[1] The term machine teaching can be used in other contexts. Zhu [35] defines it as the "inverse problem to ML" where one finds the optimal training set, given a particular learning algorithm and a target model.

ID	Concept	Age	Highest Degree	ML Exposure	Gender	Num. Samples per Task	
						Labeling	**Correction**
F01	Bird's nest in use	18-29	Masters	Fair	Male	5	5
F02	Bird's nest in use	18-29	Masters	High	Male	3	3
F03	Bird's nest in use	40-49	Bachelor	Low	Male	4	4
F04	Bird's nest in use	18-29	Bachelor	Some	Female	3	3
F05	Person playing tennis	18-29	Masters	High	Male	5	4
F06	Person playing tennis	30-39	Bachelor	Low	Female	4	4
F07	Person playing tennis	30-39	Masters	Low	Male	5	5
F08	Person playing tennis	30-39	Doctoral	Low	Female	3	3
F09	Person riding bicycle	30-39	Masters	Low	Male	4	3
F10	Person riding bicycle	40-49	Masters	Low	Female	2	2
F11	Person riding bicycle	30-39	Bachelor	Some	Male	3	4
F12	Person riding bicycle	40-49	Bachelor	Fair	Male	3	1

Figure 1. Background and concept distribution for participants in the formative study. Self reported exposure to ML varied between: `low` *"never built a ML model"*; `some`, *"took classes but never built models in practice"*; `fair`, *"occasionally builds models in practice"*; and `high`, *"often builds models in practice"*.

an image. The location of an object is often represented by a bounding box or a pixel-level demarcation. Some of the earliest works on IML focused on the task of separating foreground and background from an image [8]. Recently, these efforts go beyond object detection towards more complete descriptions of images, including objects, attributes, and relationships to other objects (*e.g.*, [15]). These systems are trained to detect a fixed set of classes of objects, relationships, and attributes, and their training datasets require thousands of crowdworkers to create detailed labels for a large number of images (*e.g.*, >300k for MS-COCO [18] and >100k for Visual Genome [15]).

Particularly important in practice is the ability to train detectors for new classes of objects, attributes and relationships. This research leverages supervised, semi-supervised and unsupervised domain adaptation techniques, and recent work has shown that these tasks can be accomplished using limited amounts of training data [23, 24, 26, 31]. In our work, we leverage the Microsoft Azure Cognitive Service for object detection [20] to accomplish this. In particular, we explore how using a teaching language can enable a subject-matter expert to efficiently create ML models for object detection.

FORMATIVE STUDY
The goal of a teaching language is to allow teachers to articulate knowledge beyond *labels*, *features* of interest and *schema* [28] to explain concepts unknown to a learning system. Given the little exploration done in this space and our goal to inform prototyping efforts, we conducted a formative role-play study to elicit how people would teach a hypothetical computational learner to detect a visible concept.

Methodology
We recruited 12 individuals (F01-F12) within a large technology company with various backgrounds to take part in the study (Fig. 1). Participants played the role of teachers to a hypothetical learning system called `Pixie`, while a study facilitator played the role of the system's interpreter, *i.e.*, the *"input and output"* of the system. Participants were told the system could understand anything the facilitator could understand, and that they were free to express themselves in their preferred medium, *e.g.*, verbal, drawings and gestures. We deliberately presented `Pixie` as a separate entity from the human facilitator so participants would consider the computational nature

of the learner and its limited contextual understanding of the world while still allowing for a full range of expression.

Sessions were about 1 hour long. We gave participants one target concept out of the following—`person riding bicycle`, `person playing tennis`, `bird's nest in use`— and asked them to explain to `Pixie` how it can *identify instances of this concept in images*. We chose concepts that were common knowledge and easily recognizable but complex in composition (*i.e.*, spanning multiple parts and requiring specific configurations between them). Sessions spanned 3 teaching tasks, and finished with an **interview**. The teaching tasks were preceded by a short **practice** round where participants performed the 3 tasks with a simpler concept (`orange`, the fruit). This gave participants an overview of what to expect. Participants received a $25 gift card as appreciation for their time. The 3 teaching tasks were as follows:

1. Foresight task (5-10 minutes). We asked participants to freely explain the given concept without visual aids. Based on related concepts introduced by their explanations, we conveyed scripted requests for clarification from `Pixie` to prompt for details, *e.g.*, *"what is a `racket`?"* and *"can you describe the relationship between `court` and `net`?"*. The script followed a framework where the facilitator conveyed questions when participants mentioned an unknown concept. `Pixie` had an understanding of general known concepts such as `person`, `circle`, and so on. We limited this by time or when the participant decided they were done.

2. Labeling task (15 minutes). We presented participants with image printouts and asked if the concept was present or not (and where), what aspects of the image were considered in their judgement, and anything else they thought was useful for `Pixie` to know. The task always started with a "positive" image (*i.e.*, with the concept present) to help ground explanations from the foresight task, with the following ones being negative and edge cases to encourage reflection. We covered as many images per participant as time allowed, ranging from 2-5 images from a total of 6 images available for each concept.

3. Correction task (15 minutes). We presented image printouts "labeled by `Pixie`" and asked participants to confirm or correct the judgements, explain factors considered, and share anything else useful for the system to know in order to fix detection issues. Participants could also pose questions to `Pixie` about what concepts it was able to see: *e.g.*, for the concept of `person playing tennis`, recognizable image elements included `person`, `racket`, `person holding racket`, `court`, and so on. The curated images depicted edge cases and detection issues to encourage explanations. We covered 1-5 images in this phase, again from a total of 6 per concept.

We video recorded sessions, transcribed and analyzed the videos for explanation patterns via thematic analysis, and conducted open coding for a sample of 4 session transcripts.

Findings
Participants were immersed in the role-play exercise and often referred to `Pixie` in the 3rd person, suggesting that the separate entity metaphor was effective. Feedback also shows that they reflected deeply about their explanations. We organize

findings under two topics: holistic considerations on the experience of providing explanations to a learning system, and perspectives on the emerging explanation language.

The Teaching Experience

Teaching workflow: from general patterns to details. We found teaching workflows to be overall consistent across the 12 participants, marked by an iterative knowledge and description refinement process. For the foresight scenario, the initial reaction was to describe the general case, its elements, and primary relationships, all which some participants called "rules": *"you kind of wanna teach the most common case and then talk about the exceptions after. At least that's how usually people learn"* (F09). They expanded and readjusted these explanations to cover edge cases and situations not initially considered by the teacher: *"I think I was being maybe unrealistically conservative and always qualifying things that 'usually would be like this', 'usually would be like that', I think I started doing that more after I've seen the counter examples that would not be filling the rules."* (F09). Being exposed to learner feedback and to a variety of counter examples in the labeling and correction tasks were instrumental in supporting this refinement process and helped participants think deeper about the concept and re-strategize their explanations, *e.g.*, *"when I saw that there can be a bird's nest like this or which looks very unusual, then I have to adjust myself in order to train Pixie"* (F02).

Awareness. When asked if they adopted any conscious strategies to their explanations, most participants commented on trying to adjust for what they thought the learner knew and could understand (8/12 participants), *e.g.*: *"I tried to use several words to describe the same thing in hopes that it would recognize what one of those words meant even if it didn't understand what all of them meant"* (F06). When asked what knowledge would have helped them provide better explanations, they mentioned better awareness of Pixie's understanding: its known *"vocabulary"* (F12); what it sees in images (2/12); how it *"looks at the world"* (F04), *i.e.*, what it is able to look and learn from in images) (4/12); and whether it is indeed learning (3/12), *e.g.*, *"I think it would be great if I teach the system something, that it [would] show in some way that it has understood"* (F05).

Feedback. Following the above points, participants appreciated having a conversation with the system as means to improve awareness. Questions prompted by the system revealed knowledge gaps and hinted on the learner's reasoning capabilities (2/12) that helped *"deduce how the system thinks"* (F10). Prediction outlines (7/12) and answers to questions (4/12) also showed what the system was able to see and how well it learned what was taught: *"I like when Pixie shows the work"* (F09). Suggestions on how the system could be more proactive in showing what it learned included informing everything it could see in an image (F11), fetching similar examples to justify a prediction (F04), showing recognized examples of a learned concept (F05), and leveraging pre-built knowledge: *"In an ideal world, Pixie would know words in the dictionary and how to identify those things, at least the nouns"* (F12). Finally, F12 also commented on the expectation that the learner-teacher communication language be shared: *"I like it to explain in the photos the same format that I'm gonna use to communicate with it. So, if I can't use a photo, or my voice or text to explain, then I wouldn't expect it to show me that. I would expect it to show me photos"*.

Characterizing the Teaching Language

Language building blocks. Consider this excerpt from a foresight task to describe person playing tennis: *"A person playing tennis holds a racket and hits a yellow-green neon ball. They often wear shorts or a sweatband on their wrist to control their grip on the racket. They mostly play on a green rectangular court that has a roughly waist-high net, so a shorter net that touches the ground."* (F06). It starts with a description of rules for the general case (*e.g.*, *"person holds a racket"*, *"person hits a yellow-green neon ball"*). Structural elements include the presence of (noun) ***concepts*** (*e.g.*, person, racket, ball, sweatband), ***relationships*** (*e.g.*, person holds racket, person on court), and ***attributes*** (*e.g.*, yellow-green, rectangular); these elements are consistent with existing efforts to generate datasets of image descriptions [13, 15].

We also discuss other semantic constraints that could be supported by a teaching language. First, we noted an extensive use of ***uncertainty qualifiers*** (*often*, *mostly*, *roughly*) to make descriptions less stringent and denote importance of certain explanations, *e.g*: *"I had to consciously say, 'okay, this is the minimum of what needs to be met for someone to be playing tennis', and 'these are the things that are nice to have'; they're optional, but they wouldn't necessarily make or break something making that statement true."* (F08). We also noted ***conditional*** statements, *e.g.*, *"But to detect if [a bird's nest is] in use you'd have to see if there are eggs inside a nest"* (F01), and ***cardinality*** constraints *e.g.*,*"A bicycle has two wheels"* (F10), both meant to describe discriminative heuristics. Participants often felt compelled to give these procedural-like explanations. We also observed ***similarity*** associations to leverage other known concepts, *e.g.*: *"bird's nest would typically be somewhat shaped like a disc"* (F01). We finally noted uses of ***negation***; while most explanations tended to be of what things "are" (following past observations of positive bias [4]), some things were easier to explain in terms of what they "are not", *e.g.*: *"The ground here is what we call grass, so it's not paint, it's not gray or dark colored as in most other pictures."* (F09).

Explanation challenges. Some concepts were more difficult to explain. These include (a) abstract notions such as aesthetically pleasing (F04) and natural vs. man made (F01); (b) shapeless or highly irregular shaped entities like ground (F09, F10) and twig (F02); and (c) pattern-evoking concepts like water (F04) and racket mesh (F01). When explanations failed, participants resorted to teaching via **image samples** instead: *"These [abstract concepts] are easier to explain through images rather than words."* (F01). Also, in contrast to a desire to craft rule-like explanations with some generalizable power, participants often preferred explaining an instance in an image: *"It's easier to explain in the context of one image, but it's harder if I know that it's gonna apply that to everything, to feel like I'm teaching it the right way of recognizing that thing"* (F06).

Using images to teach (and learn) about concepts. Beyond explanations in the context of generalizable descriptions of a concept, we also observed explanation patterns for the judgement of image samples and predictions. When faced with an image and asked to explain the rationale for a concept's presence, participants tended to *revisit* explanations provided earlier and assess their occurrence the image as evidence to support their decision, *e.g.*, *"Tennis rules require the players to use a ball that is a tennis ball, as well as a racket in order to touch a tennis ball. And that's not happening here"* (F05). This behavior led participants to reassess their explanations often, which in turn also led to a more refined understanding of the concept they were teaching, *e.g.*: *"The cheat sheet that I wrote down at the end where I was starting to... I very distinctly remember having a fork in the road mental moment of if I continue down this path, I will have rendered myself hypocritical to everything I said before and completely flip flop and change my mind on what is and is not a person playing tennis"* (F08). Participants also appreciated having learner feedback presented in the context of an image. Many commented on the usefulness of the prediction outlines (7/12) and learner's feedback on what else it sees in an image (5/12).

We did not give participants the choice to select image printouts during the study, but some still shared ideas on how those choices could support teaching, *e.g.*: *"If I get a bigger variety, and I can kinda pick and choose the ones that I feel are appropriate for my subjective understanding of playing tennis"* (F05). Testing the learner was also suggested: *"Let's say if it's I just do an online search and then I just give it a picture and say, show me where the nest is or show me where oranges are. Something like what happens in a school or a university, you teach something and you give a test"* (F01).

Design Goals for Interactive Teaching Experiences
We summarized our formative findings with the following design goals for teaching object detection systems:

(**E**)**xpressiveness.** The system should allow for expressive explanations to be crafted. Language elements to consider include *image samples*, *concepts*, *relationships*, *attributes*, *uncertainty qualifiers*, as well as support to *conditionals*, *cardinality*, *similarity* and *negation*. This is consistent with machine teaching conditions of expressiveness of language [28].

(**S**)**amples.** The system should support explanations to be provided in the context of one or more images, allowing the choice of what image samples to teach to the learning system.

(**I**)**nterchange.** The system should support a two-way communication between teacher and learner; teachers can inquire about learner's knowledge and learner can ask for feedback and communicate what it knows. This communication should ideally be supported by a common shared language.

(**A**)**wareness.** The system should provide the means for the teacher to keep track of a learner's progress and current learning state (*e.g.*, current estimated performance and number of samples left to teach until learner can be tested).

BUILDING `Pixie`
Following the formative study, we designed an interactive prototype implementing core aspects of the teaching workflow for a proof-of-concept of our hypothetical learning system, `Pixie`. It covers an *end-to-end teaching experience* of building an object detection model for visible concepts. Our design efforts aimed to: (a) validate findings of the formative study via an application of the design goals to the implementation of an interactive teaching experience; and (b) fill gaps left by the formative study via a larger pool of teaching samples and the use of actual computational learning systems for a more realistic experience. We justify design decisions by referencing our design goals: (**E**), (**S**), (**I**), (**A**).

`Pixie` provides an expressive teaching language including **concepts**, **relationships**, select semantic constraints (*uncertainty* and *cardinality*) under the moniker **modifiers**. We will cover how teachers can leverage this language and system features to *(1) browse samples*, *(2) define concepts*, *(3) provide examples*, and *(4) assess a learner's knowledge*.
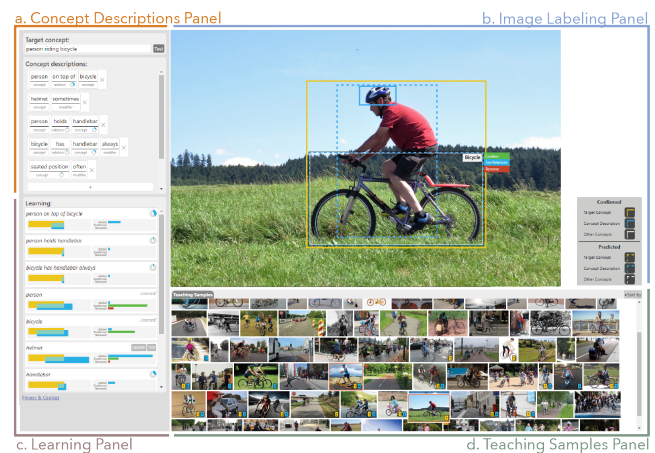

Figure 2. Overview of `Pixie`'s 4 main panels. *Center image by Martin Büdenbender from Pixabay.*

A Teaching Workflow
`Pixie`'s interface encompasses 4 main areas (Fig. 2):

(a) **Concept Descriptions panel**, where teachers can declare explanations in the form of concepts and relations (**E**);

(b) **Image Labeling panel**, to provide examples (*i.e.*, *labels*) to declared concepts and relations from images (**S**)(**I**);

(c) **Learning panel**, to monitor teaching progress (**A**);

(d) **Teaching Samples panel**, an image gallery where teachers can browse images *samples* to teach (**S**).

A workflow encompasses the following activities:

(1) Browsing samples. All four panels are empty at startup. A teacher begins by entering the name of the primary concept they want to teach in the **Target Concept** (TC) box: *e.g.*, `person riding bicycle` (Fig. 2 (a)). The system then populates the Teaching Samples panel (Fig. 2 (d)) with related images by using the TC as a text query to Bing Images [21]. Teachers can work with image samples by clicking on their respective thumbnails, which then appear on the Image Labeling

panel (<u>S</u>) (Fig. 2 (b)). Once an image is selected (*e.g.*, biker on grass, Figs. 2 and 3), `Pixie` presents it alongside concepts it already knows about (*e.g.*, `person` and `bicycle`), with each concept instance enclosed in white dashed *bounding boxes* (<u>I</u>) (Fig. 3(a)). Dashed outlines indicate that the system made a **prediction**, which can be confirmed by the teacher later.
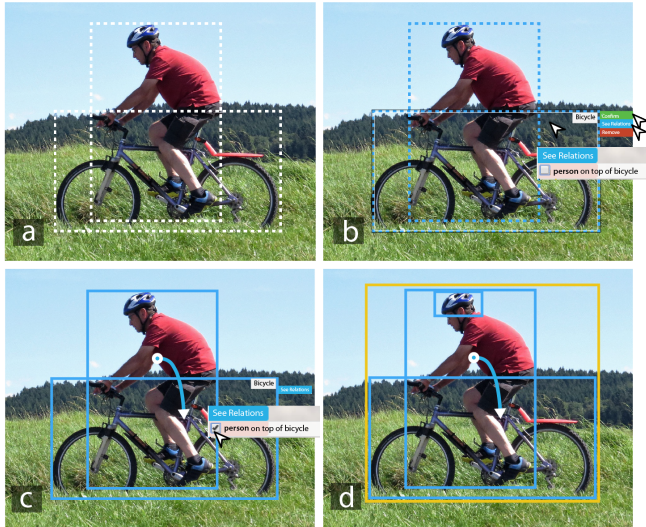

Figure 3. Labeling steps: (a) concepts known by `Pixie`, `person` and `bicycle`, before they are added to CDs (dashed white lines); (b) `person` and `bicycle` after they are added to CDs (dashed blue lines), plus visible floating menu for user to confirm predictions and add relations; (c) predictions are confirmed (solid blue lines) and relationship is added (arrow for `person on top of bicycle`); (d) teacher-added label for `helmet` (solid blue line) and `person riding bicycle` (solid yellow line, TC).

(2) Defining concepts. After exploring the Teaching Samples and learning what concepts the system already knows about (*e.g.*, `person` and `bicycle`), the teacher may want to introduce new ones (*e.g.*, `helmet`) via **concept descriptions (<u>I</u>)**. A concept description (CD) is what we call an explanation block that implements the system's teaching language building blocks, *i.e.*, **concepts**, **relations** and (optional) **modifiers (<u>E</u>)**. Possible configurations include:

– `concept` (+ `modifiers`);
– `concept` + `relation` + `concept` (+ `modifiers`).

The teacher decides to add two CDs as follows:

 A relationship between the known concepts of `person` and `bicycle`.
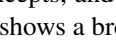
 Introduces a new concept (`helmet`) and adds an uncertainty qualifier (`sometimes`) to denote that it may not appear as often.

New CDs are added via the button, one building block at a time (Fig. 2 (a)). After adding the CDs above, the predicted bounding boxes for `person` and `bicycle` turn blue to indicate that these concepts are now listed in the Concept Descriptions panel and are therefore "important to know" (Fig. 3(b)).

(3) Teaching examples. When hovering over a bounding box, a floating menu appears (Fig. 3(b)). From this menu, the teacher can confirm that `Pixie`'s predictions for `person` and

`bicycle` are correct ("Confirm") or not ("Remove"); bounding box lines go from dashed to solid, indicating they have been checked by the teacher (<u>I</u>) (Fig. 3(c)). From the same menu, under "See Relations", teachers can add a **relation** label between `person` and `bicycle` as outlined in the CDs, which is displayed as an arrow (Fig. 3(c)). The figure shows an unlabeled instance of `helmet` and of the target concept itself, `person riding bicycle`, which are not outlined by Pixie since it hasn't learned enough about them to make predictions yet. Teachers can add new labels as bounding boxes via click & drag (Fig. 3(d)); like "confirmed" labels, these boxes have solid outlines, with TC labels outlined in yellow (<u>I</u>)(<u>S</u>).

(4) Assessing a learner's knowledge. `Pixie` requires 16 image samples[2] before it can make predictions about new concepts (*e.g.*, `helmet` and the TC concept itself, `person riding bicycle`). Teachers can monitor labeling progress via the Learning Panel (<u>A</u>) (Fig. 2 (c)), which lists all the **learnable** components taught in CDs, *i.e.*, concepts and relations that generate predictions. Each Learning Panel entry for a new concept will feature either a *"Learned!"* indicator for default concepts `Pixie` recognizes, or a progress indicator 🔵 showing how many image samples are needed until `Pixie` is able to predict it. When enough samples are taught, the indicator will spin while the underlying detection models are being trained for the new learnable. After learning is done, teachers can then assess the learned concept or relation via the "Test" button (as shown for `helmet` on Fig. 2(c)). It opens a new tab on the Teaching Samples panel with images sampled from the web [21] found to contain instances of the tested concept, which lets teachers assess how well the system has learned and helps guide refinement with a fresh set of image samples.

Teachers can get an overview of learning status and progress via a number of visual indicators (<u>A</u>) (Fig. 2(c)). An overlap widget shows occurrence of image samples containing confirmed labels for the TC (left yellow rectangle), the learnable (right blue rectangle), and both (middle green rectangle), conveying the prevalence of that learnable with respect to the TC. It can reveal correlations, confirm or challenge teacher expectations for co-occurrence of concepts, and outline discrepancies. The bar chart shows a breakdown of teacher labels for the learnable, including teacher-provided labels (top, blue), teacher-confirmed labels (center, green), and teacher-removed labels (bottom, red). This view shows how the learner has fared in predicting this concept throughout the teaching and suggests expected performance.

Prediction Pipeline

While current object detection technology to fully support the proposed teaching language (*i.e.*, CD-based TC prediction) doesn't yet exist, we leveraged a limited form of concept learning to promote a believable teaching experience. We used two complementary off-the-shelf object recognizers producing predictions solely on labels. The first one uses a Mask R-CNN model [11] trained on the MS-COCO dataset [2, 18] which covers 91 common object categories (*e.g.*, `person`, `bicycle`,

---

[2]This number is a function of the Custom Vision service used to train new concepts.

ID	Age	Highest Degree	ML Exposure	Gender
E1	40-49	Bachelor	Low	Male
E2	30-39	Masters	Fair	Male
E3	30-39	Bachelor	Some	Female
E4	40-49	Bachelor	Fair	Male
E5	40-49	(undisclosed)	Fair	Male
E6	40-49	Masters	Some	Male
E7	30-39	Doctoral	Low	Female
E8	18-29	Bachelor	Low	Female

Figure 4. Background of participants in the evaluation study.

and `car`) and helped emulate a learner's "pre-existing knowledge". To support new teacher concepts we use Custom Vision [20], a cloud-based service for fast image classification and object recognition capable of producing predictions with as few as 15 labeled image samples. While Custom Vision has lower prediction performance, it allows teachers to define new concepts and fetch predictions within reasonable time frames for an interactive session. These choices influenced our interface and language design: while the Mask R-CNN model produces detailed polygonal masks, Custom Vision only supports bounding box masks. To ensure a uniform language vocabulary for both learner and teacher (**I**), we decided user labels to be bounding boxes. Similarly, relation labels were fed to Custom Vision as the union of bounding boxes for its encompassing concepts. Finally, we also used image search services (*Bing Images* [21] and *Visual Search* [22]) to collect a variety of domain relevant images from the web.

EVALUATION

We conducted a qualitative evaluation of our prototype as a *design probe* with 8 participants (E1-E8) with diverse backgrounds and ML experience (Fig. 4). Our goal was to validate our derived teaching language and design goals in a more realistic usage scenario.

Method

Sessions started with a **walkthrough** of the system (25 min), guiding users through a pre-loaded teaching scenario for the TC of `person playing tennis` and featuring a few examples of valid CDs. We then introduced users to the TC to be taught, `person riding bicycle`, and did a brief **foresight exercise** (similar to the one in the formative study, without the scripted learner responses) to help inform initial CDs. Participants then had 40 minutes to think aloud while teaching the TC from scratch (**teaching session**). We then loaded a "test" dataset containing a set of "challenging" images and gave them 10 minutes to experiment with them (**testing session**). We finished with an **interview** (10-20 min). Participants received a $50 gift card as appreciation for their time.

We used the same initial set of 56 image samples (in random order) during the teaching session. We collected top results from a search for *"person riding bicycle"* on Bing Images [21], which yielded a good mix of positive, negative, and edge cases. For the testing session, we manually curated an additional set of 29 complex edge cases to challenge teacher assumptions.

We sought to motivate participants to express rich knowledge they thought was important, even though the underlying off-the-shelf learners only leveraged labels. We asked participants

to (a) outline CDs to explain the TC and (b) provide enough labels for the target concept so that the system can make predictions, while (c) providing labels for new elements introduced in the concept descriptions if time allowed. They were encouraged to create CDs for things they found useful for the system to know, and were told that the system would "take everything they taught into account" but it was "ultimately responsible for its own learning".

We audio and screen recorded sessions, and analyzed their transcriptions via affinity diagramming; emerging themes guided our discussions under ***Findings***.

Findings

Participants summarized their overall teaching experience as "good" and "interesting" (4/8) but also "surprisingly challenging" (3/8) as they uncovered unexpected depth and nuance in the task of describing concepts: *"I was under the impression that a bicycle is a clear concept, but then I realized no"* (E7). We discuss this complexity under ***Refining Concepts***.

Participants were also able to follow the general teaching pattern of defining a TC plus accompanying CDs and labels from a pool of image samples. While some claimed needing time to *"get acclimated with the metaphor"* (3/8) many expressed appreciation, stating it *"all makes sense"* (E1), *"much more full-fledged than expected"* (E5) and *"overall the functionality seems perfect"* (E5). Still, we found that concepts and relationships were often not expressive enough for participants to cover things they found important for the system to know. We discuss limitations under ***Language Building Blocks***.

All participants completed the main tasks within the allotted time, defining an average of 6 CDs each (*min*=4, *max*=12) and providing enough labels to train the TC. However, 40 minutes was not enough to extensively label other CDs, and only 3 had the opportunity to test CDs beyond the TC. We further discuss findings on teacher-learner dialogue under ***Awareness and Feedback*** and choices of teaching images under ***Selecting Image Samples to Teach***.

Refining Concepts

The most prominent aspect of the teaching experience was how deeply participants thought about the TC and CDs. They realized it was an activity than required more thought and perhaps mental effort than initially expected. The first challenge was making one's own *implicit knowledge* about the TC *explicit*, which (for example) involved finding out what instances qualified as a person riding a bicycle or not and which of the relevant visual aspects to make that judgement should lead to potential CDs. A participant claimed that thinking in such a CD driven manner helped find the *essence* of the target concept: *"I think that I had a gut understanding of what a person riding a bicycle was, and then when I was faced with some of these edge cases (..) that felt like I was responding to a clearer and clearer picture of an initial gut instinct along the way, and the process of having to break it down into the concept descriptions seemed like it was a helpful way of clearing away some of the fluff"* (E8). And while not everybody reasoned under a CD mindset from the get-go, exposure to varied image samples eventually led to that direction, *e.g.*: *"When I looked*

at the images, I wasn't consciously trying to come up with 'what concepts can I come up with'. And as you go serially through the images, it's when you encounter an image that is a counter example of the target concept, that's when it triggers something in your head. So maybe looking for those things, and I think the tool actually provides you with the capability to do that" (E6). Externalizing that implicit understanding could still be too challenging at times, *e.g.*: *"This person is sitting. It's a scooter [with a seat] (..) and I said it was a bike. This [other scooter] is the exact same contraption, but without a seat, and somehow that's a difference to me. Don't ask me to explain that one. I can't"* (E2).

After identifying key concepts to explain, the next step was deciding how to best translate them into CDs. There were multiple ways of describing and decomposing these aspects, and determining which way was better was challenging. This involved considering the right level of detail, *e.g.*, `person on top of bicycle` vs. `person holding handlebars` (E5), and framing, *e.g.*, whether `foot on ground` should be a concept or a relationship (E6). Some aspects were difficult to express with just concepts and relationships; we discuss them under ***Language Building Blocks***. In the end, the greater challenge was figuring out which CDs *made sense*. While every participant had their own personal notion of what qualified as a sensible teaching concept, we found that navigating the image samples was helpful in determining what was *worth teaching*, *e.g.*, *"this probably will be the only instance of that that you'll see in a lot of this training. How valuable is it to do that kind of tagging?"* (E4). It also helped highlight what concepts *"proved to be valuable a lot of the time"* (E2) and to reassess the value of existing CDs, *e.g.*, *"in the back [I was] thinking about `person holding handlebar`. (..) I hadn't gotten to the point of realization that there will be counter examples of people with bicycles, but not actually riding the bicycle, [that] could be holding onto the handlebar as well"* (E6).

Language Building Blocks

As a whole, participants found that coming up with CDs was not overly complex, but not straightforward either: from a 1-5 scale (low to high complexity), it averaged 2.5 (median 2.75, mode 3, min 1, and max 3.5). Part of that perception stemmed from the challenges discussed above plus an initial unfamiliarity with the CD mindset (E3, E7), but several issues arose from limitations of the teaching language itself. Participants wanted to express configurations for concepts, *e.g.*, denoting that a `person` is `in seated position` instead of defining `seated position` as a concept (4/8). This behaviour underscores the limitations of a language with only concepts, relationships and modifiers as building blocks. Participants also struggled to describe background concepts that are difficult to box, such as `outdoors`, `greenery` and `road` (3/8) and resorted to defining parts of these concepts as workaround, *e.g.*, `tree` (E2) and `bike lane marking` (E8). In particular, participants also wanted to explain what the target concept was *"not"* and *"what to ignore"* (4/8), *e.g.*, *"unicycles are not bicycles"* (E1), and `person standing beside bicycle` is not a `person riding bicycle` (2/8).

Modifiers were not used as often as expected. Few participants used them consistently in their CDs (2/8), whereas others didn't use it (3/8) or used them only once (3/8). We expected they would want to weight certain concept descriptions against the TC overall, but instead, we saw a more prominent desire to qualify particular *instances* (3/8), *e.g.*, denoting a weaker "weight" for tandem bikes to convey that they are not canonical (E1). Some felt strongly enough about certain labels to also want to express what factors (CDs) played a role, *e.g.*, *"Because the person is not sitting on the bicycle, I'm going to use that information that sitting on the bicycle or not sitting on the bicycle is important to the distinction of this."* (E2).

Regarding image labeling, the overall experience was found generally *"easy to know what to do"* (E7), and participants quickly understood the notion of drawing boxes for concepts and confirming or removing predictions. Adding relations was not difficult to understand but some participants struggled with usability issues to visually differentiate the many potential connections to a concept (2/8) and tended to avoid adding relationships on busy images for the sake of time. But more importantly, participants took very naturally to having teaching labels and predictions coexisting in the same image, which speaks to the value of enabling a shared language between learner and teacher (**I**).

Awareness and Feedback

Available tools to assess learning included system predictions for learnables and the Learning panel. While participants often leveraged predictions to gauge how well the system was learning, most participants stated not using or forgetting about the Learning panel for the majority of the session (7/8). This was largely attributed to the short duration of the study and a focus on labeling (5/8) but many stated the information could have been helpful had they remembered to use it (4/8).

Prediction accuracy of the two ML components that `Pixie` used varied, leading to different perceptions of performance ranging from *"overtrained"*(E2) and *"often wrong"* (E4, E5) to *"overall good job"* (E3) and *"got it over time"* (E1). We highlight two important reactions (edited for brevity) on teaching. One was on the value of teaching and verifying custom concepts which increased confidence in the system: *"I really liked that I was able to do something like teach it `wheel`. And after 16 samples, it was showing wheels back to me. (..) I would expect with (..), better teaching, and (..) having this conversation with the system, (..) I have hope and faith that the system would start understanding my target concepts better"* (E5). The second one covers the perception that one's own teaching impacts how well the system learns: *"As I was training (..) I got that warm, fuzzy feeling that it was thinking about the concept in the way I was, and then with the test [session] that is just completely off base. So I didn't feel it was thinking about it, but I'm sure that had a lot to do with (..) my concept descriptions [being] kind of all over the map. The way I was thinking of the target concept itself, at least in terms of how I was describing it with the concept descriptions was getting a little muddled as well."* (E6). This reaction suggests that the teaching experience can evoke a sense of ownership and responsibility over the system's performance. Seeing this kind

of rationale even when the underlying ML components are just standard label-based models suggests that the design of the interactive experience was conducive to a teaching mindset.

We also asked for any missing feedback that could have improved the teaching experience. Suggestions ranged from (a) better awareness of what the system knows, *e.g.*, making the donut chart more prominent (E3, E5) or known/learnt concepts (E5), to (b) actual teaching support and making sense of what is *important* to teach, *e.g.*, *"It's hard to gauge the importance of what is left to learn"* (E2), *"should it have been vital that I look at these white boxes and say 'confirm those' even though they're not part of the target concept or scenario?"* (E1) and *"There's a bunch of these examples I would feel me removing false positives to be an important step, and I don't feel that, without you telling me, that the interface is really prompting me to do that"* (E5); and (c) the impact of teaching actions to particular predictions, *e.g.*, *"So if I said there aren't handlebars in an image and it decided handlebars were really really really important, and I removed this [handlebar label], would this [prediction for TC] go away?"* (E2).

Selecting Image Samples to Teach
Earlier we discussed the importance of image samples to guide overall reasoning about the TC and CDs. As for how participants chose samples during the teaching sessions, the predominant strategy was focusing on easy positive judgements (3/8) and less busy images (3/8) to more quickly get all TC labeling done within the 40 minutes. Other strategies included seeking diversity in the TC (3/8), seeking examples to train specific CDs (2/8), and visiting a few images from the teaching samples before defining the first CDs (2/8). We believe the latter strategies would be more common in time unconstrained scenarios. In addition, 2 participants expressed appreciation for the thumbnail preview that helped them to choose samples. While we did not assess scenarios without freedom to sample, the variety of image sample choices across participants suggests that teacher agency over the process is important to allow more control over what gets taught.

DISCUSSION

In this section, we briefly revisit our research questions and discuss further ramifications of our answers. We outline design lessons and recommendations in **bold** where relevant.

RQ1. *What are the core elements of a teaching language to allow expressing a wide range of concepts to a machine learner?*
The language we derived from our formative study proved to be expressive to describe a wide variety of concepts, but not all. Expanding the language building blocks is an important direction of future work, and we discuss potential next steps as informed by our research under ***Language Expressiveness***.

RQ2. *How does one build this language into an interactive experience supporting a person's teaching to a machine learner?*
We were able to translate our findings from our formative studies into a concrete teaching experience that showcased ways in which a system can help a machine teacher express knowledge through a language in support of an object detection task. We further elaborate on aspects of the teaching experience that are complementary to the teaching language, discussed under

Supporting Choice and Judgement of Image Samples and ***Visual Representation of Concepts***.

RQ3. *How do teachers understand and use this language in the context of a machine teaching experience?*
Participants understood and used the majority of the teaching language we devised. The extent of a participant's ML background did not seem to significantly affect teaching workflows, which alludes to the machine teaching tenet of leveraging "a human's inherent capacity to teach" [28]. A core aspect informed by our studies is the nuanced nature of this teaching, which calls for the inclusion of cognitive aids and for supporting awareness of the learner. While interested in the language's and usage's impact on object detection efficacy, we frame those considerations beyond the scope of this paper and leave them as the subject of future work. We discuss further recommendations under ***Language Expressiveness*** and ***Awareness and Interchange***.

Language Expressiveness
In order to strike a balance between breadth and depth of functionality, we chose to implement a simplified language supporting concepts and relationships between concepts. In our evaluation study, we found that teachers were able to convey a wide variety of concept and relationship descriptions related to the target concept using this simplified language, but still missed the ability to express notions that would have been more directly supported with the **addition of more language constructs**. Candidates for inclusion include: attributes, relationships beyond pairwise ones, exclusive classes (*e.g.*, `unicycle` + `bicycle` + `tricycle`), *"what to ignore"*, structural configurations (*e.g.*, "person standing in a particular way"), and a notion of foreground/background. Further work is needed to assess constructs on a wider range of object detection tasks. There is also an opportunity to explore the value of rich descriptor-based features for object detection to improve prediction accuracy [5] and explainability [3].

In both the formative study and the evaluation, we encouraged a foresight exercise in which the subjects explored aspects of the TC. While we did not assess how they would do without it, we believe this task had a positive effect in the evaluation study. Particularly, it kickstarted the process to ideate CDs, and was a process that "made sense" to participants. We thus recommend that teaching systems should **support some form of foresight exercise to guide teachers**.

While the foresight exercise is helpful in bootstrapping teachers, it may also reinforce initial imprecise conceptions around the TC. After exploring samples, participants often felt their original characterization required adjustments. While the prototype supported adding and deleting CDs, future work should seek to better **support modification, adaptation and evaluation of explanations after they've been created**. Teachers should feel encouraged to refine concepts to better reflect the various changes in their mental models.

A noteworthy finding was the reduced interest for additional information about the system capabilities in the evaluation compared to the formative study. We posit this was due to the abstract nature of the computational learner in the formative

study. With a concrete language and system, participants seemed more comfortable exploring system capabilities and working within the strictures of the language.

Awareness and Interchange

A positive emerging aspect of the teaching language was the ability to express the **learner's awareness regarding what the learner can see and understand** (*i.e.*, what it knows, how well it is learning). Our prototype aimed to give awareness to the teacher in a number of ways. First, during the creation of CDs, it would try to match the teachers concepts with the current list of known concepts and populated the Learning panel with all auxiliary concepts and their learning status. In the context of an image, the system would label known concepts, which helped teachers learn about concepts the system knows and better understand the quality of the object detection.

Prediction feedback also helped establish the notion of a **shared language**, in order to reduce potential misunderstanding between the computational learner and teacher. Having a shared language enabled the teacher to both understand what the computational learner sees in a picture as well as indicating where they could provide additional insight.

One observation is that the value of labeling concepts known to the system has temporal importance, and that priorities may change and evolve throughout the teaching process. While our evaluation study did not demonstrate the utility of the Learning panel, we underscore its importance during more extended usage. Future work should explore the design space of this type of feedback and evaluating its impact on the teaching process. This highlights the importance of **providing temporal awareness of the teaching journey and learner's progress**.

Our evaluation study also informed the challenge of providing guidance and awareness about what is "worth teaching". It would be helpful to explore how to **measure and provide awareness to the teacher of the impact of teaching actions**. We provided partial support via the "Test" feature (showing how well a learnable is being recognized), but features that assess the impact of individual CDs on predictions of the target concept could be particularly useful.

Supporting Choice and Judgement of Image Samples

We observed teaching to be a non-linear, non-trivial process. Our prototype let teachers control the flow of this process by switching between images. We also found that teachers used the system to revisit previously viewed images. Unlike many types of documents and objects to which people apply machine learning methods, thumbnail images provide a good "handle" by which users could select images to revisit or examine. We did not compare the experience for a larger numbers of images, but posit that the **language can serve as a tool to search and navigate the corpora of images**.

When confronted with images that challenged their understanding of the TC, participants often wanted to specify what factors from the list of CDs had more of an impact in judging an image. We see an opportunity to improve teacher support in navigating this inherent ambiguity and to allow for more nuanced judgements beyond "yes" or "no".

Visual Representation of Concepts

One important aspect of a language for object detection is how one expresses the visual localization of the concepts and relationships within an image. The choice of bounding boxes, free-form outlines or pixel-level demarcation has a significant impact on the image annotation language and the system leveraging the language. While our choice of bounding boxes as a language utterance allowed us to leverage existing computational learning systems, it also introduced challenges. There were instances of confusion when areas for two or more concepts overlapped too much, *e.g.*, individually labeling two cyclists sharing a tandem bike. On the other hand, the visual and structural simplicity of the boxes made the labeling process more intuitive and effortless. Future investigations are in order to explore and evaluate **flexible but simple ways of marking and labeling regions, entities, concepts and relationships in images**.

In the formative study, participants preferred tight outlines for labeling concepts, specified via an imprecise lasso. Current ML methods do not allow for imprecise localization information. Understanding the trade-offs between imprecise feedback and ML performance (effort vs. accuracy) may also be a relevant direction for future work.

CONCLUSIONS AND FUTURE WORK

This work aimed to characterize and assess an expressive language for teaching a computational learner how to detect objects in images. We applied human-centered methods to define language building blocks and infer design goals to guide the development of teaching experiences for object detection. We leverage these results to build `Pixie`, an end-to-end prototype that we used as a design probe to evaluate our teaching language and design goals. Our work showed how the use of the teaching language has the potential to enhance the creation of object detection models by supporting an interactive experience where teachers grow an understanding of a target concept through a knowledge-refining process and the ability to thoughtfully explore sets of possible samples to label.

One of the most exciting findings is that participants understood and were engaged with the teaching process, suggesting that teaching object detection this way is an enjoyable experience despite the higher cognitive tasks at play. On the other hand, our exploration was limited to characterizing teaching workflows for a single teacher. Past research in collaborative sensemaking and asynchronous hand-off [34] underscores the importance of shared knowledge representations between users. We posit that mediation between multiple teachers could leverage the very teaching language used for humans and machines, with adaptations to be assessed by future work.

Finally, we look forward to the creation of machine learning models that can operate in a compound manner, *i.e.*, leveraging prior knowledge on existing concepts to inform new ones. Beyond allowing for an end-to-end realization of our teaching language, they carry significant potential to better support explainability and model maintenance and reuse [28].

REFERENCES

[1] 2020. Language. In *The Cambridge Dictionary*. Cambridge University Press. `https://dictionary.cambridge.org/dictionary/english/language`

[2] Waleed Abdulla. 2017. Mask R-CNN for object detection and instance segmentation on Keras and TensorFlow. `https://github.com/matterport/Mask_RCNN`. (2017).

[3] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160. DOI:`http://dx.doi.org/10.1109/ACCESS.2018.2870052`

[4] Saleema Amershi, Maya Cakmak, William Bradley Knox, and Todd Kulesza. 2014. Power to the people: The role of humans in interactive machine learning. *AI Magazine* 35, 4 (2014), 105–120. DOI:`http://dx.doi.org/10.1609/aimag.v35i4.2513`

[5] Steve Branson, Catherine Wah, Boris Babenko, Florian Schroff, Peter Welinder, Pietro Perona, and Serge Belongie. 2010. Visual Recognition with Humans in the Loop. In *European Conference on Computer Vision (ECCV)* (2010-01-01). Heraklion, Crete. DOI:`http://dx.doi.org/10.1007/978-3-642-15561-1_32`

[6] Anind K Dey, Stephanie Rosenthal, and Manuela Veloso. 2009. Using interaction to improve intelligence: how intelligent systems should ask users for input. In *Workshop on Intelligence and Interaction: IJCAI*.

[7] John J Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 8. DOI:`http://dx.doi.org/10.1145/3185517`

[8] Jerry Alan Fails and Dan R. Olsen, Jr. 2003. Interactive Machine Learning. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI '03)*. ACM, New York, NY, USA, 39–45. DOI:`http://dx.doi.org/10.1145/604045.604056`

[9] Rebecca Fiebrink and Perry R. Cook. 2010. The wekinator: a system for real-time, interactive machine learning in music. In *Proceedings of The Eleventh International Society for Music Information Retrieval Conference (ISMIR)*.

[10] James Fogarty, Desney Tan, Ashish Kapoor, and Simon Winder. 2008. CueFlik: interactive concept learning in image search. In *Proceedings of the sigchi conference on human factors in computing systems*. ACM, 29–38. DOI:`http://dx.doi.org/10.1145/1357054.1357061`

[11] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. 2017. Mask R-CNN. In *Proceedings of the IEEE international conference on computer vision*. 2961–2969. DOI:`http://dx.doi.org/10.1109/ICCV.2017.322`

[12] Andreas Holzinger, Markus Plass, Michael Kickmeier-Rust, Katharina Holzinger, Gloria Cerasela Crişan, Camelia-M. Pintea, and Vasile Palade. 2019. Interactive machine learning: experimental evidence for the human in the algorithmic loop. *Applied Intelligence* 49, 7 (01 Jul 2019), 2401–2414. DOI:`http://dx.doi.org/10.1007/s10489-018-1361-5`

[13] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2901–2910. DOI:`http://dx.doi.org/10.1109/CVPR.2017.215`

[14] Tim Kraska. 2018. Northstar: An Interactive Data Science System. *Proc. VLDB Endow.* 11, 12 (Aug. 2018), 2150–2164. DOI:`http://dx.doi.org/10.14778/3229863.3240493`

[15] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. `https://arxiv.org/abs/1602.07332`

[16] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell me more?: the effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 1–10. DOI:`http://dx.doi.org/10.1145/2207676.2207678`

[17] Todd Kulesza, Simone Stumpf, Margaret Burnett, Weng-Keen Wong, Yann Riche, Travis Moore, Ian Oberst, Amber Shinsel, and Kevin McIntosh. 2010. Explanatory debugging: Supporting end-user debugging of machine-learned programs. In *2010 IEEE Symposium on Visual Languages and Human-Centric Computing*. IEEE, 41–48. DOI:`http://dx.doi.org/10.1109/VLHCC.2010.15`

[18] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *European conference on computer vision*. Springer, 740–755. DOI:`http://dx.doi.org/10.1007/978-3-319-10602-1_48`

[19] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.). Curran Associates, Inc., 4765–4774.

[20] Microsoft. 2019a. What is Azure Custom Vision? `https://docs.microsoft.com/en-us/azure/cognitive-services/custom-vision-service/home`. (2019).

[21] Microsoft. 2019b. What is the Bing Image Search API? `https://docs.microsoft.com/en-gb/azure/cognitive-services/bing-image-search/overview`. (2019).

[22] Microsoft. 2019c. What is the Bing Visual Search API? `https://docs.microsoft.com/en-us/azure/cognitive-services/bing-visual-search/overview`. (2019).

[23] Maxime Oquab, Leon Bottou, Ivan Laptev, and Josef Sivic. 2014. Learning and Transferring Mid-Level Image Representations using Convolutional Neural Networks. In *Proc. Computer Vision and Pattern Recognition*. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). DOI: `http://dx.doi.org/10.1109/CVPR.2014.222`

[24] Joseph Redmon and Ali Farhadi. 2018. YOLOv3: An Incremental Improvement. *arXiv* (2018). `https://arxiv.org/abs/1804.02767`

[25] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*. 1135–1144. DOI: `http://dx.doi.org/10.1145/2939672.2939778`

[26] Aruni RoyChowdhury, Prithvijit Chakrabarty, Ashish Singh, SouYoung Jin, Huaizu Jiang, Liangliang Cao, and Erik Learned-Miller. 2019. Automatic adaptation of object detectors to new domains using self-training. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. DOI: `http://dx.doi.org/10.1109/CVPR.2019.00087`

[27] Burr Settles. 2012. *Active Learning*. Morgan & Claypool Publishers. DOI: `http://dx.doi.org/10.2200/S00429ED1V01Y201207AIM018`

[28] Patrice Y Simard, Saleema Amershi, David M Chickering, Alicia Edelman Pelton, Soroush Ghorashi, Christopher Meek, Gonzalo Ramos, Jina Suh, Johan Verwey, Mo Wang, and others. 2017. Machine Teaching: A New Paradigm for Building Machine Learning

Systems. *arXiv preprint arXiv:1707.06742* (2017). `https://arxiv.org/abs/1707.06742`

[29] Simone Stumpf, Vidya Rajaram, Lida Li, Margaret Burnett, Thomas Dietterich, Erin Sullivan, Russell Drummond, and Jonathan Herlocker. 2007. Toward harnessing user feedback for machine learning. In *Proceedings of the 12th international conference on Intelligent user interfaces*. ACM, 82–91. DOI: `http://dx.doi.org/10.1145/1216295.1216316`

[30] Emily Wall, Soroush Ghorashi, and Gonzalo Ramos. 2019. Using Expert Patterns in Assisted Interactive Machine Learning: A Study in Machine Teaching. In *Proceedings of the 17th IFIP TC 13 International Conference on Human-Computer Interaction*. Springer International Publishing. DOI: `http://dx.doi.org/10.1007/978-3-030-29387-1_34`

[31] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. 2014. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems 27*. Curran Associates, Inc., 3320–3328.

[32] Jason Yosinski, Jeff Clune, Anh Nguyen, Thomas Fuchs, and Hod Lipson. 2015. Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579* (2015). `https://arxiv.org/abs/1506.06579`

[33] Matthew D Zeiler and Rob Fergus. 2014. Visualizing and understanding convolutional networks. In *European conference on computer vision*. Springer, 818–833. DOI: `http://dx.doi.org/10.1007/978-3-319-10590-1_53`

[34] Jian Zhao, Michael Glueck, Petra Isenberg, Fanny Chevalier, and Azam Khan. 2017. Supporting handoff in asynchronous collaborative sensemaking using knowledge-transfer graphs. *IEEE transactions on visualization and computer graphics* 24, 1 (2017), 340–350.

[35] X. Zhu. 2015. Machine Teaching: an Inverse Problem to Machine Learning and an Approach Toward Optimal Education. In *The Twenty-Ninth AAAI Conference on Artificial Intelligence*.