Artificial Intelligence Review (2023) 56:3005–3054
https://doi.org/10.1007/s10462-022-10246-w

Human-in-the-loop machine learning: a state of the art

Eduardo Mosqueira-Rey[1] · Elena Hernández-Pereira[1] · David Alonso-Ríos[1] · José Bobes-Bascarán[1] · Ángel Fernández-Leal[1]

Published online: 17 August 2022

Abstract
Researchers are defining new types of interactions between humans and machine learning algorithms generically called human-in-the-loop machine learning. Depending on who is in control of the learning process, we can identify: active learning, in which the system remains in control; interactive machine learning, in which there is a closer interaction between users and learning systems; and machine teaching, where human domain experts have control over the learning process. Aside from control, humans can also be involved in the learning process in other ways. In curriculum learning human domain experts try to impose some structure on the examples presented to improve the learning; in explainable AI the focus is on the ability of the model to explain to humans why a given solution was chosen. This collaboration between AI models and humans should not be limited only to the learning process; if we go further, we can see other terms that arise such as Usable and Useful AI. In this paper we review the state of the art of the techniques involved in the new forms of relationship between humans and ML algorithms. Our contribution is not merely listing the different approaches, but to provide definitions clarifying confusing, varied and sometimes contradictory terms; to elucidate and determine the boundaries between the different methods; and to correlate all the techniques searching for the connections and influences between them.

Keywords Human-in-the-loop machine learning · Active learning · Interactive machine learning · Machine teaching · Curriculum learning · Explainable AI

✉ Eduardo Mosqueira-Rey
eduardo@udc.es

Elena Hernández-Pereira
elena.hernandez@udc.es

David Alonso-Ríos
david.alonso@udc.es

José Bobes-Bascarán
jose.bobes@udc.es

Ángel Fernández-Leal
angel.fleal@udc.es

[1] Department of Computer Science and Information Technologies, Universidade da Coruña (CITIC), Campus de Elviña, 15071 A Coruña, Spain

1 Introduction

There is currently a great demand for machine learning (ML) solutions. This is because the advances that have occurred in recent years around this technology have popularized it and have brought it closer to the general public. But building machine learning systems is a complex process that requires deep knowledge of machine learning techniques.

Usually, humans are required at various points in the loop of the machine learning process but following a kind of monolithic conception in which the machine learning algorithm is modeled, built, tested and then offered to the public without further changes.

Models that are developed under this scenario might run the risk of not scaling well, becoming static, being hard to evaluate, and degrading their performance due to changes in the context they are deployed into. Also, due to the limitations of the dominating connectionist approach, they usually lack logical reasoning and the possibility of identifying causal relations (Holmberg et al. 2020).

Researchers are defining new types of interactions between humans and machine learning algorithms, which we can group under the umbrella term of *Human-in-the-loop machine learning* (HITL-ML) (Munro 2020). The idea is not only to make machine learning more accurate or to obtain the desired accuracy faster, but also to make humans more effective and more efficient.

Depending on who is in control of the learning process, we can identify different approaches to HITL-ML (Holmberg et al. 2020):

- **Active learning (AL)** (Settles 2009), in which the system remains in control of the learning process and treats humans as oracles to annotate unlabeled data.
- **Interactive machine learning (IML)** (Amershi et al. 2014), in which there is a closer interaction between users and learning systems, with people interactively supplying information in a more focused, frequent, and incremental way compared to traditional machine learning.
- **Machine teaching (MT)** (Simard et al. 2017; Ramos et al. 2020), where human domain experts have control over the learning process by delimiting the knowledge that they intend to transfer to the machine learning model.

Aside from control, humans can also be involved in the learning process in other ways. For example, human learning has inspired different algorithms designs throughout the development of machine learning. As an outstanding feature of human learning, curriculum, or learning in a meaningful order, has been exploited and transferred to machine learning, which forms the subdiscipline named **curriculum learning (CL)** (Bengio et al. 2009). This idea is focused on trying to impose some structure on the training set to accelerate and improve the learning and it constitutes another approach to HITL-ML.

Also, we must also bear in mind that, in certain domains, it is advisable that the algorithms should explain their results to humans. We are not only interested in the ability of an algorithm to solve a problem with a given accuracy, but also in the ability to explain why a given solution was chosen. This is called **Explainable AI (XAI)** (Adadi and Berrada 2018) and is a research field that aims to make the results of AI systems more understandable to humans. Currently, it has been noted that the humans' role has not been sufficiently studied in existing approaches to explainability (Abdul et al. 2018).

Finally, we have to take into account not only the development of AI or ML models but also the design of the interactions and behaviors that compose the human experience

around the AI models (van Allen 2018). This leads us to the definition of two new terms related to the relationship between humans and AI models that go beyond cooperation in learning, called "**Usable AI**" and "**Useful AI**" (Xu 2019), and which are fundamental to ensure that an AI model is successful. Usable AI focuses on ensuring that AI systems are usable by the people interacting with them. Useful AI goes further and tries to make AI models useful in a broad sense, i.e., useful to the society in which they are embedded, approaching AI from a human perspective by considering human conditions and contexts.

In this paper we review the state of the art of the techniques involved in the new forms of relationship between humans and ML algorithms focusing mainly in the different strategies on how to incorporate humans into the learning process. The contribution of this work is not merely to list papers within each discipline, but to provide definitions of each term and to clarify confusing, varied and sometimes contradictory definitions—e.g., the term Machine Teaching has been used in the literature to define very different and sometimes unrelated techniques—. It is also intended to clarify and determine the boundaries between the different approaches, which are not always very clear—e.g., IML comes from AL adding new levels of interactivity, but this can become confused or mixed with interactive MT—. An attempt is also made to correlate all the techniques with each other and to see the relationships and influences that they have between them. This aspect will be commented in each section and will be discussed in the final section of discussion and conclusions.

The field of Human-in-the-Loop ML is quite broad, so going into detail on each of the techniques would be unfeasible. For this reason, in each of them we focus on giving historical perspectives, offering definitions, describing the main methods involved and their applications. Within each section, reference is also made to reviews of these techniques that exist in the literature, to offer a starting point for those who want to learn more details about them. Thus, for example, (Dudley and Kristensson 2018) review aspects related to the user interface in IML, (Ramos et al. 2020) reviews the human aspects involved in interactive MT. Sometimes these reviews offer perspectives that do not always coincide as occurs with XAI goals which are classified differently in recent review papers (Barredo Arrieta et al. 2020; Minh et al. 2021; Meske et al. 2022; Das and Rad 2020).

In Fig. 1 we can see a mind map with the structure of the paper, and in each of its sections we will introduce a mind map that helps to summarize the contents exposed within that particular section.

Thus, the paper is structured as follows: First, we begin with an explanation of the different types of learning with human collaboration: active learning (AL)—Sect. 2—,

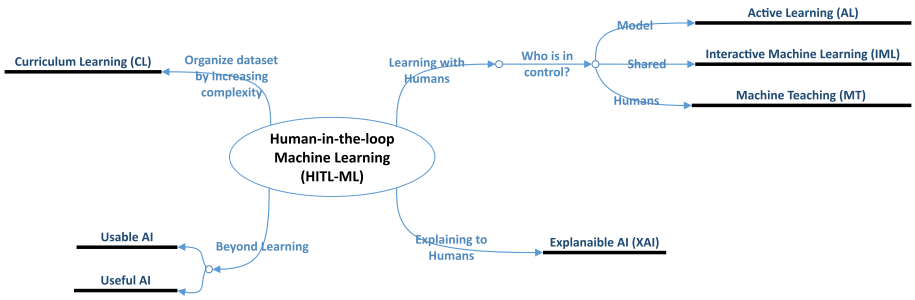

Fig. 1 Human-in-the-loop machine learning (HITL-ML) mind map

interactive machine learning (IML)—Sect. 3—and Machine Teaching (MT)—Sect. 4—. This will be followed by a discussion on curriculum learning (CL)—Sect. 5—since it is a technique that will be used, to a greater or lesser extent, in the techniques commented. The following section is devoted to describing the process where ML models explain their results to humans through Explainable AI (XAI)—Sect. 6—. We will then describe more briefly the relationship between humans and AI and ML systems in terms that go beyond learning and defining usable AI and useful AI—Sect. 7—. Finally, we will end the paper with a chapter for discussion and conclusions—Sect. 8— in which we highlight the relationships between the different techniques and discuss trends and future developments.

2 Active learning (AL)

When we described the human-in-the-loop machine learning (HITL-ML) approach in the introduction we mentioned that the inclusion of humans in the learning process could be done at different levels depending on who was in control of the process. The first of these levels is active learning (AL) in which the system remains in control and uses humans as oracles to annotate data. In this section we will describe the different definitions of AL, the process to update a ML model following this approach commenting on the different strategies that can be taken. We end up with a brief review of some AL applications, the different issues that can emerge when applying AL, and how this technique is connected with other ML techniques. A mind map of AL can be consulted in Fig. 2.

2.1 AL definitions

AL is a machine learning approach in which a learner requests an oracle (who acts as a teacher) to label selected examples that are not clear and that will provide relevant information to the learning process. As a result, the learner improves its learning performance

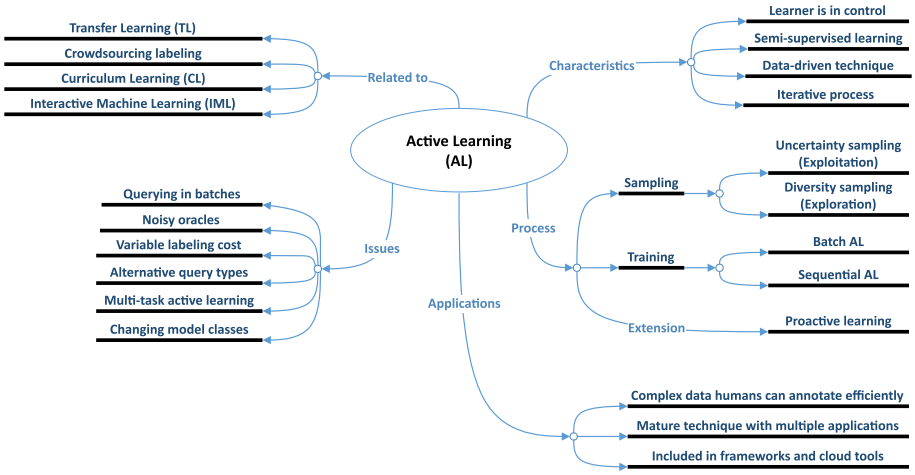

Fig. 2 Active learning (AL) mind map

using less training examples. It is very effective in settings where there is a lot of unlabeled data available, but the annotation task is expensive or time-consuming.

In this technique, the **learner is in control** of the data, and queries an entity with extensive knowledge of the domain (typically a human expert) for annotating unlabeled examples.

Therefore, AL is a kind of **semi-supervised learning** as it uses both labeled and unlabeled data. New examples get annotated in an iterative and incremental process, where a query strategy is used to choose an example to be queried, and once labeled by an oracle, will result in a model accuracy increment.

It was inspired by the family of instructional techniques with the same name in the education literature (Bonwell and Eison 1991) whose intention is to make the student a partner in the learning process and thus not being overly dependent on the teacher. The source of knowledge could be a set of positives examples and/or an oracle as proposed by Valiant (1984) and Angluin (1987). While the former focuses on the knowledge acquisition issue, the latter describes some alternatives on query construction. One of the first applications of this technique to machine learning can be found in Sammut and Banerji (1986) in which AL is used to enable the learner to take an active role in acquiring the new concepts.

For further reading, we refer to Settles (2009) and Olsson (2009) for detailed active learning literature surveys.

2.2 AL process

Active learning uses an **iterative process** for obtaining training data, unlike passive learning, where all labeled data is provided in advance. It is said that the learner is curious and requests information from the oracle based on different query strategies.

AL is a **data-driven technique** as it relies on the data to get the highest performance. The acquisition of unlabeled examples is less expensive than the labeled ones. By using a mechanism that helps selecting the most relevant examples, the system reduces the amount of data required to train the model, while maximizing (at least maintaining) its accuracy, at a lower cost (Settles 2011).

The AL process is as follows, the data set is divided into two groups of examples: the ones that are labeled, and the ones that are still unlabeled (i.e., the label is unknown). The model iteratively select a new example (or set of examples) from the unlabeled group and provide it to the oracle so that it gets labeled. The system then trains the model using the new data until the desired performance or a stop condition is achieved (see Fig. 3).

Here it is important to distinguish several processes that occur within active learning and that are sometimes confused. The first is the **sampling process** also known as the **query strategy**. This process consists of selecting those instances to be labeled by the human expert.

Angluin (1987) was one of the very first authors that cite some alternatives on query construction. Another notable work is Cohn et al. (1994) that described an example-based approach called *selective sampling*, as a rudimentary form of active learning that is suited for concept-learning problems. The samples are selected sequentially from a region of uncertainty, which is the area of the domain where misclassification is still possible. For each example, the region of uncertainty is recalculated, and new examples are picked from that region. With this approach, as more examples are added to the model, the uncertainty decreases without decreasing the efficiency.

In this regard, Settles (2009) mentioned three main sampling strategies:

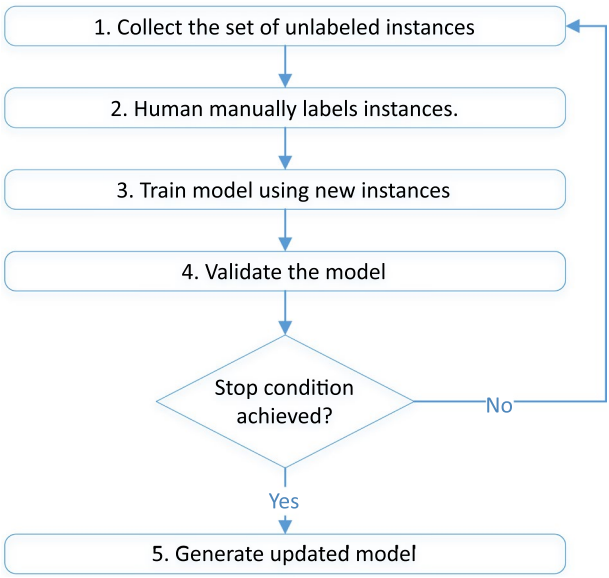

Fig. 3 Steps taken in order to update the model in AL

- **Membership query synthesis**: The learner may request labels for any unlabeled instance in the input space, including queries that the learner generates de novo.
- **Stream-based selective sampling**: Also called sequential sampling, in which each unlabeled instance is drawn one at a time from the data source, and the learner must decide whether to query or discard it.
- **Pool-based sampling**: The entire collection of data (or a subset of it) is evaluated and ranked in order to select the best element to annotate.

More recently, Munro (2020) distinguishes three types of sampling strategies: random, uncertainty and diversity. The random query strategy is the simplest as the data to be labeled are randomly selected. The other two strategies are more interesting since they describe a well-known dilemma: exploitation vs. exploration (Hills et al. 2015).

- **Uncertainty sampling (Exploitation)**: It selects instances which have the least label certainty under the current trained model. In this category we found:
  - **Least confidence**, which takes the example with the lowest confidence in their most likely class label.
  - **Margin of confidence**, that uses the smallest difference between the top two highest probabilities for each possible label.
  - **Ratio of confidence**, which uses the ratio between the top two most confident predictions.
  - **Entropy**, that uses the difference between all predictions.

- **Diversity sampling (Exploration)**: It selects unlabeled items that are rare or unseen in the training data to increase the picture of the problem space. Here, we found:

- **Model-based outliers**, that samples for low activation (e.g. hidden layers).
- **Cluster-based sampling**, which uses unsupervised learning to cluster the data to find outliers that are not part of any trend.
- **Representative sampling**, that finds items most representative of the target domain.
- **Real-world diversity**, which increases fairness with data supporting real-world diversity.

As we can see, while exploitation focuses on improving the efficiency using existing products or data, exploration goes beyond the known data samples to enhance the diversity of the data. The latter is relevant if we want the model to generalize (and we want it).

Other aspect related with the AL process is **how many new instances are labeled before training again the model**. In this regard Rubens et al. (2015) identify two main approaches:

- **Batch**: several examples get labeled until the model is trained again.
- **Sequential**: the system is retrained after each new element is labeled given immediate feedback to the user.

Here we can identify a trade-off between the two alternatives. Sequential training is important, for example, when working with recommender systems, since users expect to get an updated list of recommendations based on their last annotation, on the other hand, allowing the user to rate several items, or several features of an item before readjusting the model is more efficient both computationally and in terms of the cost associated with the interactions with humans.

Finally, the AL process we have described here is considered ideal: "The oracle is assumed to be infallible (never wrong), indefatigable (always answers), individual (only one oracle), and insensitive to costs (always free or always charges the same)". Some authors like Donmez and Carbonell (2008) proposed an extension to the AL concept, called **Proactive Learning**, which is a generalization that relaxes several assumptions of the process seeking to cover a more realistic scenario. In this case, the oracle is probabilistic (may err), reluctant (may refuse to answer), plural (different oracles), and the costs are variable (per oracle, per instance).

2.3 Applications of AL

As a general rule, the fields of application of AL are those where **the cost of annotating data is high, but these are tasks that humans generally do well**, such as interpreting images or processing natural language.

AL is of special interest when the labeling example process is expensive or time-consuming, and it also applies on the scenario of a scarce number of examples (e.g., rare diseases). The active learner aims to achieve high accuracy using as few labeled instances as possible, thereby minimizing the cost of obtaining labeled data. As a result, the learner improves its learning performance (i.e., maximizing accuracy).

Regarding image and video classification, in the medical context we can find AL in the classification of radiology reports (Hoi et al. 2006; Nguyen and Patrick 2014). In biological research, it has been used in recognizing multiple types of plankton using images (Luo et al. 2004). An extensive survey in medical image analysis, which serves as the basis for clinical decision making, is performed in Budd et al. (2021).

Regarding Natural Language Processing (NLP) we refer to the survey of Olsson (2009) as a starting reference. A modern reference is De Angeli et al. (2021) that uses AL for the classification of cancer pathology reports.

As we can see, **AL is a mature technique** and due to its versatility it has been applied in a diverse number of settings. For example, we could find it at the basis of recommender systems (Rubens et al. 2015), for the construction of a reward estimation models (Lopes et al. 2009), and as the base for predicting molecular energetics (Smith et al. 2018) or predicting a heart disease (El-Hasnony et al. 2022) just to mention some of them.

In response to these multiple applications, AL has been included in software programming frameworks so that it can be integrated in custom developments (Jamieson et al. 2015; Reyes et al. 2016; Tang et al. 2019). AL can also be found incorporated in many cloud tools such as QnA Maker (Microsoft 2022), a cloud-based Natural Language Processing (NLP) service by Microsoft or into the Appen platform used for AI data sourcing, data annotation and model evaluation by humans.

2.4 AL shortcomings

Even if AL has produced good results in many scenarios, it is not free of issues, as some base assumptions about this technique do not always hold. Some of these issues or limitations have been described by Settles (2011) and Donmez and Carbonell (2008) and we quote them below.

- **Querying in batches**. It is often assumed a pool-based scenario in which the learner would select instances to be queried one at a time, re-train the model and using the new generated model, repeat the process. As the training process is usually expensive and it is not feasible to re-train for every single instance queried, batches of instances are selected allowing less training steps and making the process more efficient.
- **Noisy oracles**. In most experiments it is assumed that the quality of labeled data from the oracle is high. Even if a human acts as the oracle, some instances are implicitly difficult, both for models and humans. Furthermore, humans can become distracted or fatigued over time, which introduces variability in the quality of their annotations. The use of multiple non-experts could overcome this issue, but still there are some decisions to be made as how to decide an oracle label is trustworthy, when to query several oracles or when to get new queries for a noisy example.
- **Variable labeling cost**. The cost of obtaining new labels has been assumed uniform and fixed. Moreover, the cost of misclassification has been ignored in many experiments. A cost-sensitive framework based on meta-features can be set to include the variability in the model.
- **Alternative query types**. Refers to the assumption that the query unit is always the same type as the target concept to be learned. The usual approach of membership query is used in many systems, but some other types of queries can be considered as multiple-instance or querying features.
- **Multi-task active learning**. It is assumed that there is only one learner trying to solve a single task, but sometimes the same data could be labeled for various sub-tasks at the same time. When using classification of not mutually exclusive categories, a learner can decide to query for several of them at a time.
- **Changing model classes**. In some scenarios the model does not contain a representative set of examples of a real problem, and the unknown data remains unexplored.

When re-using a model in a different problem it can become problematic. It is also the case when new knowledge is available in the future and there is a need to incorporate this information into an existing model. The system should be prepared so that it can incorporate new classes, or change the existing ones.

2.5 AL related techniques

We could also find AL **combined with other techniques** to produce better global results. For example, **Transfer Learning (TL)** (Zhuang et al. 2021) is a technique used to create high-performance learners trained with data that is easily obtained and then transfer their knowledge to solve real-world machine learning scenarios in which training data is expensive or difficult to collect (Weiss et al. 2016). TL can be used as an alternative to AL to overcome the lack of training examples (Aggarwal et al. 2014), but it can also be used in conjunction with AL to avoid a cold start in the creation of the model. Nevertheless, the use of TL does not change the general process of AL, since pretrained models still require additional human labels to achieve accurate results in their tasks. However, a substantial head start in labeling can influence the choice of active learning strategy to use (Munro 2020). Thus, in general, TL is considered the basis of some of the most advanced active learning strategies proposed.

AL can be also related with **crowdsourcing labeling** services such as the Amazon Mechanical Turk (Amazon 2022). Crowdsourcing services offer the acquisition of non-expert annotations at low cost outsourcing small annotation tasks to a large group of freelance workers. A consequence of using non-expert annotators is a lower annotation quality that requires of quality control strategies. AL and crowdsourcing are complementary approaches: AL reduces the number of annotations used while crowdsourcing reduces the cost per annotation. Combined, the two approaches could substantially lower the cost of creating training sets (Laws et al. 2011; Zhao et al. 2020).

We can also relate AL with curriculum learning (CL) (see Sect. 5). Humans do learn better if the examples used to train them are sorted and organized so that they get gradually more complex. This process of creating an ordered sequence of examples to be provided to the learner at different stages of the learning process can help improving the learner performance. When using CL on an AL approach, instead of taking the examples near the decision surface, the focus should be on choosing the examples that the learner could potentially label, and gradually add new examples near the decision border.

Finally, we can connect AL with the technique described in the following section, **interactive machine learning (IML)**. As we will see, IML is a generalization of an active learning approach in which the control is shared between humans and learning models.

3 Interactive machine learning (IML)

The following level inside human-in-the-loop machine learning (HITL-ML) regarding who is in control of the learning is interactive machine learning (IML) in which there is a closer interaction between users and learning systems, with people interactively supplying information in a more focused, frequent, and incremental way compared to traditional machine learning. In this section we will describe the different definitions of IML, its differences with AL and several applications in which IML was employed. A mind map of IML can be consulted in Fig. 4.

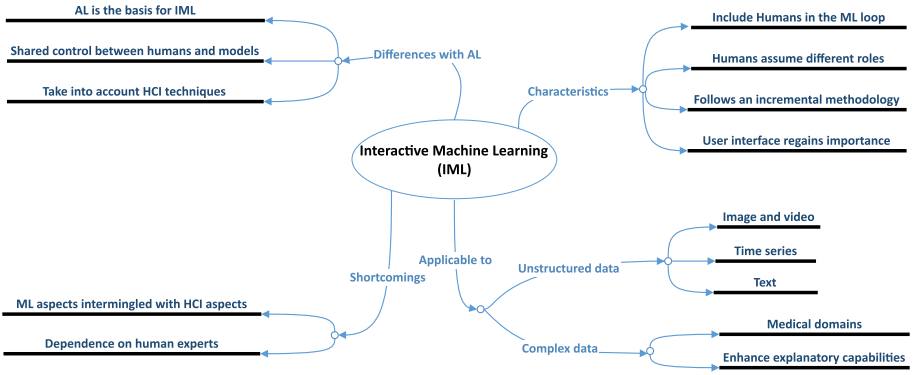

Fig. 4 Interactive machine learning (IML) mind map

3.1 IML definitions

The paper that first describes interactive machine learning (IML) is the work of Ware et al. (2001). These authors describe IML as a method for interactively constructing decision tree classifiers. Although their definition of IML is more restricted than the current notion of the term, they highlight the basic idea that is behind most IML approaches: letting users (experts and non-experts) build classifiers.

Following this initial approach to IML we can cite the work of Talbot et al. (2009) in which the authors built a system called *EnsembleMatrix*, that was used to visualize the confusion matrices of various classifiers and allow users to interact with these visualizations to better understand the models and to set up model combination strategies to obtain better results. Some of these authors (Kapoor et al. 2010) continued on this path with, *ManiMatrix*, an interactive system that allowed interactive refinement of classification boundaries in a multiclass setting.

Nevertheless, the work highly cited as seminal paper of IML is Fails and Olsen (2003) where the human designer trains, corrects and teaches interactively the model until desired results are met. Here, authors do not only pose the question of introducing humans into the machine learning loop, but they also contrast the concept of IML with the so called *Classical Machine-Learning* (CML) concept (that in previous sections we have defined as *passive learning*), where training is performed off-line, trying to optimize learning at the expense of longer training times.

The idea of including humans in the loop, thus changing the working methodology, continues in Porter et al. (2013). This work proposes that humans and computers should work together on the same task doing what each of them does best at any specific moment. There are different methodologies according to the position humans have within the workflow. On the one hand, humans can go to the end of the flow, correcting the results of a machine learning system—e.g., using humans to validate, clean and correct the results. On the other hand, humans can be used first, performing identification and annotation tasks that are simple for them but complicated for machines—e.g., interactive image segmentation, in which humans provide input with basic annotation tools. Authors go further and define as a promise of IML to have systems where this dialogue between machines and humans is more dynamic and optimized to the abilities of each one, in the same way that crowdsourcing is enabling humans to be cost-effective in tasks traditionally performed by machines.

Following these works, we can find others that provide more up-to-date definitions of IML. For example, Holzinger (2016) defined IML as "algorithms that can interact with agents and can optimize their learning behavior through these interactions, where the agents can also be human". Although the definition seems to downplay the fact that the external agents may be human, in his paper, Holzinger defends the use of "*human-in-the-loop*" for complex domains such as health informatics, where biomedical data sets are full of uncertainty and incompleteness, and the problems to solve are hard.

The relationship between humans and ML models in IML is also highlighted by Jiang et al. (2019) stating that IML is "an iterative learning process that tightly couples a human with a machine learner" or Ramos et al. (2020) defining IML as "the process in which a person (or people) engages with a learning algorithm, in an interactive loop, to generate useful artifacts". Later, the authors describe these artifacts as data, insights about data, or machine-learned models.

Ramos et al. (2020) also comment that the roles that humans can play in IML can be different, as they can be: ML experts, data scientists, crowdsource workers or domain experts. These different roles affect the form and function of the IML systems.

In this regard, Yang et al. (2018) include into IML people who are not formally trained in ML, that is, non-experts, and focus the research field in the development of tools that allow these non-experts to actively build ML solutions to serve their needs in the real world.

Jiang et al. (2019) also included another aspect that is important in IML: how the model is updated. They stated that the process should be iterative. In this way, Amershi et al. (2014) focus the difference between traditional ML and IML in how the model is updated. In IML the updates are faster (as an immediate response to user input), focused (centered in a particular aspect of the model) and incremental (the model is changed continually, with small updates). Dudley and Kristensson (2018) also emphasize the iterative part in their definition of IML as "an interaction paradigm in which a user or user group iteratively builds and refines a mathematical model to describe a concept through iterative cycles of input and review".

Finally, Dudley and Kristensson (2018) described the fundamental parts of an IML system as: users, model, data and interface. The first three were already present in classical ML systems, although here the users' role can be different and can include users with no deep understanding of ML techniques. But what is new here is the interface part. A classic ML system must have an interface, but it is a passive one; in IML the interface is responsible for the bidirectional feedback between the other three components and for the authors, the interface design is critical to the success of the IML process.

From all of these definitions we can extract the main features that underpin an IML system:

- **Humans in the ML loop**. They have been assigned to tasks at which they are more efficient than machines.
- **Humans assuming different roles**. They can be domain experts, non-expert users, data scientists, etc.
- **Incremental methodology**. The model is updated iteratively and incrementally.
- **The importance of the user interface**. It influences how learning takes place and conditions learning outcomes.

3.2 Differences with active learning (AL)

But, what is the difference between active learning and interactive machine learning? We have seen in previous sections that AL differs from classical or passive learning in that it is an interactive process in which the model poses queries, usually in the form of unlabeled data instances that have to be labeled by an oracle that, normally, is a human annotator. Since there is an interactive process inside AL, it can be confused with IML which is a similar technique but with its own characteristics.

Holzinger (2016) includes AL as one of the three pillars that form the basis of IML—the other two are Reinforcement Learning and Preference Learning—and proposes to use IML in fields in which there are insufficient training samples following an "expert-in-the-loop" approach.

According to Amershi et al. (2014): "Although active learning results in faster convergence, users get frustrated from having to answer the learner's long stream of questions and not having control over the interaction". Amershi cited several studies which reveal that users do not like to behave like simple oracles, and that human factors, such as interruptibility or frustration, should be taken into account in active learning techniques.

For Dudley and Kristensson (2018) both AL and IML focus on selecting new points for labeling by the user, but the key distinction is that in AL the selection is driven by the model and in IML the selection is driven by the user.

This leads to another difference between AL and IML systems: how to evaluate them. Fiebrink et al. (2011) already stated that in an IML system the evaluation of the models should go beyond their accuracy and should include subjective judgments of properties such as cost, confidence, complexity, etc. AL focuses on building better models in an *algorithm-centered* evaluation, but in IML systems we have to take into account human factors, so there is also a *human-centered* evaluation, focusing on the utility and effectiveness of the application for end-users. Boukhelifa et al. (2018) stated that coupling both evaluations in IML systems can bring forth insights that can play an important role in addressing the "black-box" effect of machine learning algorithms.

In Fig. 5 we can see a schematic representation of the IML process, we can see that it is a freer scheme than the one of AL represented in Fig. 3. In this diagram interactivity becomes more important and there are different tasks that the human can do interactively (such as selecting examples, labeling cases, etc.) depending on which is more appropriate at each step (Suh et al. 2019).

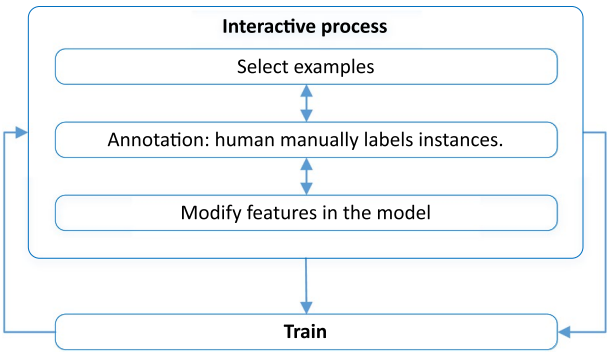

Fig. 5 Schematic representation of the IML process

Therefore, we can summarize the differences between AL and IML in the following points.

- AL is the basis for IML.
- The difference relies more in who has the control of the learning process and not in the interactivity of the approach. In AL the model retains the control and uses the human as an oracle; in IML there is a closer interaction between users and learning systems, so the control is shared.
- Since the interaction is closer in IML, we need to take into account Human-Computer Interaction techniques (HCI), something that is not so important in AL.
- In IML, humans perform more tasks other than labeling data in a freer and less structured process.

3.3 Applications of IML

Before starting to describe IML applications, it is useful to consider a brief description of the data according to their structure. In this respect, we can classify data as follows:

- **Structured data**: also known as *fully-structured*, is data that follows a predefined data model or schema (Sint et al. 2009; Abiteboul et al. 2000). A typical example is data that resides in tables in a relational database (or a similar structure like Excel tables or Pandas DataFrames).
- **Unstructured data**: is data that has no identifiable structure (Sint et al. 2009; Blumberg and Atre 2003), does not have a predefined model or does not fit into relational databases (Rusu et al. 2013). These include binary files such as image, video and audio files, and certain types of text documents. Non-relational or NoSQL databases are the best fit for managing this data.
- **Semi-structured data**: is a middle category between the other two and it is more complicated to define. We can describe it as data that does not conform with a data model or structure, but contains tags or markers that add semantics to that data (Rusu et al. 2013). Abiteboul et al. (2000) described it as "schemaless or self-describing terms that indicate that there is no separate description of the type or structure of the data" and Sint et al. (2009) state that "although this type of data does not require a schema, it does not mean that the definition is not possible, it is rather optional". This includes, for example, tagged text formats such as XML, JSON or YAML.

Even with these definitions, there are gray areas and corner cases that cast doubts on how to classify certain types of data. For example, structured data can contain unstructured elements, such as text documents or BLOBs (binary large objects).

When classical machine learning systems faced a problem with unstructured data, due to the limitations of these systems when dealing with raw data, a feature engineering phase is necessary to convert this raw data (e.g. pixel values of an image) into a suitable internal representation (e.g. direction of edges over the image) from which the learning model could detect or classify patterns (e.g. distinguish ones from zeros in images of handwritten digits).

Deep learning (LeCun et al. 2015), on the other hand, allows a machine to be fed with raw data and to automatically discover the representations needed for detection or classification. Deep learning models are composed of several layers that amplify aspects of

the input that are important for discriminate elements and suppress irrelevant variations. As noted by LeCun et al. (2015): "The key aspect of deep learning is that these layers of features are not designed by human engineers: they are learned from data using a general-purpose learning procedure".

This does not mean that feature engineering has no place in deep learning. It is often necessary to inject some form of prior knowledge in a deep model. But this little previous engineering is a competitive advantage over other learning models that is also fueled by the increases amounts of available data and computational performance that exists nowadays. For these reasons, deep learning systems have performed outstandingly well with unstructured data such as images, temporal signals, natural language processing and, in general, with any type of high-dimensional data.

Highlighting the structured and unstructured data division, and also emphasizing deep learning models ability to deal with the latter are the reasons why the application of IML seems to be especially useful in giving additional structure to something that does not have it. For example, one of the most comprehensive and recent papers we can find on applications of IML is that of Dudley and Kristensson (2018). The categorization used in this work is based on the underlying data type and therefore we can find the applications classified in the following categories: text, images, time series data, raw numerical data and, finally, assisted processing of structured information. As we can see, four out of five categories deal with unstructured data.

Let us look in detail at some of these applications in various application domains. We will highlight here those applications that are most relevant, other applications can be found in the aforementioned review by Dudley and Kristensson (2018), in the work of Jiang et al. (2019)—that classify the IML applications in a task-oriented taxonomy rather than in a data-oriented classification—or the work by Meza Martínez et al. (2019)—that defines an *integrative theoretical framework* for IML identifying five dimensions of application, namely *classification*, *clustering*, *information retrieval*, *regression* and *teaching intelligent agents* with 15 subdimensions inside them. We will organize the applications according to the type of data they use: image and video, time series data, text and, finally, complex data.

3.3.1 Image and video

Image classification has been one of the most successful fields in machine learning in general, and so it is in IML. One of the earliest works on the topic (Fails and Olsen 2003) develops the Crayons tool, a system that uses IML to create image classifiers. Another notable early example is CueFlik (Fogarty et al. 2008) a web image search application in which users create their own rules for classifying images giving examples and counterexamples of images that fulfill and do not fulfill the rules respectively.

But one application with images that has been very successful is using IML for interactive image segmentation. The idea is to facilitate the knowledge elicitation process by using experts as users of the IML tool and having them mark on the image they are shown content relevant to the model (the identification of a tumor, a face, etc.). One of the works that highlights the importance of this type of collaboration in IML was Porter et al. (2013) that remarks that "interactive image segmentation is an important tool in biomedical imaging, material science, geology, manufacturing, and food inspection".

Among the recent developments within this technique we can name the ilastik tool (Berg et al. 2019) that contains pre-defined workflows for image segmentation, object classification, counting and tracking and that allows non-expert users to interactively provide

annotations to steep the learning curve. Another tool is AIDE (Kellenberger et al. 2020), an acronym that stands for *Annotation Interface for Data-driven Ecology*. AIDE is an image annotation framework for ecological surveys that integrates closely users and machine learning models into the learning loop.

Finally, Jiang et al. (2019) made a thorough survey of existing IML works in the visual analytics community, focusing on those applications in which interactive visualizations allow users to interactively train machine learning models. These works include: visual pattern mining, interactive anomaly detection, interactive information retrieval, visual topic analysis and other tasks related with ML techniques such as visual cluster analysis and interactive processes of dimensionality reduction, classification, regression and model analysis.

3.3.2 Time-series data

Segmentation can be applied not only to images but also to video. For example, Kabra et al. (2013) use an interactive system that allows expert users, in this case biologists, to observe videos of different animals and allows them to add labels to frames in which they observe certain animals' behaviors. These labels are then transmitted to the underlying machine learning system. Video, as well as sound or biomedical signals, have a temporal component that makes us classify these applications within the time-series data section.

With respect to music, IML was applied in the field of composition. An example is the development of the Wekinator tool (Fiebrink and Cook 2010; Fiebrink 2011), a software system that enables the application of music information retrieval techniques based on machine learning, to real-time musical performance. Subsequent work originated from this research led to the development of the BeatBox tool (Hipke et al. 2014), a system that enables end-user creation of custom beatbox recognizers and interactive adaptation of recognizers to an end user's technique, environment, and musical goals.

But when we talk about sound we do not have to limit ourselves only to music. There are works within the interactive sound recognition area (Ishibashi et al. 2020) or in the Spoken Language Understanding (SLU) area (Begeja et al. 2004). We can find also hybrid sound-video applications, for example for the recognition of musical gestures (Visi and Tanaka 2021).

Finally, when we talk about time series, the idea that generally comes to mind is biomedical signals. (ECG, EEG, etc.). Here we also find applications of IML in the field of electromyography (EMG) analysis (Zbyszynski et al. 2020) or in the development of brain-computer interfaces (Kosmyna et al. 2015).

3.3.3 Text

One of the first applications that engages users interactively in the task of text processing is the work of Heimerl et al. (2012). It compares three approaches for interactive classifier training in a user study, incorporating active learning to various degrees in order to reduce the labeling effort as well as to increase effectiveness. They also add interactive visualization for letting users explore the status of the classifier and for judging its quality in iterative feedback loops, which is more like an IML approach than an AL approach.

On the other hand, the work of Wallace et al. (2012) is the first research publication to cite explicitly the application of IML to text processing with the development of the tool *abstrackr*, a system for facilitating citation screening for systematic reviews.

Another typical example that is often cited is the work of Šavelka et al. (2015) that applies IML to relevance assessment in statutory analysis, developing a framework in which a single human expert cooperates with a machine learning text classification algorithm.

Kim et al. (2015) investigate an efficient, accurate and scalable representation of high-dimensional, complex data points that aids human reasoning when working with ML models. The authors work with text documents from news articles, but the contents of these documents were complex in nature.

3.3.4 Complex data

The last paragraph leads us to another type of data on which the use of IML may be appropriate, that is, *complex data*. Complex data is defined by Castle (2017) as data that is *big*—i.e. we are dealing with large amounts of data that pose a challenge in terms of the computational resources needed to process them—and that come from many disparate sources—i.e. messy data, from multiple data sets that follow a different internal logic or structure. Carlson (2015) pointed that "in fact, relatively small data sets can often exhibit complexity making them difficult to analyze with traditional approaches". Tolls (2018) described complex data as "data whose type, structure and heterogeneity make it difficult to analyze" and remarks that this type of data can encompass both large and small data sets.

There are domains, such as medicine, where it is common to find complex data to feed our machine learning systems. Holzinger and Jurisica (2014) stated that the central problem in healthcare and biomedical research is that biomedical data models are characterized by significant complexity provoking *information overload* (Berghel 1997), that is "drowning in data, yet starving for knowledge".

This complex data in medical domains combine vast amounts of diverse data, including structured, semi-structured and weakly structured data and unstructured information (Holzinger and Jurisica 2014). In this domain, interactivity with humans is used to generate hypothesis as well as for extract relationships and information from the data. For this reason it is common to see applications of IML in medical domains, as we have already mentioned (Kosmyna et al. 2015; Porter et al. 2013; Berg et al. 2019). Another example of these applications is Fadhil and Wang (2018) that introduce an application of interactive machine learning (IML) in a telemedicine system, to enable automatic and personalized interventions for lifestyle promotion.

Holzinger et al. (2019) stated that IML is particularly suitable in the medical domain. This is due to the fact that ML algorithms, especially the deep learning ones, require large amounts of data to be able to infer models. In medicine we can find big databases with clinical cases, for example, The Cancer Genome Atlas (TCGA) (Liu et al. 2018b; Tomczak et al. 2015) but when we focus on a particular type of cancer, the data available may not be sufficient to train a deep learning model. A *doctor-in-the-loop* approach can use human expertise and long-term experience to fill the gaps in large amounts of data or deal with complex data (Holzinger 2016).

In addition, we can add the problem of the black-box nature of many of the ML models; although we understand the underlying mathematical principles of such models, they lack an explicit declarative knowledge representation that eases the interpretation of the decisional process (Holzinger et al. 2017), something that has great importance in clinical practice. IML can be a means to solve these problems: incorporating human knowledge

and skills we can improve the quality of the learning models and build them with less data, and this human component also enables features as re-traceability and explainability that mitigate the black-box problem. At this respect, Teso and Kersting (2019) developed a framework called *explanatory interactive learning* in which at each step, the learner queries users but also explains that query to them. Users then answer the query but also correct the explanation. The idea is to enhance the explanatory power of ML algorithms and, consequently, the trust that the users put in them. We will review Explainable AI (XAI) in Sect. 6.

3.4 IML shortcomings

Since IML is based on AL it shares some of its shortcomings, but it also adds some of its own. The most obvious one is that the increased interactivity causes ML aspects to be intermingled with HCI aspects. According to Michael et al. (2020) this entails much more effort in the development of applications, since they must be built and studied uniquely. Perhaps the search for methodologies and theoretical frameworks for IML systems, such as the one proposed by Meza Martínez et al. (2019) may be the solution in the future.

Another drawback that arises in IML and that it also shares with other interactive models is the dependence on the presence of human experts. The revolution and paradigm shift brought about by the development of techniques such as deep learning (LeCun et al. 2015) was largely based on taking humans out of the equation in exchange for substantially increasing computational requirements. IML systems promise to lower these computational requirements and make learning more efficient, in exchange for bringing humans back into the equation with the problems associated to them (availability, attention, interactivity, different expertise, etc.).

4 Machine teaching (MT)

Machine Teaching (MT) is another approach to transferring knowledge from (initially) humans to computers. If the approaches discussed in the previous sections have been differentiated by who is in control of the learning process, the MT paradigm places the responsibility firmly on the teacher.

Even though the MT paradigm is quite different in nature to the other paradigms described in this paper—and represents an alternative to them—there are many common factors. Over time, the process has become iterative and incremental; occasionally, it has sought inspiration in other approaches, such as active learning; and at times it has ended up obtaining results that are comparable to other techniques, such as curriculum learning (which is defined in the next chapter).

The term Machine Teaching has meant different things at different times, but today it is mainly used in the field of machine learning to describe the idea of a teacher who teaches an ML model to an ML algorithm. The teacher is meant to be a human, although algorithms simulating teachers have become more common lately.

In Fig. 6 we can see a mind map of MT in which we try to distinguish between a more classical interpretation of the term (used in intelligent systems designed to teach humans) to a more current conception of it where we have humans acting as teachers of machines, or even machines themselves being teachers of other machines.

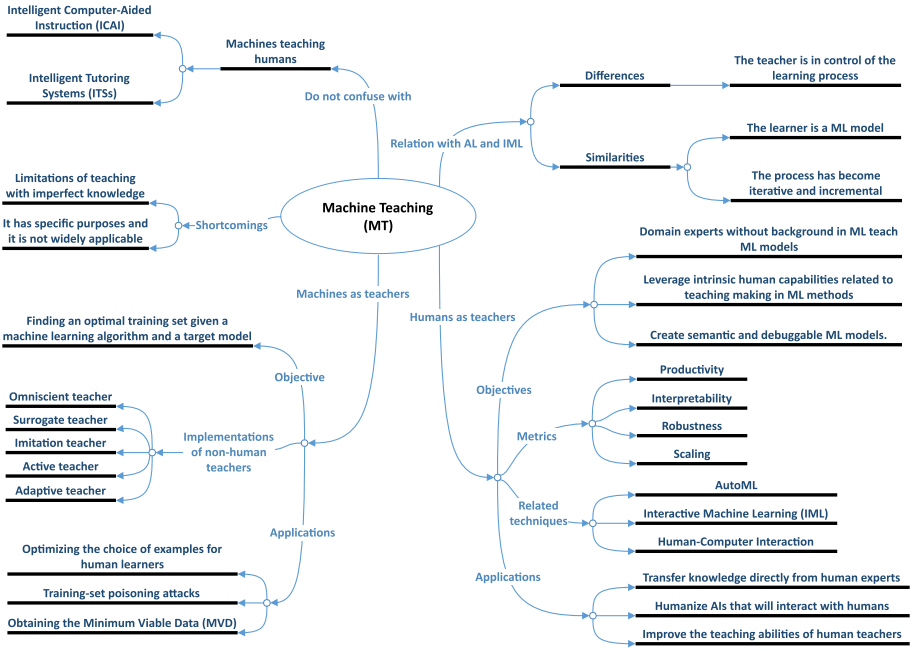

Fig. 6 Machine Teaching (MT) mind map

4.1 Antecedents of machine teaching

The idea of learning algorithms who are taught by humans was introduced several decades ago. An important author is Angluin (1987), who theorizes about human experts trying to communicate their method to an expert system. Ideally, the examples should be "central" or "crucial", in order to help the system converge to a correct hypothesis. She notes that the source for learning general rules from examples had historically been arbitrary or random, and aims instead for learning methods in which the source of examples is *helpful*. She also uses the terms *learner* and *teacher* in the same way they have been used by the other approaches discussed in previous sections of this paper. In her case, the learning algorithm is the *learner*, and the source of examples the *teacher*. This leads to the concept of the *Minimally Adequate Teacher*, who is expected to answer questions from the learner correctly. Angluin (1987) uses context-free languages as her field of application and the questions are related to sets and membership. The answers can be yes, no, or a counterexample, depending on the type of question.

Two tangentially related terms that should be mentioned in a prehistory of Machine Teaching are Intelligent Computer-Aided Instruction (ICAI) and Intelligent Tutoring Systems (ITSs), which are considered synonymous to all intents and purposes. An overview was written in Nwana (1990). The goal is to use artificial intelligence techniques to improve education (e.g., learning mathematics). In these paradigms, the roles are typically reversed, so the teacher is a computer and the learner is a human. In fact, since the term "Machine Teaching" is such an ordinary combination of words, some of the early uses of the term that can be found refer precisely to machines teaching humans [e.g., Weimer (2010) or Johns et al. (2015)]. In the rest of this section, however, we will focus primarily on the question of transferring knowledge to machine learning models.

4.2 Humans as teachers

One of the first authors to use the term *Machine Teaching* specifically as a related but alternative approach to machine learning is Diamant (2006), who criticized the attempts of conventional image-processing paradigms to extract high-level semantics from low-level processing stages. Instead, he argues that semantics is not an inherent property of an image but a property of a human observer watching it. Diamant uses the term *world ontology* to describe this knowledge about the outer world, suggesting that a vision machine can be provided with a replica of the ontology, which does not have to be entirely full.

He also objects to describing knowledge transfer (in the human and animal world) as a "learning process" (which could be considered the original inspiration for machine learning). Rather, Diamant (2009) notes that teaching in the natural world (e.g., in animal herds) does not mean human-like mentoring but can involve a specific semantic transference of knowledge, a quasi-mechanistic transmission of information from a teacher to a pupil, or from one community member to another. In this interpretation, machine learning and Machine Teaching are not simply two techniques for arriving at the same goal but two radically different approaches.

Machine teaching has recently attracted the attention of big multinational corporations, such as Microsoft. A team of Microsoft researchers (Simard et al. 2017) considered that the current processes for building machine learning systems require practitioners with deep knowledge of machine learning, and in order to meet the growing demand for ML systems it is necessary to significantly increase the number of individuals that can teach machines. This means that building ML systems should be available to domain experts with little or no ML expertise.

This team of researchers considers machine teaching a discipline in its own right, which they describe as living at the intersection of the human-computer interaction, machine learning, visualization, systems and engineering fields. In fact, machine teaching is considered such a big paradigm shift away from machine learning as to treat algorithms as swappable pieces. Whereas machine learning aims to create new algorithms to improve the accuracy of the learners, Machine Teaching is focused on the *efficacy* of the *teachers*. The metrics for measuring performance would include *productivity*, *interpretability*, *robustness*, and *scaling with the complexity of the problem or the number of contributors* (Simard et al. 2017). The latter is key, as machine learning can be costly in terms of time and expertise, whereas Machine Teaching proponents want to open up and democratize the field. We can emphasize two ways in which this is manifested:

- Involving domain experts who do not necessarily have a background in machine learning.
- Helping to address problems in which labeled data is hard to find.

How is information delivered? Zhu et al. (2018) define two dimensions of Machine Teaching, as follows:

- **Batch teaching**, in which the teacher gives the learner a training set, the order is unimportant and elements may or may not be duplicated.
- **Sequential teaching**, in which the student learns in a sequence and the order matters and should be optimized by the teacher.

Here it should be noted that the terms batch and sequential are used with a slightly different meaning than when talking about active learning (Sect. 2). In that chapter a typical description of the batch process vs. the sequential ML process was given. In batch the process runs in batches to be more efficient, while the sequential model is less computationally efficient but tries to get immediate feedback on the results. Zhu et al. (2018) adds a new dimension here that has to do not only with the number of cases presented to the model but also with their ordering. Thus, in this case, batch means that the cases are not ordered, and sequential means that the cases are ordered and are supplied one by one according to the established order.

In terms of actual implementation, Machine Teaching began as a batch process in which the learner was fed examples in one shot, which inevitably placed the focus on the *size* of the data set, as we have discussed previously. Later, Simard et al. (2017) implemented the Machine Teaching process as a "never-ending loop", which was used by Wall et al. (2019) in what is perhaps the first study of actual end-users engaging in this type of process. Wall et al. (2019) establish a loop involving a human teacher, a machine learning model, and a large set of unlabeled data called the *sampling set*.

The teacher begins by exploring the sampling set and then adds or edits labels or features. These are used to train the ML model. If it becomes necessary to fix errors (such as mislabeling errors, learner errors, or representation errors), the teacher goes back to adding or editing labels or features. Otherwise, we go back to the beginning of the loop and start a new iteration until it is decided that we are done for now.

The teacher should be able to articulate the relevant concepts that explain why a document is labeled as a member of a specific class, and the learner should have a consistent underlying learning algorithm. There is no predetermined test or ground-truth set that assesses the quality of the model. Evaluation can be judged based on the number of correct predictions in a dynamically-generated set of positively predicted documents (Wall et al. 2019).

Ramos et al. (2020) go a step further and talk about a human-centered approach to building machine-learned models which they call Interactive Machine Teaching (IMT). They choose that name because they see IMT as an IML process in which the human-in-the-loop takes the role of a teacher, and their goal is to create a machine-learned model, so the name is therefore analogous to IML. But this term is arguably confusing (as Ramos et al. themselves acknowledge in their paper) with other acronyms that use the term IMT to stand for different things. In the next section of this paper we cite Iterative Machine Teaching (not interactive), where machines teach other machines and the term Interactive Machine Teaching is also mixed up, as we have expressed before, with machines teaching humans. For this reason we have titled this section "Humans as teachers" in order to avoid this confusion.

Ramos et al. (2020) is a good summary of the state of the art of MT in which humans act as teachers of machines. They argue that this technique enable people to leverage intrinsic human capabilities related to teaching making machine learning methods accessible to subject-matter experts and allowing the creation of semantic and debuggable ML models. A schematic representation of this process can be shown in Fig. 7 in which we can see a teaching loop in which the human expert prepares a curriculum of cases, explains the knowledge contained in it to the learner and, afterwards, reviews the model predictions and its reasoning.

This possibility of giving non-ML experts access to ML construction has already been explored in a number of ways. For example, Automated machine learning tools (AutoML) provide methods and processes to make ML model creation and evaluation easier, and

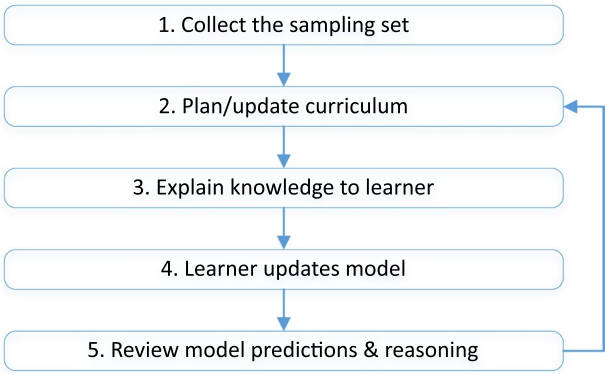

Fig. 7 Schematic representation of the MT process

ultimately, available for non-machine learning experts. This set of tools seeks to automate the decision on what learning algorithms to use, what hyper-parameters to select or which features are more relevant for a certain model. Furthermore, they provide a means of model evaluation and optimization. Many cloud platforms (e.g., AWS, Azure, Google) are including AutoML tools as part of their ML stacks. The problem with the AutoML approach is that it produces black-box models that are inadequate for scenarios requiring transparency, and also can be impractical to use in cases where there are not significant labeled data available.

4.3 Machines as teachers

Zhu (2015) draws on the pioneering work of Angluin (1987) and, like Diamant (2009), explicitly uses the term Machine Teaching, which he similarly describes as an inverse (in an almost mathematical sense) problem to machine learning. But the focus is quite different in this case, as Zhu is deliberately uninterested in "hard wiring" the knowledge into the learner. Instead, he chooses to give a very specific definition to the term Machine Teaching, defining it as "the problem of finding an optimal training set given a machine learning algorithm and a target model" (Zhu 2015).

In this version of Machine Teaching, the teacher knows the target model in advance, and also the student's algorithm, and must "teach" the former to the latter somehow. In practice, this means designing the optimal training set (typically the smallest one, but not always). This type of Machine Teaching also differs from related approaches, such as active learning. The main difference is that the teacher knows the target model upfront and does not need to explore.

Liu et al. (2017) coined the term *Iterative Machine Teaching* as an evolution of MT in which the learner's model is continuously updated by an *iterative algorithm*. That is, the process consists of iterations and the teacher is not a human but an algorithm. Again, this is very easily confused with the term *Interactive Machine Teaching* seen above, that is why we have preferred to call this section "Machines as teachers".

Essentially, the learner remains passive and the teacher observes, influences, and communicates with the learner, choosing an example and feeding it to the learner, who runs a fixed iterative algorithm using that example. Rather than focusing on the size of the entire data set, Iterative Machine Teaching focuses on questions of *sequence* [sometimes obtaining results that are similar to curriculum learning, explained in Sect. 5, as discovered by

Liu et al. (2017)], *convergence*, and the so-called *iterative teaching dimension*, which is the smallest number of examples (or rounds) necessary for learning.

The literature offers several implementations of non-human teachers for the Iterative Machine Teaching paradigm. A non-exhaustive list of examples would include Liu et al. (2017) (classified according to the information the teacher has about the learner), Liu et al. (2018c) (inspired by active learning), and Chen et al. (2018) (based on adaptivity). A summary of these implementations would be as follows:

- **Omniscient teacher** (Liu et al. 2017), which has total access to the characteristics of the learner, namely, feature space, model, loss function and optimization algorithm. Here, an example's difficulty is calculated using the learner's loss function (as the norm of the gradient of the squared loss function) and usefulness is calculated using the discrepancies between the learner's weights and those of the teacher while considering the difficulty of the example.
- **Surrogate teacher** (Liu et al. 2017), which has access just to the loss function. An example's difficulty is calculated as in the omniscient teacher, but an example's usefulness must be calculated in a different way, because only the loss function is available. Here, the usefulness is calculated as the learner's loss minus the teacher's loss (considering again the difficulty of the example).
- **Imitation teacher** (Liu et al. 2017), which does not have access to the performance of the learner. Thus, the teacher needs a copy of the learner to be used as a reference for the selection of examples. In this copy, the teacher would have access to the learning parameters as if it were an omniscient teacher.
- **Active teacher** (Liu et al. 2018c), which is inspired by human teaching, real-world exams and active learning. The teacher is not directly observing the student, but can actively make queries with a few samples in each iteration. The student will return its predictions and the teacher will estimate the student's status based on this feedback, and also determine which example must be provided next.
- **Adaptive teacher** (Chen et al. 2018), which uses adaptivity and observes the learner's hypothesis at every time step. It would be the opposite of a *non-adaptive teacher* who does not receive any feedback during teaching and only knows the initial hypothesis of the learner. Chen et al. (2018) propose this model as an improvement over the *omniscient*, *surrogate* and *imitation* teachers above, finding that adaptivity plays a key role and the learner's transitions are smooth and interpretable, with the learner transitioning to the next hypothesis according to some local preference (i.e., dependent on the current hypothesis).

4.4 Applications of MT

As mentioned above, the current conception of Machine Teaching typically consists in humans acting as teachers of machines, who are usually the learners. The main usefulness of this approach is to take advantage of the inherent abilities of humans for teaching, in order to allow people without a machine learning background to transfer knowledge to a computer system similarly to how they would teach another human. An example of this approach is PICL (Ramos et al. 2020).

Continuing this approach of humans teaching machines, MT can be used in fields such as robotics. For example, Sena et al. (2018) and Sena and Howard (2020) use MT to teach robot manipulators by example, rather than by programming them, which would also have

the benefit of helping humans become better teachers. In a software context, we can find the work of Peng et al. (2021) in which task bots (task-oriented dialog systems) communicate with users through natural language and are trained by humans via MT, which, according to the authors, reduces the cost of fine-tuning. The idea is to use human teachers to make these bots more convincing because ultimately they are going to interact with other human users.

As can be seen, the field of education and training is a natural fit for this approach. When we discussed the antecedents of MT, we briefly mentioned Intelligent Tutoring Systems (ITSs), in which the machine acted as the teacher and the human was the learner. Machine Teaching now lets us add a layer to this process, in which a human teacher will first train the computer system in order to make it a better teacher. After it has been trained, this system will be used to train other humans. For example, Weitekamp et al. (2020) have used this approach to help develop and evolve AI tutors who teach humans. Taken to a workplace setting, but inspired by challenges in data-driven online education, Singla et al. (2014) focus on teaching workers how to classify in crowdsourcing services, with the goal of improving their accuracy.

We have also previously discussed the possibility of machines teaching machines. In this case, the target model is known in advance by the teacher, so the practical applications must necessarily be different. For example, Zhu (2015) uses it to deal with malicious uses, such as training-set attacks (Mei and Zhu 2015) or training-set poisoning (Zhu et al. 2018). Here, attackers pollute the training data so that a specific learning algorithm produces a model that is beneficial to them (e.g. manipulating spam filters to avoid the detection of malicious emails). Other applications of machines teaching machines are more methodological in nature. Mosqueira-Rey et al. (2021) suggest using this technique for obtaining the "*Minimum Viable Data (MVD)*" for training a learning model. MVD is a term coined by van Allen (2018) that refers to the minimum data needed to train machine learning models. The name is borrowed from the agile world, in which we have the idea of a "*Minimum Viable Product (MVP)*", a product with just enough features to satisfy early customers, and to provide feedback for future product development.

4.5 MT shortcomings

In light of the typical applications of Machine Teaching that were outlined above, it should be obvious that MT has very specific purposes and should only be used for particular types of tasks. For example, if the target model is already known in advance, there must be other reasons for using MT, such as making the learner more human-like, or improving the teaching abilities of human teachers.

Even when MT is the most promising approach to solving a problem, some basic conditions must still be met. For example, Devidze et al. (2020) discuss the limitations of teaching with imperfect knowledge, explaining some assumptions that are generally made, which may or may not be actually true. For example, it is often assumed that the teacher has perfect knowledge of:

- The learner (e.g., a computational model of the dynamics of learning, and parameters representing the initial knowledge and rate of learning).
- The task specification (e.g., a comprehensive representation of the task and ground-truth data).

In fact, the teacher's knowledge is often incomplete or flawed, and the resulting curriculum is often far for optimal. Devidze et al. (2020) have analyzed this type of situation and conducted an experimental evaluation of teaching with imperfect knowledge. One of their most interesting conclusions is that assessing the learning rate correctly is significantly more important than the prior knowledge of the learner.

5 Curriculum learning (CL)

As we have seen in previous sections, in order to teach complex tasks, teachers are often required to organize the concepts that constitute the final task, taking into account its complexity. This organization leads to what is known as a curriculum. The student is gradually introduced to the concepts in the curriculum by increasing complexity, in order to take advantage of previously learned concepts and ease the abstraction of new ones.

In many traditional machine learning paradigms, the learner (the *student*) estimates an objective function using a set of training label examples (supplied by the *teacher*). These examples are randomly presented to the model, ignoring the complexities of data samples and the learning status of the current model. So the following question arises: could the curriculum-like training strategy benefit machine learning? According to Wang et al. (2021), the power of introducing curriculum into machine learning depends on how the curriculum is designed for specific applications and data sets. Its advantages can be summarized as improving the model performance and accelerating the training process, which cover the two most significant requirements in major machine learning research.

The idea of training a learning machine with a curriculum can be traced back at least to Elman (1993). The basic insight is to *start small*, learn easier aspects of the task, and then gradually increase the difficulty level. With this idea in mind, Bengio et al. (2009) confirm that machine learning algorithms can benefit from a curriculum strategy and that a well chosen curriculum strategy can help to find better local minima of a non-convex training criterion. Bengio proposes the term curriculum learning (CL) as the training strategy that trains a machine learning model with a curriculum. With experiments on supervised visual and language learning tasks, the author showed that some curriculum strategies work better than others, that some are useless for some tasks, and that better results could be obtained on specific data sets with more appropriate curriculum strategies.

In Fig. 8 we can see a mind map with the aspects related to curriculum learning that we will detail below.

5.1 CL perspectives and motivations

To explain the CL advantages, two perspectives should be analyzed: *optimization problem* and *data distribution*. The results of this analysis allow to establish two motivations for applying CL: to *guide* and to *denoise*. From the **optimization problem** point of view, Bengio et al. (2009) bring up that CL can be seen as a particular *continuation method*, an optimization strategy for non-convex criteria which first optimizes a smoother (and also easier) version of the problem to reveal the "global picture", and then gradually consider less smoothing versions, until the target objective of interest. In this way, continuation methods *guide* the training towards better regions in parameter space.

On the other hand, from the perspective of **data distribution** and due to the big data collection which brings noisy data that is less cognizable or wrongly annotated, the CL

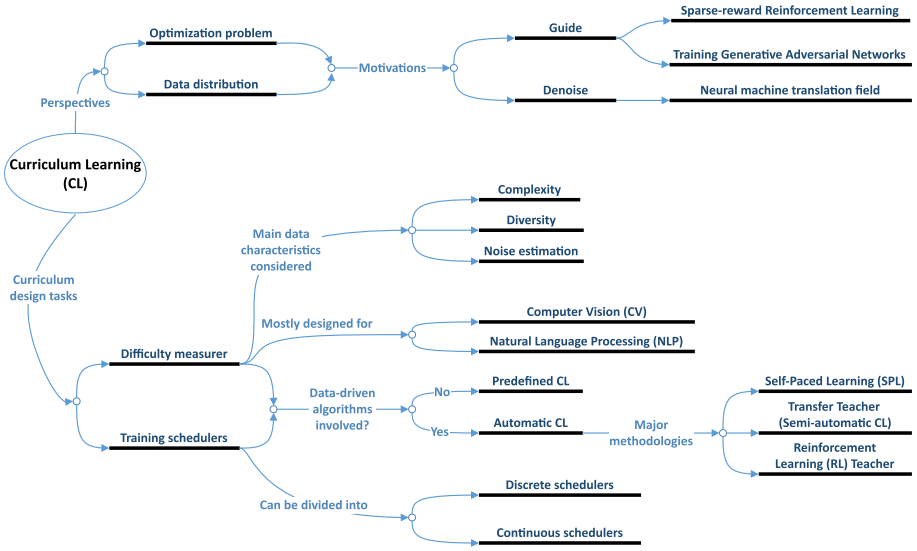

Fig. 8 Curriculum learning (CL) mind map

strategy considers this noisy data as harder examples in the data sets and the cleaner data as the easier ones. Since this strategy motivates training on easier examples, an intuitive hypothesis that reveals the *denoising* efficacy of CL on noisy data is established: CL learner wastes less time with the harder and noisier examples (Bengio et al. 2009).

Two are the motivations for applying CL, that is, to **guide**, regularizing the training towards better regions in parameter space, as from the perspective of the optimization problem; and to **denoise**, focusing on high-confidence easier area to mitigate the interference of noisy data as from the perspective of data distribution. In this sense, most of the applications of CL can be classified into these two groups (Wang et al. 2021). The guide motivation group involves difficult tasks where direct training on them results in poor performance or slow convergence. CL strategies guide the training from easier tasks to the target ones, being examples sparse-reward Reinforcement Learning (Florensa et al. 2017) or training Generative Adversarial Networks (Soviany et al. 2020). It also concerns tasks where the target distribution is quite different from the training distribution, and a good curriculum helps to guide the training to adapt to the target distribution. In this case, domain adaptation setting (Zhang et al. 2019b) and imbalanced classification (Wang et al. 2019) are representative settings. The *denoise* motivation group involves noisy or heterogeneous training data sets and CL strategies could help making the training faster, more robust, and more generalizable, being a popular application the neural machine translation field (Kumar et al. 2019).

The way curriculum strategies have been defined leaves a lot of work to the teacher. To minimize the amount of teacher effort involved keeping the advantages of a curriculum strategy, it is natural to consider a form of active selection of examples similar to what humans do. Curriculum learning endeavors to impose some structure on the training set. This structure basically relies on identifying *hard* and *easy* examples, and trust in this distinction in order to teach the learner. It would be advantageous for a learner to focus on *interesting* examples, which would be standing near the frontier of the learner's knowledge

and abilities, neither too easy nor too hard. This approach could be used to at least automate the pace at which a learner would move along a predefined curriculum.

5.2 Curriculum design

To organize a curriculum for students, teachers need to deal with two challenging tasks:

- Arrange the material taking into account its complexity or difficulty, a knowledge that is not available in the training set in most of machine learning paradigms. This task is referred to as scoring function or **difficulty measurer** (Wang et al. 2021).
- Guide the pace at which the material is presented, known as pacing function or **training scheduler** (Wang et al. 2021).

Therefore, a general framework for curriculum design consists of these two main components. The nature of the difficulty measurer and the training scheduler lead us to two different CL categories:

- Specifically, when both the difficulty measurer and the training scheduler are designed by human prior knowledge with no data-driven algorithms involved, the CL method is called **predefined CL**.
- If any (or both) of the two components are learned by data-driven models or algorithms, then the CL method is known as **automatic CL**.

5.2.1 Difficulty measurers

Researchers have manually designed various difficulty measurers mainly based on the data characteristics of specific tasks and most of them designed for image in Computer Vision (CV) and text data in Natural Language Processing (NLP). Among data characteristics, complexity, diversity and noise estimation are considered. **Complexity** represents the structural complexity of a particular example, such that examples with higher complexity have more dimensions and are thus harder to be captured by the models (e.g. sentence length in NLP tasks). **Diversity** stands for the distributional diversity of a group of examples (e.g. regular or irregular shapes in CV tasks) where a larger value of diversity means the data is more diverse and is more difficult for model learning. Larger diversity sometimes also makes the data noisier. So, another characteristic to be studied is **noise estimation** which estimates the noise level of examples and defines cleaner data as easier.

Other interesting difficulty measurers include human-annotation based Image Difficulty Scores (Soviany et al. 2020; Ionescu et al. 2016) which are proposed to measure the difficulty of an image by collecting the response times of human annotators answering "Yes" or "No" at identifying objects in images. Intuitively, longer response time corresponds to harder image examples. These measures can be considered separately but also correlated. For example, high complexity and high diversity bring more degrees of freedom to the data, which needs a model with larger capacity and bigger effort of training.

5.2.2 Training schedulers

While predefined difficulty measurers differ among diverse data types and tasks, the existing predefined training schedulers are usually data/task agnostic and can be divided into

discrete and *continuous* schedulers. The difference lies in the method of adjustment of the training data subset.

Discrete schedulers adjust the training data subset after every fixed number (> 1) of epochs or convergence on the current data subset and are commonly used due to their simplicity and effectiveness (Bengio et al. 2009; Spitkovsky et al. 2010; Ionescu et al. 2016).

Continuous schedulers adjust the training data subset at every epoch by mapping training epoch number to a proportion of the easiest examples available at each epoch. This mapping function can be a linear function (Hacohen and Weinshall 2019), a root function (Platanios et al. 2019), or even a geometric progression function (Penha and Hauff 2020). There is also a special group of continuous schedulers, referred to as *distribution shift*, that start on an initial distribution and gradually move to a target distribution (Liu et al. 2018a).

5.2.3 Automatic CL

Despite the simplicity and effectiveness of predefined CL, some limitations appear (Wang et al. 2021): (a) it is difficult to find the best combination of difficulty measurers and training scheduler for a specific task and its data set; (b) both of them stay fixed during training process ignoring the feedback of the current model; (c) expert domain knowledge is needed for designing a predefined difficulty measurer; (d) easy examples for humans are not always easy for models, since their decision boundaries are basically different; (e) the best hyperparameters of training schedulers are hard to find; and, (f) the performance of several predefined difficulty measurers is sensitive to the initial learning rate.

The limitations of predefined CL have prevented CL from being explored in more diverse applications, so automatic CL methods were introduced to overcome these limitations and to reduce the need of human teachers in the curriculum design. Three major methodologies exist: self-paced learning (SPL), transfer teacher, and Reinforcement Learning (RL) Teacher.

Self-paced learning (SPL) methods let the model (student) act as a teacher and measure the difficulty of training examples according to its losses on them. This method, initially proposed by Kumar et al. (2010), automates the difficulty measurer by taking the example-wise training loss of the current model as criteria and training the model at each iteration with the proportion of data with the lowest training losses. This proportion gradually grows to the whole training set. The SPL method embeds the curriculum design into the learning objective of the original machine learning tasks (Liu et al. 2021).

Transfer teacher methods let a stronger teacher model act as the teacher and measure the difficulty of training examples according to the teacher's performance on them. It comes up from the idea of human education where the student finds it hard to measure the difficulty of the materials if he understands a little about them. So, it is advantageous to ask a mature teacher to help the student organizing an easy-to-hard curriculum. The transfer teacher method is a semi-automatic CL method introduced by Weinshall et al. (2018). In particular, the training examples are sorted based on the performance of a pre-trained network on a larger data set, adjusted to the data set at hand, and then its knowledge is transferred applying a predefined training scheduler to finish the CL design. Most of these methods are loss-based ones which do not need domain knowledge and are closely related to SPL (Xu et al. 2020). Nevertheless, in the NLP literature, there exist some methods which used transfer teacher based on cross-entropy (Zhou et al. 2020). Hacohen and Weinshall (2019) include the idea of curriculum learning by bootstrapping based on self-tutoring as a different scoring function. Here, the network is trained without curriculum and the

resulting classifier is used to rank the training data in order to train the same network again from scratch.

Reinforcement Learning (RL) Teacher methods adopt reinforcement learning models as the teacher to play dynamic data selection according to the feedback from the student. These methods are based on the idea of an ideal teaching strategy where both the teacher and the student are involved and improve together: the student provides feedback to the teacher, who adjusts the teaching action accordingly. At each training epoch, the RL Teacher dynamically selects examples for training (*action* in the RL schemes) according to the student's feedback (*state* and *reward* in the RL schemes). From the RL Teacher sets, the teacher model as both the difficulty measurer and training scheduler by dynamically considering the student's feedback. Both traditional RL and Deep RL models are leveraged in these designs (Kumar et al. 2019; Matiisen et al. 2020), where the Deep RL models are stronger in performance but more time-consuming and harder to train.

5.3 How to select a CL method

There is no guidance for the selection of a CL methodology in real-world applications, so Wang et al. (2021) offer some conclusions from empirical studies that although scarce, compare and analyze different CL methods. Cirik et al. (2016) showed that predefined CL benefits more when smaller models are applied and the size of the training set is limited. Zhang et al. (2018) concluded that predefined CL is highly sensitive to the choices of difficult measurer and the hyperparameters. Hacohen and Weinshall (2019) demonstrated that transfer teacher is the most robust automatic CL, and that the advantage of CL is more effective when the task is difficult.

As it can be seen, the best selection among different CL categories needs further empirical studies. This selection could be guided by the knowledge about the data set and the task goal. If expert domain knowledge is available, then predefined CL methods are more suitable to design a knowledge-driven curriculum. On the contrary, if we have no prior assumptions on the data, then automatic CL methods are more suitable to learn a data-driven curriculum adaptive to the data set and task goal. Nevertheless, some hybrid CL methods are designed which adopt different CL methods making them complement to each other (Jiang et al. 2015; Zhang et al. 2019a). Wang et al. (2021) present as an interesting idea for future research to embed human prior on sample importance into the fully data-driven CL methods which has been explored in Wang et al. (2020).

As it has been shown, CL is related to Transfer Learning and Multi-task Learning but is also connected to active learning (AL) (Cohn et al. 1996). Both of them involve dynamic data selection but their goals are quite different, as it can be seen. In AL, an active learner improves performance with fewer labeled data through questions to an expert to annotate unlabeled instances for further training. CL improves performance and accelerates convergence in supervised, weakly-supervised, and unsupervised settings, while AL is designed for label-saving training in the semi-supervised setting.

6 Explainable AI (XAI)

As stated previously, humans can carry out the learning process of machine learning systems as teachers and this fact can affect the performance of these systems. Nevertheless, when decisions derived from sophisticated AI-powered systems affect humans' lives (as in

e.g. medicine, law or defense), there is an emerging need for understanding how such decisions are furnished by AI methods (Goodman and Flaxman 2017).

While at very first AI systems were easily interpretable, the rise of opaque models, such as Deep Neural Networks (DNNs), has raised questions about how trustworthy these systems are, preventing users from tracing the logic behind predictions. The danger is in creating and using decisions that are not justifiable, legitimate, or that simply do not allow obtaining detailed explanations of their behavior.

In order to avoid limiting the effectiveness of the current generation of AI systems, eXplainable AI (XAI) (Gunning 2017) proposes creating a suite of machine learning techniques that (1) produce more explainable models while maintaining a high level of learning performance, and (2) enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

In this section we will try to define the concept of Explainable AI (XAI) and the different techniques that exist, focusing on deep learning. A mind map of XAI can be found at Fig. 9.

6.1 What is explainable artificial intelligence?

The most commonly used nomenclature used in XAI communities includes terms such as understandability, comprehensibility, interpretability, explainability and transparency. In all the above terms, understandability comes up as the most important concept in XAI and it has to be considered on the one hand as model understandability and on the other, human understandability. This is the reason why the definition of XAI refers to the concept of audience, defined as the users of the model, as the cognitive skills and pursued goal of those users have to be taken into account jointly with the intelligibility and comprehensibility

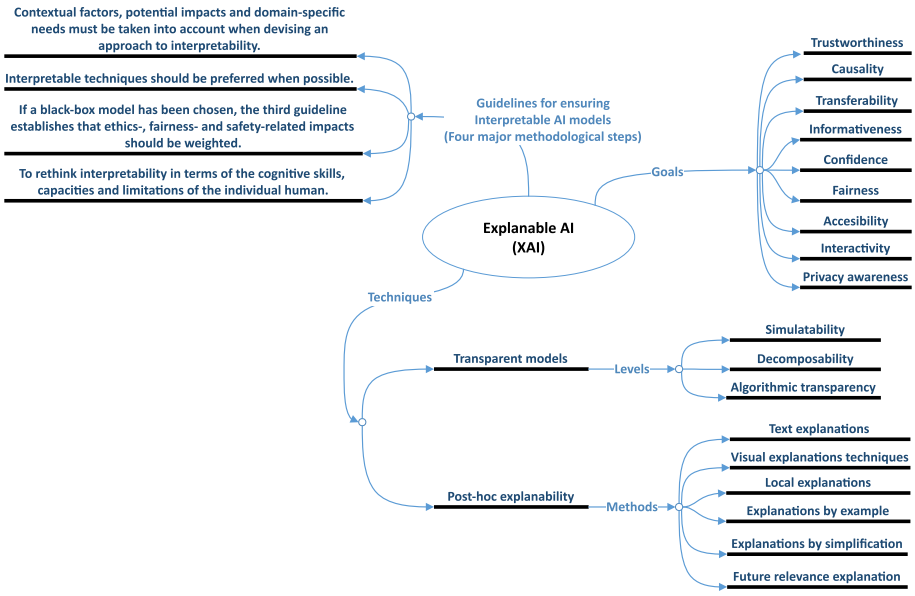

Fig. 9 Explainable AI (XAI) mind map

of the model in use. This prominent role taken by understandability makes the concept of audience the cornerstone of XAI (Barredo Arrieta et al. 2020).

Let's take as a starting point the definition of the term XAI given by Gunning (2017): "XAI will create a suite of machine learning techniques that enables human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners".

A model can be explained, but the interpretability of the model is something that comes from the design of the model itself. Bearing this in mind, explainable AI can be defined as follows: "Given an audience, an explainable Artificial Intelligence is one that produces details or reasons to make its functioning clear or easy to understand".

This definition, a first contribution of Barredo Arrieta et al. (2020), implicitly assumes that the ease of understanding and clarity targeted by XAI techniques for the model at hand, results in different application purposes, such as a better trustworthiness of the model's output by the audience.

The research activity around XAI has exposed different goals to draw from the achievement of an explainable model. In Barredo Arrieta et al. (2020), authors synthesize and enumerate definitions for these XAI goals, so as to settle a first classification criteria:

- **Trustworthiness**: it might be considered as the confidence of whether a model will act as intended when facing a given problem. Although it should most certainly be a property of any explainable model, it does not imply that every trustworthy model can be considered explainable on its own.
- **Causality**: considered as the inference of causal relationships from observational data. An explainable ML model could validate the results provided by causality inference techniques, or provide a first intuition of possible causal relationships within the available data.
- **Transferability**: the mere understanding of the inner relations taking place within a model facilitates the ability of a user to reuse this knowledge in another problem. Transferability should also fall between the resulting properties of an explainable model, but again, not every transferable model should be considered as explainable.
- **Informativeness**: as ML models are used with the ultimate intention of supporting decision making, a great deal of information is needed in order to be able to relate the user's decision to the solution given by the model, and to avoid falling in misconception pitfalls.
- **Confidence**: as a generalization of robustness and stability, confidence should always be assessed on a model in which reliability is expected.
- **Fairness**: from a social standpoint, explainability can be considered as the capacity to reach and guarantee fairness in ML models, and it should be considered as a bridge to avoid the unfair or unethical use of algorithm's outputs.
- **Accessibility**: explainable models will ease the burden felt by non-technical or non-expert users when having to deal with algorithms that seem incomprehensible at first sight.
- **Interactivity**: this goal is related to fields in which the end users are of great importance, and their ability to tweak and interact with the models is what ensures success.
- **Privacy awareness**: the ability to explain the inner relations of a trained model by non-authorized third parties may also compromise the differential privacy of the data origin.

Nevertheless, there exists more recent reviews with another criteria (Minh et al. 2021; Meske et al. 2022; Das and Rad 2020). For example, Minh et al. (2021) mentions as the

XAI nomenclature: *understandability* that is linked to informativeness, causality, transferability, fairness and confidence; *comprehensibility*, as dependent on the users' ability to perceive the knowledge that is learned by a model and is linked to informativeness and accessibility; and *interpretability*, as estimating the level that the users can comprehend the outputs of the AI models and is linked to trustworthiness, causality, transferability, informativeness, fairness and privacy awareness. In this work, the absence of transparency in the XAI nomenclature is noteworthy but it can be explained as Minh et al use transparency and explainability indistinctly.

As we can see, the terms and classification criteria are similar but there is no unified definition nor specific goals of XAI, because it is usually associated with the efforts and initiatives to establish transparent AI and solve the trust concerns instead of being a standard concept.

6.2 XAI techniques

Taking the previous classification criteria into account, XAI techniques can be organized as follows. The first distinction made in the literature is among models that are interpretable by design—transparent models—and those that can be explained by means of external XAI techniques—post-hoc explainability—.

Transparency can be considered at the level of the entire model, at the level of individual components such as parameters, and at the level of the training algorithm, resulting in the following categorization:

- **Simulatability** denotes the ability of a model of being simulated or thought about strictly by a human. Rule based systems do not fulfill this characteristic whereas a single perceptron neural network does. This aspect aligns with the fact that sparse linear models are more interpretable than dense ones. Providing a decomposable model with simulatability requires that the model has to be self-contained enough for a human to think and reason about it as a whole.
- **Decomposability** is the ability to explain each of the parts of a model (input, parameter and calculation). It requires every input to be readily interpretable. For an algorithmically transparent model to be decomposable, every part of the model must be understandable by a human without the need for additional tools.
- **Algorithmic transparency** deals with the ability of the user to understand the process followed by the model to produce any given output from its input data. The main constraint for algorithmically transparent models is that the model has to be fully explorable by means of mathematical analysis and methods.

These levels of transparency were introduced by Lipton (2018) but they are a real statement in XAI surveys (Barredo Arrieta et al. 2020; Minh et al. 2021).

Post-hoc explainability focuses on models that are not easily interpretable by design and have to enhance their interpretability turning to methods such as:

- **Text explanations** which also include methods generating symbols that represent the functioning of the model. These symbols may describe the logic of the algorithm using a semantic mapping from model to symbols.

- **Visual explanation techniques** that visualize the model's behavior. Many of these techniques rely on dimensionality reduction techniques for a human interpretable visualization.
- **Local explanations** which segment the solution space and give explanations to less complex solution subspaces that are relevant for the whole model.
- **Explanations by example** that extract representative data examples which relate to the result generated by a certain model that reflect the inner relationships and correlations found by the model.
- **Explanations by simplification** in which a new system is rebuilt based on the trained model to be explained. This new model tries to optimize its resemblance to its antecedent functioning, while reducing its complexity, and keeping a similar performance score.
- **Feature relevance explanation** that clarifies the inner functioning of a model by computing a relevance score for its variables. These scores quantify the sensitivity a feature has upon the output of the model.

Among transparent machine learning models are linear/logistic regression, decision trees, k-nearest neighbors, rule-based learning (every model that generates rules to characterize the data it is intended to learn from), general additive models and Bayesian models.

When ML models do not meet any of the criteria imposed to declare them transparent, an independent method must be developed and applied to the model to explain its decisions. These methods are divided between (1) those that are designed to be applied to ML models of any kind, and (2) those that are designed for a specific ML model.

The first class are called *model-agnostic techniques* and are conceived to be joined to any model to obtain some information from its prediction procedure. Taking into account the techniques referenced above, model-agnostic techniques rely on explanation by simplification, feature relevance explanation and visualization techniques. As creating visualizations from just inputs and outputs from an opaque model is a complex task, all visualization methods falling into this category work along with feature relevance techniques, which provide information that is exhibited to the end user.

The second class – model-specific techniques –, is divided into two main branches: those dealing with shallow models which refers to all ML models that do not depend on layered structures of neural processing units; and those developed for deep learning models such as convolutional neural networks, recurrent neural networks, and hybrid schemes of deep neural networks and transparent models. Within shallow ML models, there are strictly interpretable (transparent) approaches as k-nearest neighbors and decision trees, and models that rely on more sophisticated learning algorithms that require additional layers of explanation, as tree ensembles, random forests and Support Vector Machines (SVMs). For tree ensembles, the additional layers of explanation found in the literature are explanation by simplification and feature relevance techniques (Hara and Hayashi 2018). For SVMs, post-hoc explainability techniques applied are explanation by simplification, local explanations, visualizations and explanations by example (Barakat and Bradley 2007; Chen et al. 2007; Gaonkar et al. 2015).

6.3 Explainability in deep learning

It is out of the scope of this revision to describe in detail explainability in all ML models. For this reason, the focus will be on the most complex ML model: deep learning.

Explainability in deep learning is specified through post-hoc local explanations and feature relevance techniques. This section reviews explainability methods proposed for the most used deep learning models, namely multi-layer neural networks, Convolutional Neural Networks (CNN) and Recurrent Neural Networks (RNN). For multi-layer neural networks, the adopted methods seen in the literature are model simplification approaches, feature relevance estimators, text explanations, local explanations and model visualizations (Che et al. 2017; R. Traoré and Caselles-Dupré 2019; Montavon et al. 2017).

CNNs constitute the main models in all computer vision tasks. Their structure entails extremely complex internal relations that are very difficult to explain. Fortunately, including explainability in CNNs is easier than in other types of models, as the human cognitive skills favors the understanding of visual data. Recent works that include explainability in CNNs can be divided into: (1) those that try to understand the decision process by mapping back the output in the input space to see which parts of the input were discriminative for the output; and (2) those that try to delve inside the network and interpret how the intermediate layers see the external world, not necessarily related to any specific input, but in general. The most adopted approach to explainability in CNNs are visualization mixed with feature relevance methods.

For the first class, the most salient techniques are reconstruction of the layers activations occluding regions of an image, or modification of the network architecture (Zeiler et al. 2011; Zeiler and Fergus 2014; Zhou et al. 2016; Selvaraju et al. 2017). For the second class, authors proposed the reconstruction of images from the internal CNN representations (Mahendran and Vedaldi 2015; Nguyen et al. 2016; Bau et al. 2017).

As occurs with CNNs in the visual domain, RNNs have lately been used extensively for predictive problems over inherently sequential data, emphasizing natural language processing and time series analysis. These types of data exhibit long-term dependencies that are too complex to be captured by an ML model. RNNs are able to retrieve such time-dependent relationships by formulating the retention of knowledge in the neuron as another parametric characteristic that can be learned from data. The few contributions made for explaining RNN models can be divided into two groups: (1) explainability by understanding what an RNN model has learned, mainly via feature relevance methods (Arras et al. 2017; Che et al. 2015); and (2) explainability by modifying RNN architectures to provide insights about the decisions they make via local explanations (Krakovna and Doshi-Velez 2016; Choi et al. 2016).

Finally, it is worth mentioning that the use of background knowledge in the form of logical statements or constraints in Knowledge Bases (KBs) has shown to not only improve explainability but also performance with respect to purely data-driven approaches (d'Avila Garcez et al. 2019). This hybrid approach provides robustness to the learning system when errors are present in the training data labels. Other approaches have shown to be able to jointly learn and reason with both symbolic and sub-symbolic representations and inference. The interesting aspect is that this blend allows for expressive probabilistic-logical reasoning in an end-to-end fashion. A successful use case is on dietary recommendations, where explanations are extracted from the reasoning behind (non-deep but KB-based) models (Donadello et al. 2019).

A remarkable perspective on hybrid XAI models consists of enriching black-box models knowledge with that one of transparent ones. In particular, this can be done by constraining the neural network thanks to a semantic KB and bias-prone concepts (Bennetot et al. 2019), or by stacking ensembles jointly encompassing white and black-box models (Loyola-González 2019).

6.4 Guidelines for ensuring interpretable AI models

It is very important to take into account the interests, demands and requirements of the users interacting with a system to be explained, from the designers of the system to the decision makers using its outputs and users going through the consequences of decisions made therefrom.

Given the need for having the human in the loop, some attempts at determining the procedural guidelines to implement and explain AI systems have been recently published. Among them, the study in Leslie (2019) determines that the incorporation and consideration of explainability in practical AI design and deployment workflows should comprise four major methodological steps:

1. **Contextual factors, potential impacts and domain-specific needs must be taken into account when devising an approach to interpretability**. These include a thorough understanding of the purpose for which the AI model is built, the complexity of explanations that are required by the audience, and the performance and interpretability levels of existing technology, models and methods.
2. **Interpretable techniques should be preferred when possible**. When considering explainability in the development of an AI system, the decision of which XAI approach should be chosen should measure domain-specific risks and needs, the available data resources and existing domain knowledge, and the suitability of the ML model to meet the requirements of the computational task to be addressed. It is in the confluence of these three design drivers where the authors in Leslie (2019) recommend first the consideration of standard interpretable models rather than sophisticated yet opaque modeling methods.
3. If a black-box model has been chosen, the third guideline establishes that **ethics-, fairness- and safety-related impacts should be weighted**. Specifically, responsibility in the design and implementation of the AI system should be ensured by checking whether such identified impacts can be mitigated and counteracted by supplementing the system with XAI tools that provide the level of explainability required by the domain in which it is deployed. To this end, the third guideline suggests (1) a detailed articulation, examination and evaluation of the applicable explanatory strategies, (2) the analysis of whether the coverage and scope of the available explanatory approaches match the requirements of the domain and application context where the model is to be deployed; and (3) the formulation of an interpretability action plan that sets forth the explanation delivery strategy, including a detailed time frame for the execution of the plan, and a clearance of the roles and responsibilities of the team involved in the workflow.
4. Finally, the fourth guideline encourages to **rethink interpretability in terms of the cognitive skills, capacities and limitations of the individual human**. This fact is one of the objectives of studies on measures of explainability that consider human mental models, the accesibility of the audience to vocabularies of explanatory outcomes, and other means to involve the expertise of the audience into the decision of what explanations should provide.

7 Usable and useful AI

When we talk about including humans in the machine learning loop, we basically mean including them as part of the learning process. We have seen that this can be done at several levels, humans can be oracles who are responsible for labeling data unknown to the model (AL), they can interact with the model in an active way in the learning process (IML) or they can act as teachers of the model, trying to transfer their domain expertise to the model (MT). Finally, explainable AI, although not part of the learning process itself, is about AI models being able to explain learning outcomes to humans, justifying their conclusions.

But this interactivity with humans should not be limited only to the learning process, if we go further we can see that the relationship of humans with AI continues. Humans become users of AI systems, so they are not only looking for proper technical performance, but also for the system to be easy to use and useful to achieve the objectives they have set.

This leads us to the definition of two new terms related to the relationship between humans and AI models that go beyond cooperation in learning, called "*usable AI*" and "*useful AI*", and which are fundamental to ensure that an AI model is successful.

7.1 Usable AI

Usable AI can be defined as an AI solution that is easy to learn and use via optimal user experience (UX) created by effective Human-Computer Interaction (HCI) design (Xu 2019). We can therefore speak of usability in the learning process and usability in the use of the system itself.

7.1.1 Usable in its learning process

Usability in the learning process starts with data usability. Just because the data is available does not mean that the data is suitable for use in a machine learning process. To do so, they must comply with a series of characteristics, among which we mention (Koesten and Simperl 2021):

- **Usable**: usability in the most limited context, i.e., that we can use them because they are the right size, we have the right permissions, their license allows it, they do not contain sensitive information, etc.
- **Relevant**: that cover the topic of interest at the right level of detail.
- **Quality**: here we would include concepts such as completeness, provenance, accuracy, cleanliness, consistency of formatting, etc.
- **Reusable**: so that they can be used in different studies. Here we would incorporate aspects such as that they are easily understandable, that there are different ways of accessing them, that there is a management of the changes produced in the data, as well as a collaborative nature in the data work processes.

Then we have to take usability into account in the learning process itself. Usability is always context-dependent, and within the context we can highlight the roles that humans play in the learning process, because the tools or interfaces that we design for

them must be appropriate to their level of experience and knowledge. This way we can identify several types of *experts* that can be required at various positions in the loop of the machine learning process:

- **ML experts**: They are experts with extensive knowledge in ML techniques. In supervised learning they select the data, label the data, classify them into training data and testing data, extract the features needed to feed the machine learning algorithm, create the model and refine it if the performance obtained is not optimal, etc. In unsupervised learning, machine learning experts are required to interpret the clusters identified by the model so data can be converted into knowledge.
- **Domain experts**: In many domains, the designers of machine-learning-based systems do not themselves hold the expertise required to create training data. In such projects, the collaboration of *domain experts* is necessary.
- **Data experts**: A data expert or data scientist is a multi-disciplinary scientist that uses methods, processes and algorithms to extract knowledge from data.

As learning processes become more and more interactive, usability within them becomes more and more important.

7.1.2 Usable in its functioning

We have seen in recent years that the performance of AI systems in general, and ML and Deep Learning (DL) algorithms in particular, is achieving (and surpassing in some fields) human level performance. As these systems become more reliable and easier to work with, designers can embed them into products as AI modules allowing the interaction with people (e.g. such as voice recognition or object detection systems).

Due to this adoption of artificial intelligence and machine learning techniques in user-facing products, some authors highlight the interest of the HCI field to discuss the implications of such adoption from different points of view (Churchill et al. 2018).

In previous sections we described how intelligent systems can learn better and deal with unknown situations if they work closely with humans during the learning process. But when the system is deployed in its production environment, the learning process ends and the system is not able to improve anymore its performance, and moreover, it could degrade being no longer valid if the context changes and new information is available. This situation is neglecting one important source of knowledge, the final users.

More and more researchers are realizing the importance of studying users of intelligent systems and how these systems can benefit and learn interactively from their end-users. Once the systems are deployed they can receive from their users corrections that can be used to generate additional training data, enabling an incremental improvement of the AI performance (Lindvall et al. 2018).

A curious approach of learning from users is a system created by von Ahn and Dabbish (2004). In this case the authors developed a two-player guessing game that created labeled training data as a side effect of playing. That is, the users are generating data but to another aim different from the game that they are playing. In this case, the users determined the contents of images by providing meaningful labels for them. This allowed the authors to have proper labels associated with each image that can be later used to perform more accurate image search, improve the accessibility of sites, and help users block inappropriate images.

This is similar to the use of CAPTCHAs when we are asked to identify objects in certain images. CAPTCHA (an acronym of "Completely Automated Public Turing test to tell Computers and Humans Apart") is a type of challenge-response test used in computing to determine whether or not the user is human (von Ahn et al. 2004). O'Malley (2018) stated that "Google has been training AI using the users' responses to CAPTCHAS for years without them knowing that they were doing so".

In all these techniques, the design of HCI plays a key role in achieving intelligent systems that continuously improve through use (Lindvall et al. 2018) and in which the designer's goal is to develop the interaction between the user and the product.

7.2 Useful AI

Other authors have gone beyond the concept of "usability" of AI. For example, van Allen (2018) indicates that we have to take into account not only the development of artificial intelligence or machine learning models but also the design of the interactions and behaviors that compose the human experience around the AI models. Wong (2018) stated that the role of the designer is to contribute a humanist perspective that takes into account social, political, ethical, cultural, and environmental aspects that are not normally associated with AI development but are necessary to include AI into daily human-to-computer interactions.

In this way we can say that AI not only has to be usable, but also has to be useful. Xu (2019) define *Useful AI* as an AI solution that can provide the functions required to satisfy target users' needs in the valid usage scenarios of their work and life.

Useful AI is part of a broader movement known as Human-centered AI (HAI) which refers to approaching AI from a human perspective by considering human conditions and contexts (Yang et al. 2021; Shneiderman 2020).

At the core of useful AI lies the concept of trust, a user of an AI system will always ask himself the question: "Can I trust you?", and this reflexive skepticism directly affects users' trust and decision-making efficiency, thus also affecting the adoption of AI solutions (Xu 2019). This lead us to the concept of "Trustworthy AI".

7.2.1 Trustworthy AI

Because of the black-box effect, some of the AI solutions are not explainable and comprehensible to users. Although we have already introduced Explainable AI (XAI) in this paper, the XAI version created for data scientists is incomprehensible to most non-expert users. The ultimate goal of XAI should be to ensure that target users can understand the outputs, thus helping them improve their decision making efficiency (Xu 2019).

Furthermore, the AI model can be developed with the aim of being self-explanatory so that users can understand why certain decisions are being made. This quality would help to create a trustworthy model for those scenarios where the transparency should be present (e.g.: finance, health, etc.).

Trust is a critical concept in system design because an imperfect AI is likely to be rejected unless a reasonable level of trust is generated between humans and the system. When trust is not generated, people do not accept decisions made by the system. In other words, it is important to design trustworthy interactions between humans and AI to provide a positive user experience. Incorrect or not, users of a system form their trust based on their own sense-making process. By providing an explanation of an AI process, the gap between

users' own sense-making and the actual AI algorithm is reduced and users' understanding increases accordingly, which leads them to trust the AI (Muir 1987).

Related with this issue is the principle of trust by privacy. When creating AI systems, which are fueled by data, privacy and security aspects are an inherent part of the system's life cycle. This maximizes respecting people's right to privacy and their personal data.

In short, we can say that an AI system is trustworthy if it meets the following properties, which come from the world of trustworthy computing and which is open to new additions that may be necessary in the future (Wing 2021):

- **Accuracy**: How well does the AI system do on new (unseen) data compared to data on which it was trained and tested?
- **Robustness**: How sensitive is the system's outcome to a change in the input?
- **Fairness**: Are the system outcomes unbiased?
- **Accountability**: Who or what is responsible for the system's outcome?
- **Transparency**: Is it clear to an external observer how the system's outcome was produced?
- **Explainability**: Can the system's outcome be justified with an explanation that a human can understand and/or that is meaningful to the end user?
- **Ethical**: Was the data collected in an ethical manner? Will the system's outcome be used in an ethical manner?

8 Discussion and conclusions

We have divided the discussion and conclusions into two sections. In the first, we analyze the relationships between the different techniques by analyzing their similarities, differences and mutual influences. In the second section we will analyze the current and future trends in the field of human-in-the-loop machine learning.

8.1 Relationships between the different techniques

In any learning process, two fundamental roles must be considered: the teacher, who wants to teach and provides a set of training examples, and the learner, who wants to learn and estimates an objective function using the set of examples provided by the teacher. In this paper, we have described how humans act as teachers of an ML model in different approaches (i.e., AL, IML, MT) carrying out tasks such as identifying elements and decomposing concepts/features in order to build up more complex ones.

The main difference between these approaches is the behavior of teachers and learners: Does the learner ask questions about what he or she does not know? or is the teacher the one who takes the initiative and provides the most appropriate example at a given time? In other words, the difference lies in who is in control of the learning process.

In the case of **active learning (AL)**, humans act as a teacher who is requested by the learner (i.e., the model) to label examples that are not clear and that will provide relevant information. Therefore, in AL the model remains in control and uses the human as an oracle.

In **interactive machine learning (IML)** there is a closer interaction between users and the learning system, so the control is shared. In this case humans can be assigned to tasks in the ML loop at which they are more efficient than machines, so that they

can play different roles—ML experts, data scientists, crowd-source workers or domain experts—that affect the form and function of the IML systems. Also different methodologies can be established according to the position humans have within the workflow: humans can go to the end of the flow, correcting the results of a machine learning system—e.g., using humans to validate, clean and correct the results—, or humans can act first, performing identification and annotation tasks that are simple for them but complicated for machines—e.g., interactive image segmentation, in which humans provide input with basic annotation tools—.

Finally, in **Machine Teaching (MT)** the control of the learning process relies on human experts that have the aim to transfer their expertise to an intelligent system. Therefore they have to carefully choose the examples they want to transfer to the learner so that the learning process runs smoothly and progresses towards its final goal.

Interactivity is important in AL, IML and MT, but what distinguishes these techniques is not the degree of interactivity, but the intended use of it. In AL the communication is from the learner to the teacher, the system has to be able to display the data in a way that is easy to understand for the teacher and has to follow a questioning strategy that avoids user boredom and frustration, for example, trying not to deliver poor quality or unrepresentative data that would make them lose interest in the system (Mosqueira-Rey et al. 2021). In IML, interactivity depends very much on the type of strategy followed and the objectives to be achieved. As we have seen in the section on IML applications, many of the developments are about using humans to help give structure to unstructured data, such as images, videos, and time-series data. In the case of MT, interactivity is used so that teachers, who are domain experts, do not necessarily have to have ML knowledge. Therefore, they can transfer their domain knowledge to an ML model following an educational strategy similar to the one they would follow if the student were a human (using examples at increasing levels of difficulty, correcting errors and responding proactively to what the student is learning).

This process of sorting the examples by difficulty in MT to make the learning technique more efficient leads us to the so-called **curriculum learning (CL)**. CL is a technique for tidying up the curriculum, and is related to Transfer Learning and Multi-task Learning but is also connected to AL. Both of them involve dynamic data selection but their goals are quite different. AL is designed for a semi-supervised setting in which an active learner improves its performance with fewer labeled data through questions to an expert that annotates unlabeled instances for further training. Therefore, in AL the process focuses on the examples that the teacher could potentially label for gradually adding new examples near the decision border. On the other hand, CL improves performance and accelerates convergence in supervised, weakly-supervised, and unsupervised settings, taking the examples near the decision surface.

This preparation of the curriculum can be seen also in MT in no-batch approaches, first with human teachers because they usually organize the examples in increasing order of difficulty in order to improve the learner's learning, and obviously with non-human teachers that follow an iterative and incremental process preparing the materials in a very similar way to CL but in this case, focusing on identifying the smallest number of examples (or rounds) necessary for learning.

A logical consequence of studying the role of humans in learning tasks is that it should be noted that, in addition to being included in the loop, they can also be at the end of it trying to interpret what the models have learned. This circumstance implies using **Explainable AI (XAI)** methods which propose creating a suite of ML techniques that produce more explainable models and enable humans to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners.

Moreover, as we have also mentioned, adding experts to the learning loop helps to connect some techniques with XAI, since including experts, in principle, favors knowledge improvement. This is especially useful in critical areas such as healthcare, where we seek to avoid the black-box effect of most ML systems (e.g. Deep Learning).

Another logical consequence of the study of the role of humans is that it must be taken into account that to make ML more efficient it is not enough just to include experts in the learning tasks. It also has to be kept in mind that an AI software has to be usable and useful. **Usable** in the sense that it is easy to use, not only during the learning process but also by its end users when interacting with the system. **Useful** in a broad sense, meaning not only offering adequate results, but also reliable ones, taking into consideration features such as robustness, fairness, accountability, transparency, explainability, ethical, etc.

These relationships between the different techniques are summarized in the following Fig. 10, which is nothing more than an adaptation of Fig. 1 to which connections have been added to show graphically how some techniques influence or relate to others.

8.2 Trends and future developments

After analyzing the different approaches to human-in-the-loop machine learning (HITL-ML) together we have identified some trends (Mosqueira-Rey et al. 2022) and future developments in the field.

The first trend is that interactivity has increasing importance in the development of ML models, because as we move towards more human control we also move towards to more interactivity. For example, when in AL a learner requests a human oracle to label examples, the questions must be presented in a way that the human understands them. This implies that the characteristics of usability in the learning process (clarity, consistency, efficiency, etc.) is particularly relevant. As interactivity increases, these aspects become more important to the point that, in the case of MT, the measurement of metrics, such as productivity, interpretability, robustness, and scaling is considered essential to check whether a system is successful.

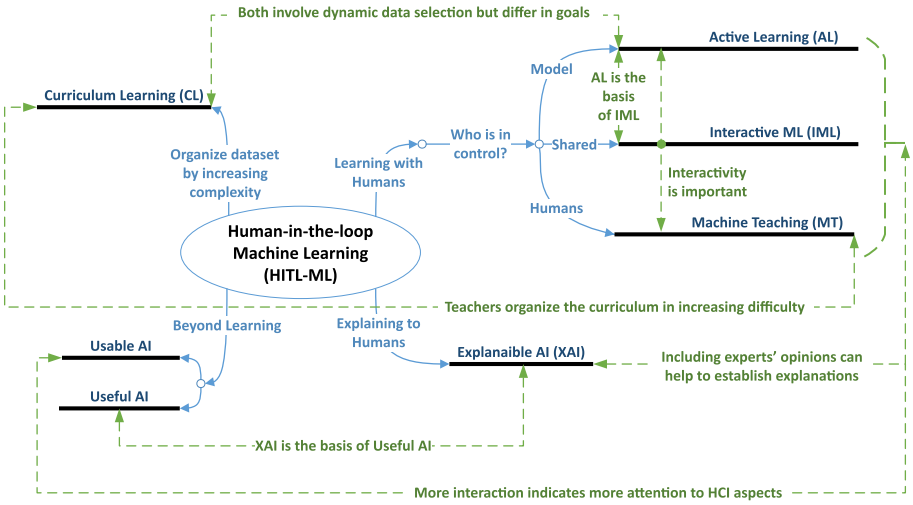

Fig. 10 Human-in-the-loop machine learning (HITL-ML-relations) mind map

Another trend that can be seen in the integration of humans into the ML learning loop is **a trend toward the automation of the majority of the tasks carried out**. The idea pursued here is that an ML model can be created by someone without (or with little) knowledge of ML. We can see this idea behind the concept of AutoML (He et al. 2021), a set of tools that seek to automate the decision on what learning algorithms to use, what hyper-parameters to select, or which features are more relevant for a certain model, providing means of model evaluation and optimization. These tools are often based on cloud developments, offered through a "Software as a service" (SaaS) model, in this case renamed as "machine learning as a Service (MLaaS)" (Ribeiro et al. 2015). The main advantage of the cloud is that it provides a platform allowing users to focus on the problem itself, without having to worry about the infrastructure. The concept of MLaaS evolved to MLOps (Treveil et al. 2020), that is, applying the same principles of DevOps to machine learning, which led to the emergence of automated data management, model training/deployment, and monitoring. As we can see, the tools available to researchers are moving away from the ad-hoc and experimental approach to a more engineering perspective (Mosqueira-Rey et al. 2022).

The final trend identified is to put the focus of attention on the domain-specific problem and not so much on the technique needed to implement it. The MLaaS and MLOps approaches require ML expertise, and if you are offering ML-specific information you are not really acting as a teacher, because you are not transferring knowledge about the topic but about the technique. MT aims to go a step further, the idea is to follow an human-centered approach, where the teacher is a domain expert that designs an ML model without ML knowledge. In this way, expert knowledge can be transferred directly to the machine. The rationale here is that a teacher gives information about labels but also semantic information about why these labels are used, as well as assessing performance. In other words, the goal is to take advantage of the abilities we humans have when it comes to sharing knowledge among us, and using them to transfer knowledge to a machine.

As a final thought we consider that the inclusion of humans in the loop of ML, and the concepts of usability and usefulness in AI software has led to the emergence of a broader movement known as Human-centered AI (HAI) (Xu 2019) which refers to approaching AI from a human perspective by considering human conditions and contexts. In this context, it should be considered that the first two waves of AI failed not only due to the lack of mature technologies, but also because they left human needs unsatisfied. AI is starting to satisfy them and provide a positive user experience (UX) for a variety of application scenarios in the third wave. It is also starting to deliver mature business models with useful AI in which people started to consider the inclusion of human aspects such as ethics, interpretability, fairness, etc. Thus, the third wave of AI can be characterized by its technological improvement but also by its human-centered approach.

In short, we can say that the next frontier of AI is not only technological but also humanistic and ethical, which opens up a wide range of research lines in this field, such as those exploring the concept of Trustworthy AI.

Acknowledgements This work has been supported by the State Research Agency of the Spanish Government, Grant (PID2019-107194GB-I00/AEI/10.13039/501100011033) and by the Xunta de Galicia, Grant (ED431C 2022/44) with the European Union ERDF funds. We wish to acknowledge the support received from the Centro de Investigación de Galicia "CITIC", funded by Xunta de Galicia and the European Union (European Regional Development Fund- Galicia 2014-2020 Program), by Grant ED431G 2019/01.

Funding Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

Declarations

Conflict of interest The authors declare that they have no conflict of interest regarding the publication of this article.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

Abdul A, Vermeulen J, Wang D et al (2018) Trends and trajectories for explainable, accountable and intelligible systems: an hci research agenda. In: Proceedings of the 2018 CHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI '18, pp 1–18, https://doi.org/10.1145/3173574.3174156

Abiteboul S, Buneman P, Suciu D (2000) Data on the web: from relations to semistructured data and XML. Morgan Kaufmann, Data Management Systems Series

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52,138-52,160. https://doi.org/10.1109/ACCESS.2018.2870052

Aggarwal CC, Kong X, Gu Q et al (2014) Active learning: a survey. Data classification: algorithms and applications. Chapman and Hall/CRC, Boca Raton, pp 599–634

Amazon (2022) Amazon mechanical turk. https://www.mturk.com/. Accessed on 23 Mar 2022

Amershi S, Cakmak M, Knox WB et al (2014) Power to the people: the role of humans in interactive machine learning. AI Magazine 35(4):105–120. https://doi.org/10.1609/aimag.v35i4.2513

Angluin D (1987) Learning regular sets from queries and counterexamples. Inf Comput 75(2):87–106. https://doi.org/10.1016/0890-5401(87)90052-6

Arras GL, Montavon, Müller KR, Samek W (2017) Explaining recurrent neural network predictions in sentiment analysis. In: EMNLP'17 workshop on computational approaches to subjectivity, sentiment and social media analysis, https://doi.org/10.18653/v1/W17-5221

Barakat NH, Bradley AP (2007) Rule extraction from support vector machines: a sequential covering approach. IEEE Trans Knowl Data Eng 19(6):729–741. https://doi.org/10.1109/TKDE.2007.190610

Barredo Arrieta A, Díaz-Rodríguez N, Del Ser J et al (2020) Explainable artificial intelligence (xai): concepts, taxonomies, opportunities and challenges toward responsible AI. Inf Fusion 58:82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Bau D, Zhou B, Khosla A et al (2017) Network dissection: Quantifying interpretability of deep visual representations. In: 2017 IEEE conference on computer vision and pattern recognition (CVPR), 3319–3327, https://doi.org/10.1109/CVPR.2017.354, https://ieeexplore.ieee.org/document/8099837

Begeja L, Renger B, Gibbon D et al (2004) Interactive machine learning techniques for improving SLU models. In: Proceedings of the HLT-NAACL 2004 workshop on spoken language understanding for conversational systems and higher level linguistic information for speech processing. Association for Computational Linguistics, Boston, Massachusetts, USA, 10–16, https://aclanthology.org/W04-3003

Bengio Y, Louradour J, Collobert R et al (2009) Curriculum learning. In: Proceedings of the 26th annual international conference on machine learning. Association for Computing Machinery, New York, NY, USA, ICML '09, 41-48, https://doi.org/10.1145/1553374.1553380

Bennetot A, Laurent JL, Chatila R et al (2019) Towards explainable neural-symbolic visual reasoning. arxiv:1909.09065

Berg S, Kutra D, Kroeger T et al (2019) Ilastik: interactive machine learning for (bio)image analysis. Nat Methods 16(12):1226–1232. https://doi.org/10.1038/s41592-019-0582-9

Berghel H (1997) Cyberspace 2000: dealing with information overload. Commun ACM 40(2):19–24. https://doi.org/10.1145/253671.253680

Blumberg R, Atre S (2003) The problem with unstructured data. DM Rev 13(42–49):62

Bonwell CC, Eison JA (1991) Active learning: creating excitement in the classroom. 1991 ASHE-ERIC higher education reports. ERIC Clearinghouse on Higher Education, The George Washington University, One Dupont Circle, Suite 630, Washington, DC 20036-1183

Boukhelifa N, Bezerianos A, Lutton E (2018) Evaluation of interactive machine learning systems. In: Zhou J, Chen F (eds) Human and machine learning: visible, explainable, trustworthy and transparent. Springer, Cham, pp 341–360

Budd S, Robinson EC, Kainz B (2021) A survey on active learning and human-in-the-loop deep learning for medical image analysis. Med Image Anal 71(102):062. https://doi.org/10.1016/j.media.2021.102062

Carlson G (2015) What eactly is complex data? https://www.ayasdi.com/exactly-complex-data/. Accessed on 04 Mar 2021

Castle E (2017) 7 signs you're dealing with complex data. https://www.sisense.com/blog/7-signs-youre-dealing-with-complex-data/. Accessed on 04 Mar 2022

Che Z, Purushotham S, Khemani R et al (2015) Distilling knowledge from deep networks with applications to healthcare domain. arXiv e-prints arxiv:1512.03542 [stat.ML]

Che Z, Purushotham S, Khemani R et al (2017) Interpretable deep models for ICU outcome prediction. In: AMIA annual symposium proceedings, 371–380, https://pubmed.ncbi.nlm.nih.gov/28269832/

Chen Z, Li J, Wei L (2007) A multiple kernel support vector machine scheme for feature selection and rule extraction from gene expression data of cancer tissue. Artif Intell Med 41(2):161–175. https://doi.org/10.1016/j.artmed.2007.07.008

Chen Y, Singla A, Aodha OM et al (2018) Understanding the role of adaptivity in machine teaching: The case of version space learners. In: Proceedings of the 32nd international conference on neural information processing systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'18, 1483-1493, https://dl.acm.org/doi/abs/10.5555/3326943.3327079

Choi E, Bahadori T, Schuetz A et al (2016) Retain: Interpretable predictive model in healthcare using reverse time attention mechanism. In: Proceedings of the 30th international conference on neural information processing systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'16, 3512-3520

Churchill EF, van Allen P, Kuniavsky M (2018) Designing AI. Interactions 25(6):34–37. https://doi.org/10.1145/3281764

Cirik V, Hovy E, Morency LP (2016) Visualizing and understanding curriculum learning for long short-term memory networks. arXiv e-prints arxiv:1611.06204 [cs.CL]

Cohn D, Atlas L, Ladner R (1994) Improving generalization with active learning. Mach Learn 15(2):201–221. https://doi.org/10.1007/BF00993277

Cohn DA, Ghahramani Z, Jordan MI (1996) Active learning with statistical models. J Artif Intell Res 4(1):129–145. https://doi.org/10.5555/1622737.1622744

d'Avila Garcez A, Gori M, Lamb LC et al (2019) Neural-symbolic computing: an effective methodology for principled integration of machine learning and reasoning. arXiv e-prints arxiv:1905.06088 [cs.AI]

Das A, Rad P (2020) Opportunities and challenges in explainable artificial intelligence (XAI): a survey. arXiv e-prints arxiv:2006.11371 [cs.CV]

De Angeli K, Gao S, Alawad M et al (2021) Deep active learning for classifying cancer pathology reports. BMC Bioinform 22(1):1–25

Devidze R, Mansouri F, Haug L et al (2020) Understanding the power and limitations of teaching with imperfect knowledge. In: Bessiere C (ed) Proceedings of the twenty-ninth international joint conference on artificial intelligence, IJCAI-20. International Joint Conferences on Artificial Intelligence Organization, 2647–2654, https://doi.org/10.24963/ijcai.2020/367

Diamant E (2009) Machine learning: When and where the horses went astray? In: Zhang Y (ed) Machine learning. InTech, London, pp 1–18. https://doi.org/10.5772/9156

Diamant E (2006) Learning to understand image content: Machine learning versus machine teaching alternative. In: 2006 International conference on information technology: research and education, 26–29, https://doi.org/10.1109/ITRE.2006.381526

Donadello I, Kessler F, Dragoni M et al (2019) Persuasive explanation of reasoning inferences on dietary data. In: Joint proceedings of the 6th international workshop on dataset profilling and search

and the 1st workshop on semantic explainability co-located with the 18th international semantic web conference (ISWC 2019)
Donmez P, Carbonell JG (2008) Proactive learning: Cost-sensitive active learning with multiple imperfect oracles. In: Proceedings of the 17th ACM conference on information and knowledge management. Association for Computing Machinery, New York, NY, USA, CIKM '08, 619–628, https://doi.org/10.1145/1458082.1458165
Dudley JJ, Kristensson PO (2018) A review of user interface design for interactive machine learning. ACM Trans Interact Intell Syst. https://doi.org/10.1145/3185517
El-Hasnony IM, Elzeki OM, Alshehri A et al (2022) Multi-label active learning-based machine learning model for heart disease prediction. Sensors. https://doi.org/10.3390/s22031184
Elman JL (1993) Learning and development in neural networks: the importance of starting small. Cognition 48(1):71–99. https://doi.org/10.1016/0010-0277(93)90058-4
Fadhil A, Wang Y (2018) Towards automatic & personalised mobile health interventions: an interactive machine learning perspective. arXiv e-prints arxiv:1803.01842 [cs.CY]
Fails JA, Olsen DR (2003) Interactive machine learning. In: Proceedings of the 8th international conference on intelligent user interfaces. Association for Computing Machinery, New York, NY, USA, IUI '03, 39–45, https://doi.org/10.1145/604045.604056
Fiebrink RA (2011) Real-time human interaction with supervised learning algorithms for music composition and performance. PhD thesis, Computer Science Dept. Princeton University, Princeton, NJ, USA, https://dl.acm.org/doi/book/10.5555/2125776
Fiebrink R, Cook PR (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 2010), Utrecht
Fiebrink R, Cook PR, Trueman D (2011) Human model evaluation in interactive supervised learning. In: Proceedings of the SIGCHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI '11, 147–156, https://doi.org/10.1145/1978942.1978965
Florensa C, Held D, Wulfmeier M et al (2017) Reverse curriculum generation for reinforcement learning. In: Levine S, Vanhoucke V, Goldberg K (eds) Proceedings of the 1st annual conference on robot learning, proceedings of machine learning research, vol 78. PMLR, 482–495, http://proceedings.mlr.press/v78/florensa17a.html
Fogarty J, Tan D, Kapoor A et al (2008) Cueflik: Interactive concept learning in image search. In: Proceedings of the SIGCHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI '08, 29–38, https://doi.org/10.1145/1357054.1357061
Gaonkar B, Shinohara TR, Davatzikos C (2015) Interpreting support vector machine models for multivariate group wise analysis in neuroimaging. Med Image Anal 24(1):190–204. https://doi.org/10.1016/j.media.2015.06.008
Goodman B, Flaxman S (2017) European union regulations on algorithmic decision-making and a "right to explanation." AI Mag 38(3):50–57. https://doi.org/10.1609/aimag.v38i3.2741
Gunning D (2017) Explainable artificial intelligence (xAI). Tech. rep., Defense Advanced Research Projects Agency (DARPA), https://www.darpa.mil/program/explainable-artificial-intelligence
Hacohen G, Weinshall D (2019) On the power of curriculum learning in training deep networks. In: Chaudhuri K, Salakhutdinov R (eds) Proceedings of the 36th international conference on machine learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA, Proceedings of Machine Learning Research, vol 97. PMLR, 2535–2544, http://proceedings.mlr.press/v97/hacohen19a.html
Hara S, Hayashi K (2018) Making tree ensembles interpretable: A bayesian model selection approach. In: Storkey A, Perez-Cruz F (eds) Proceedings of the twenty-first international conference on artificial intelligence and statistics, proceedings of machine learning research, vol 84. PMLR, 77–85, https://proceedings.mlr.press/v84/hara18a.html
He X, Zhao K, Chu X (2021) AutoML: a survey of the state-of-the-art. Knowl Based Syst 212(106):622. https://doi.org/10.1016/j.knosys.2020.106622
Heimerl F, Koch S, Bosch H et al (2012) Visual classifier training for text document retrieval. IEEE Trans Vis Comput Graphics 18(12):2839–2848. https://doi.org/10.1109/TVCG.2012.277
Hills TT, Todd PM, Lazer D et al (2015) Exploration versus exploitation in space, mind, and society. Trends Cogn Sci 19(1):46–54. https://doi.org/10.1016/j.tics.2014.10.004
Hipke K, Toomim M, Fiebrink R et al (2014) Beatbox: End-user interactive definition and training of recognizers for percussive vocalizations. In: Proceedings of the 2014 international working conference on advanced visual interfaces. Association for Computing Machinery, New York, NY, USA, AVI '14, 121–124, https://doi.org/10.1145/2598153.2598189

Hoi SCH, Jin R, Zhu J et al (2006) Batch mode active learning and its application to medical image classification. In: Proceedings of the 23rd international conference on machine learning. Association for Computing Machinery, New York, NY, USA, ICML '06, 417-424, https://doi.org/10.1145/1143844.1143897

Holmberg L, Davidsson P, Linde P (2020) A feature space focus in machine teaching. In: 2020 IEEE international conference on pervasive computing and communications workshops (PerCom Workshops), 1–2, https://doi.org/10.1109/PerComWorkshops48775.2020.9156175

Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inform 3(2):119–131. https://doi.org/10.1007/s40708-016-0042-6

Holzinger A, Jurisica I (2014) Knowledge discovery and data mining in biomedical informatics: the future is in integrative, interactive machine learning solutions. In: Holzinger A, Jurisica I (eds) Interactive knowledge discovery and data mining in biomedical informatics: state-of-the-art and future challenges. Springer, Berlin, Heidelberg, pp 1–18. https://doi.org/10.1007/978-3-662-43968-5_1

Holzinger A, Plass M, Kickmeier-Rust M et al (2019) Interactive machine learning: experimental evidence for the human in the algorithmic loop. Appl Intell 49(7):2401–2414. https://doi.org/10.1007/s10489-018-1361-5

Holzinger A, Biemann C, Pattichis CS, et al (2017) What do we need to build explainable AI systems for the medical domain? arXiv e-prints arxiv:1712.09923 [cs.AI]

Ionescu RT, Alexe B, Leordeanu M et al (2016) How hard can it be? estimating the difficulty of visual search in an image. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), pp 2157–2166, https://doi.org/10.1109/CVPR.2016.237

Ishibashi T, Nakao Y, Sugano Y (2020) Investigating audio data visualization for interactive sound recognition. In: Proceedings of the 25th international conference on intelligent user interfaces. Association for Computing Machinery, New York, NY, USA, IUI '20, 67–77, https://doi.org/10.1145/3377325.3377483

Jamieson KG, Jain L, Fernandez C et al (2015) Next: a system for real-world development, evaluation, and application of active learning. In: Cortes C, Lawrence N, Lee D et al (eds) Advances in neural information processing systems, vol 28. Curran Associates Inc, Red Hook

Jiang L, Meng D, Zhao Q et al (2015) Self-paced curriculum learning. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence. AAAI Press, AAAI'15, 2694-2700, https://doi.org/10.5555/2886521.2886696

Jiang L, Liu S, Chen C (2019) Recent research advances on interactive machine learning. J Vis 22(2):401–417. https://doi.org/10.1007/s12650-018-0531-1

Johns E, Mac Aodha O, Brostow GJ (2015) Becoming the expert-interactive multi-class machine teaching. In: Proceedings of the IEEE conference on computer vision and pattern recognition, 2616–2624, https://doi.org/10.1109/CVPR.2015.7298877

Kabra M, Robie AA, Rivera-Alba M et al (2013) Jaaba: interactive machine learning for automatic annotation of animal behavior. Nat Methods 10(1):64–67. https://doi.org/10.1038/nmeth.2281

Kapoor A, Lee B, Tan D et al (2010) Interactive optimization for steering machine classification. In: Proceedings of the SIGCHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI '10, 1343–1352, https://doi.org/10.1145/1753326.1753529

Kellenberger B, Tuia D, Morris D (2020) Aide: accelerating image-based ecological surveys with interactive machine learning. Methods Ecol Evol 11(12):1716–1727. https://doi.org/10.1111/2041-210X.13489

Kim B, Patel K, Rostamizadeh A et al (2015) Scalable and interpretable data representation for high-dimensional, complex data. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence, Association for the Advancement of Artificial Intelligence (AAAI), Austin, Texas USA, 1763–1769, https://ojs.aaai.org/index.php/AAAI/article/view/9474

Koesten L, Simperl E (2021) Ux of data: making data available doesn't make it usable. Interactions 28(2):97–99. https://doi.org/10.1145/3448888

Kosmyna N, Tarpin-Bernard F, Rivet B (2015) Adding human learning in brain-computer interfaces (bcis): towards a practical control modality. ACM Trans Comput-Hum Interact. https://doi.org/10.1145/2723162

Krakovna V, Doshi-Velez F (2016) Increasing the interpretability of recurrent neural networks using hidden markov models. arXiv e-prints arxiv:1606.0532 [cond-mat.soft]

Kumar M, Packer B, Koller D (2010) Self-paced learning for latent variable models. In: Lafferty J, Williams C, Shawe-Taylor J et al (eds) Advances in neural information processing systems. Curran Associates Inc, Red Hook, pp 1189–1197

Kumar G, Foster G, Cherry C et al (2019) Reinforcement learning based curriculum optimization for neural machine translation. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 2054–2061, https://doi.org/10.18653/v1/N19-1208, https://www.aclweb.org/anthology/N19-1208

Laws F, Scheible C, Schütze H (2011) Active Learning with Amazon Mechanical Turk. In: Proceedings of the conference on empirical methods in natural language processing. Association for Computational Linguistics, USA, EMNLP '11, 1546-1556, https://doi.org/10.5555/2145432.2145597

LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444. https://doi.org/10.1038/nature14539

Leslie D (2019) Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of ai systems in the public sector. 10.5281/zenodo.3240529

Lindvall M, Molin J, Löwgren J (2018) From machine learning to machine teaching: the importance of UX. Interactions 25(6):52–57. https://doi.org/10.1145/3282860

Lipton ZC (2018) The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. Queue 16(3):31–57. https://doi.org/10.1145/3236386.3241340

Liu W, Dai B, Humayun A et al (2017) Iterative machine teaching. In: Precup D, Teh YW (eds) Proceedings of the 34th international conference on machine learning, proceedings of machine learning research, vol 70. PMLR, 2149–2158, https://proceedings.mlr.press/v70/liu17b.html

Liu C, He S, Liu K et al (2018a) Curriculum learning for natural answer generation. In: Proceedings of the 27th international joint conference on artificial intelligence. AAAI Press, IJCAI'18, 4223-4229, https://doi.org/10.24963/ijcai.2018/587

Liu J, Lichtenberg T, Hoadley KA et al (2018b) An integrated TCGA pan-cancer clinical data resource to drive high-quality survival outcome analytics. Cell 173(2):400-416.e11. https://doi.org/10.1016/j.cell.2018.02.052

Liu W, Dai B, Li X et al (2018c) Towards black-box iterative machine teaching. In: Dy J, Krause A (eds) Proceedings of the 35th international conference on machine learning, proceedings of machine learning research, vol 80. PMLR, 3141–3149, https://proceedings.mlr.press/v80/liu18b.html

Liu Z, Feng X, Wang Y et al (2021) Self-paced learning enhanced neural matrix factorization for noise-aware recommendation. Knowl Based Syst 213(106):660. https://doi.org/10.1016/j.knosys.2020.106660

Lopes M, Melo F, Montesano L (2009) Active learning for reward estimation in inverse reinforcement learning. Joint European conference on machine learning and knowledge discovery in databases. Springer, Berlin Heidelberg, pp 31–46

Loyola-González O (2019) Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view. IEEE Access 7:154096–154113. https://doi.org/10.1109/ACCESS.2019.2949286

Luo T, Kramer K, Samson S et al (2004) Active learning to recognize multiple types of plankton. In: Proceedings of the 17th international conference on pattern recognition, 2004. ICPR 2004., 478–481 Vol.3, https://doi.org/10.1109/ICPR.2004.1334570

Mahendran A, Vedaldi A (2015) Understanding deep image representations by inverting them. In: 2015 IEEE conference on computer vision and pattern recognition (CVPR), 5188–5196, https://doi.org/10.1109/CVPR.2015.7299155, https://ieeexplore.ieee.org/document/7299155

Matiisen T, Oliver A, Cohen T et al (2020) Teacher-student curriculum learning. IEEE Trans Neural Netw Learn Syst 31(9):3732–3740. https://doi.org/10.1109/TNNLS.2019.2934906

Mei S, Zhu X (2015) Using machine teaching to identify optimal training-set attacks on machine learners. In: Proc. of the 29th AAAI conference on artificial intelligence, 2871–2877, https://ojs.aaai.org/index.php/AAAI/article/view/9569

Meske C, Bunde E, Schneider J et al (2022) Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. Inf Syst Manag 39(1):53–63. https://doi.org/10.1080/10580530.2020.1849465

Meza Martínez MA, Nadj M, Maedche A (2019) Towards an integrative theoretical framework of interactive machine learning systems. In: Proceedings of the 27th European conference on information systems (ECIS), Stockholm & Uppsala, Sweden, https://aisel.aisnet.org/ecis2019_rp/172

Michael CJ, Acklin D, Scheuerman J (2020) On interactive machine learning and the potential of cognitive feedback. arXiv e-prints arxiv:2003.10365 [cs.HC]

Microsoft (2022) Qna maker. https://www.qnamaker.ai/. Accessed on 23 Mar 2022

Minh D, Wang HX, Li YF et al (2021) Explainable artificial intelligence: a comprehensive review. Artif Intell Rev. https://doi.org/10.1007/s10462-021-10088-y

Montavon G, Lapuschkin S, Binder A et al (2017) Explaining nonlinear classification decisions with deep Taylor decomposition. Pattern Recognit 65:211–222. https://doi.org/10.1016/j.patcog.2016.11.008

Mosqueira-Rey E, Alonso-Ríos D, Baamonde-Lozano A (2021) Integrating iterative machine teaching and active learning into the machine learning loop. Procedia Comput Sci 192:553–562. https://doi.org/10.1016/j.procs.2021.08.057

Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D et al (2022) A classification and review of tools for developing and interacting with machine learning systems. In: Proceedings of the 37th annual ACM symposium on applied computing. Association for Computing Machinery, New York, NY, USA, 1083–1092, https://doi.org/10.1145/3477314.3507310

Muir BM (1987) Trust between humans and machines, and the design of decision aids. Int J Man-Mach Stud 27(5):527–539. https://doi.org/10.1016/S0020-7373(87)80013-5

Munro R (2020) Human-in-the-loop machine learning. Manning Publications, Shelter Island

Nguyen DHM, Patrick JD (2014) Supervised machine learning and active learning in classification of radiology reports. J Am Med Inform Assoc 21(5):893–901. https://doi.org/10.1136/amiajnl-2013-002516

Nguyen A, Dosovitskiy A, Yosinski J et al (2016) Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Proceedings of the 30th international conference on neural information processing systems. Curran Associates Inc., Red Hook, NY, USA, NIPS'16, 3395-3403, https://doi.org/10.5555/3157382.3157477

Nwana HS (1990) Intelligent tutoring systems: an overview. Artif Intell Rev 4(4):251–277. https://doi.org/10.1007/BF00168958

Olsson F (2009) A literature survey of active machine learning in the context of natural language processing. Tech. rep., Swedish Institute of Computer Science, http://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-23510

O'Malley J (2018) Captcha if you can: how you've been training ai for years without realising it. https://www.techradar.com/news/captcha-if-you-can-how-youve-been-training-ai-for-years-without-realising-it

Peng B, Li C, Li J et al (2021) Soloist: building task bots at scale with transfer learning and machine teaching. Trans Assoc Comput Linguist 9:807–824. https://doi.org/10.1162/tacl_a_00399

Penha G, Hauff C (2020) Curriculum learning strategies for IR. In: Jose JM, Yilmaz E, Magalhães J et al (eds) European conference on information retrieval: advances in information retrieval. Springer, Cham, pp 699–713. https://doi.org/10.1007/978-3-030-45439-5_46

Platanios EA, Stretcu O, Neubig G et al (2019) Competence-based curriculum learning for neural machine translation. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, pp 1162–1172, https://doi.org/10.18653/v1/N19-1119, https://www.aclweb.org/anthology/N19-1119

Porter R, Theiler J, Hush D (2013) Interactive machine learning in data exploitation. Comput Sci Eng 15(5):12–20. https://doi.org/10.1109/MCSE.2013.74

Ramos G, Meek C, Simard P et al (2020) Interactive machine teaching: a human-centered approach to building machine-learned models. Hum Comput Interact 35(5–6):413–451. https://doi.org/10.1080/07370024.2020.1734931

Reyes O, Pérez E, del Carmen Rodrıguez-Hernández M et al (2016) Jclal: a java framework for active learning. J Mach Learn Res 17:1–5

Ribeiro M, Grolinger K, Capretz MA (2015) MLaaS: Machine learning as a service. In: 2015 IEEE 14th international conference on machine learning and applications (ICMLA), 896–902, https://doi.org/10.1109/ICMLA.2015.152

Rubens N, Elahi M, Sugiyama M et al (2015) Active learning in recommender systems. In: Ricci F, Rokach L, Shapira B (eds) Recommender systems handbook. Springer, Boston, pp 809–846

Rusu O, Halcu I, Grigoriu O et al (2013) Converting unstructured and semi-structured data into knowledge. In: 013 11th RoEduNet international conference, 1–4, https://doi.org/10.1109/RoEduNet.2013.6511736

Sammut C, Banerji RB (1986) Learning concepts by asking questions. In: Michalski RS, Carbonell J, Mitchell T (eds) Machine learning: an artificial intelligence approach, vol 2. Morgan Kaufmann, Burlington, pp 167–192

Šavelka J, Trivedi G, Ashley KD (2015) Applying an interactive machine learning approach to statutory analysis. In: Rotolo A (ed) Legal knowledge and information systems, frontiers in artificial intelligence and applications, vol 279. IOS Press, Amsterdam, pp 101–110. https://doi.org/10.3233/978-1-61499-609-5-101

Selvaraju RR, Cogswell M, Das A et al (2017) Grad-cam: Visual explanations from deep networks via gradient-based localization. In: 2017 IEEE international conference on computer vision (ICCV), 618–626, https://doi.org/10.1109/ICCV.2017.74, https://ieeexplore.ieee.org/document/8237336

Sena A, Howard M (2020) Quantifying teaching behavior in robot learning from demonstration. Int J Robot Res 39(1):54–72. https://doi.org/10.1177/0278364919884623

Sena A, Zhao Y, Howard MJ (2018) Teaching human teachers to teach robot learners. In: 2018 IEEE international conference on robotics and automation (ICRA), 5675–5681, https://doi.org/10.1109/ICRA.2018.8461194

Settles B (2009) Active learning literature survey. Tech. rep., University of Wisconsin-Madison. Department of Computer Sciences, https://minds.wisconsin.edu/handle/1793/60660

Settles B (2011) From theories to queries: Active learning in practice. In: Guyon I, Cawley G, Dror G et al (eds) Active learning and experimental design workshop In conjunction with AISTATS 2010, proceedings of machine learning research, vol 16. JMLR workshop and conference proceedings, Sardinia, Italy, 1–18, http://proceedings.mlr.press/v16/settles11a.html

Shneiderman B (2020) Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Hum Comput Interact 36(6):495–504. https://doi.org/10.1080/10447318.2020.1741118

Simard PY, Amershi S, Chickering DM et al (2017) Machine teaching: A new paradigm for building machine learning systems. arXiv e-prints arxiv:1707.06742

Singla A, Bogunovic I, Bartók G et al (2014) Near-optimally teaching the crowd to classify. In: Xing EP, Jebara T (eds) Proceedings of the 31st international conference on machine learning. PMLR, Bejing, China, proceedings of machine learning research, 154–162, http://proceedings.mlr.press/v32/singla14.pdf

Sint R, Schaffert S, Stroka S et al (2009) Combining unstructured, fully structured and semi-structured information in semantic wikis. In: 4th semantic wiki workshop (SemWiki 2009) at the 6th European semantic web conference (ESWC 2009), Hersonissos, Greece, 73–87, http://ceur-ws.org/Vol-464/paper-14.pdf

Smith JS, Nebgen B, Lubbers N et al (2018) Less is more: sampling chemical space with active learning. J Chem Phys 148(24):241,733

Soviany P, Ardei C, Ionescu RT et al (2020) Image difficulty curriculum for generative adversarial networks (cugan). In: 2020 IEEE winter conference on applications of computer vision (WACV), 3452–3461, https://doi.org/10.1109/WACV45572.2020.9093408

Spitkovsky VI, Alshawi H, Jurafsky D (2010) From baby steps to leapfrog: How "less is more" in unsupervised dependency parsing. In: Human language technologies: the 2010 annual conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, Los Angeles, California, 751–759, https://www.aclweb.org/anthology/N10-1116

Suh J, Ghorashi S, Ramos G et al (2019) Anchorviz: facilitating semantic data exploration and concept discovery for interactive machine learning. ACM Trans Interact Intell Syst. https://doi.org/10.1145/3241379

Talbot J, Lee B, Kapoor A et al (2009) Ensemblematrix: Interactive visualization to support machine learning with multiple classifiers. In: Proceedings of the SIGCHI conference on human factors in computing systems. Association for Computing Machinery, New York, NY, USA, CHI '09, 1283-1292, https://doi.org/10.1145/1518701.1518895

Tang YP, Li GX, Huang SJ (2019) ALiPy: Active learning in python. Tech. rep., Nanjing University of Aeronautics and Astronautics, https://github.com/NUAA-AL/ALiPy, available as arXiv preprint arxiv:1901.03802

Teso S, Kersting K (2019) Explanatory interactive machine learning. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society. Association for Computing Machinery, New York, NY, USA, AIES '19, 239–245, https://doi.org/10.1145/3306618.3314293

Tolls V (2018) An event-based approach to modeling complex data in critical care. PhD thesis, Queen's University (Canada), https://qspace.library.queensu.ca/bitstream/handle/1974/24489/Tolls_Victoria_J_201809_MSC.pdf

Tomczak K, Czerwińska P, Wiznerowicz M (2015) The cancer genome atlas (TCGA): an immeasurable source of knowledge. Contemp Oncol 19(1A):68–77. https://doi.org/10.5114/wo.2014.47136

Treveil M, Omont N, Stenac C et al (2020) Introducing MLOps. O'Reilly Media, Sebastopol

Traoré R, Caselles-Dupré ea (2019) Discorl: continual reinforcement learning via policy distillation. arXiv e-prints arxiv:1907.05855 [cs.LG]

Valiant LG (1984) A theory of the learnable. Commun ACM 27(11):1134–1142. https://doi.org/10.1145/1968.1972

van Allen P (2018) Prototyping ways of prototyping AI. Interactions 25(6):46–51. https://doi.org/10.1145/3274566

von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of the SIGCHI conference on human factors in computing systems. ACM, New York, NY, USA, CHI '04, 319–326, https://doi.org/10.1145/985692.985733

von Ahn L, Blum M, Langford J (2004) Telling humans and computers apart automatically. Commun ACM 47(2):56–60. https://doi.org/10.1145/966389.966390

Visi FG, Tanaka A (2021) Interactive machine learning of musical gesture. In: Miranda ER (ed) Handbook of artificial intelligence for music: foundations, advanced approaches, and developments for creativity. Springer, Cham, pp 771–798. https://doi.org/10.1007/978-3-030-72116-9_27

Wall E, Ghorashi S, Ramos G (2019) Using expert patterns in assisted interactive machine learning: a study in machine teaching. In: Lamas D, Loizides F, Nacke L et al (eds) Human-computer interaction—INTERACT 2019. Springer, Berlin, pp 578–599. https://doi.org/10.1007/978-3-030-29387-1_34

Wallace BC, Small K, Brodley CE et al (2012) Deploying an interactive machine learning system in an evidence-based practice center: Abstrackr. In: Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium. Association for Computing Machinery, New York, NY, USA, IHI '12, 819–824, https://doi.org/10.1145/2110363.2110464

Wang Y, Gan W, Yang J et al (2019) Dynamic curriculum learning for imbalanced data classification. In: 2019 IEEE/CVF international conference on computer vision (ICCV), 5016–5025, https://doi.org/10.1109/ICCV.2019.00512

Wang X, Pham H, Michel P et al (2020) Optimizing data usage via differentiable rewards. In: III HD, Singh A (eds) Proceedings of the 37th international conference on machine learning, proceedings of machine learning research, vol 119. PMLR, 9983–9995, https://proceedings.mlr.press/v119/wang20p.html

Wang X, Chen Y, Zhu W (2021) A survey on curriculum learning. IEEE Trans Pattern Anal Mach Intell. https://doi.org/10.1109/TPAMI.2021.3069908

Ware M, Frank E, Holmes G et al (2001) Interactive machine learning: letting users build classifiers. Int J Hum Comput Stud 55(3):281–292. https://doi.org/10.1006/ijhc.2001.0499

Weimer M (2010) Machine teaching: a machine learning approach to technology enhanced learning. PhD thesis, Darmstadt University of Technology, http://tuprints.ulb.tu-darmstadt.de/2109/

Weinshall D, Cohen G, Amir D (2018) Curriculum learning by transfer learning: Theory and experiments with deep networks. In: Proceedings of the 35th annual international conference on machine learning, 5235–5243, http://proceedings.mlr.press/v80/weinshall18a.html

Weiss K, Khoshgoftaar TM, Wang D (2016) A survey of transfer learning. J Big data 3(1):1–40. https://doi.org/10.1186/s40537-016-0043-6

Weitekamp D, Harpstead E, Koedinger KR (2020) An interaction design for machine teaching to develop AI tutors. In: Proceedings of the 2020 CHI conference on human factors in computing systems, 1–11, https://doi.org/10.1145/3313831.3376226

Wing JM (2021) Trustworthy AI. Commun ACM 64(10):64–71. https://doi.org/10.1145/3448248

Wong JS (2018) Design and fiction: imagining civic AI. Interactions 25(6):42–45. https://doi.org/10.1145/3274568

Xu W (2019) Toward human-centered AI: a perspective from human–computer interaction. Interactions 26(4):42–46. https://doi.org/10.1145/3328485

Xu B, Zhang L, Mao Z et al (2020) Curriculum learning for natural language understanding. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 6095–6104, https://doi.org/10.18653/v1/2020.acl-main.542, https://www.aclweb.org/anthology/2020.acl-main.542

Yang Q, Suh J, Chen NC et al (2018) Grounding interactive machine learning tool design in how non-experts actually build models. In: Proceedings of the 2018 designing interactive systems conference. Association for Computing Machinery, New York, NY, USA, DIS '18, 573–584, https://doi.org/10.1145/3196709.3196729

Yang SJ, Ogata H, Matsui T et al (2021) Human-centered artificial intelligence in education: seeing the invisible through the visible. Comput Educ 2(100):008. https://doi.org/10.1016/j.caeai.2021.100008

Zbyszynski M, Tanaka A, Visi F (2020) Interactive machine learning: strategies for live performance using electromyography. In: Silva H (ed) Open source biomedical engineering. Springer, Berlin

Zeiler MD, Fergus R (2014) Visualizing and understanding convolutional networks. In: Fleet D, Pajdla T, Schiele B et al (eds) European conference on computer vision. Springer, Cham, pp 818–833. https://doi.org/10.1007/978-3-319-10590-1_53

Zeiler MD, Taylor GW, Fergus R (2011) Adaptive deconvolutional networks for mid and high level feature learning. In: 2011 International conference on computer vision, 2018–2025, https://doi.org/10.1109/ICCV.2011.6126474, https://ieeexplore.ieee.org/document/6126474

Zhang X, Kumar G, Khayrallah H et al (2018) An empirical exploration of curriculum learning for neural machine translation. arXiv e-prints arxiv:1811.00739 [cs.CL]

Zhang D, Han J, Guo G et al (2019a) Learning object detectors with semi-annotated weak labels. IEEE Trans Circuits Syst Video Technol 29(12):3622–3635. https://doi.org/10.1109/TCSVT.2018.2884173

Zhang X, Shapiro P, Kumar G et al (2019b) Curriculum learning for domain adaptation in neural machine translation. In: Proceedings of the 2019 conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Association for Computational Linguistics, Minneapolis, Minnesota, 1903–1915, https://doi.org/10.18653/v1/N19-1189, https://www.aclweb.org/anthology/N19-1189

Zhao Y, Prosperi M, Lyu T et al (2020) Integrating crowdsourcing and active learning for classification of work-life events from tweets. In: Fujita H, Fournier-Viger P, Ali M et al (eds) Trends in artificial intelligence theory and applications. Artificial intelligence practices. Springer, Cham, pp 333–344. https://doi.org/10.1007/978-3-030-55789-8_30

Zhou B, Khosla A, Lapedriza A et al (2016) Learning deep features for discriminative localization. In: 2016 IEEE conference on computer vision and pattern recognition (CVPR), 2921–2929, https://doi.org/10.1109/CVPR.2016.319, https://ieeexplore.ieee.org/document/7780688

Zhou Y, Yang B, Wong DF et al (2020) Uncertainty-aware curriculum learning for neural machine translation. In: Proceedings of the 58th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Online, 6934–6944, https://doi.org/10.18653/v1/2020.acl-main.620, https://www.aclweb.org/anthology/2020.acl-main.620

Zhu X (2015) Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In: Proceedings of the twenty-ninth AAAI conference on artificial intelligence. AAAI Press, AAAI'15, 4083–4087, https://ojs.aaai.org/index.php/AAAI/article/view/9761

Zhu X, Singla A, Zilles S et al (2018) An overview of machine teaching. arXiv e-prints arxiv:1801.05927

Zhuang F, Qi Z, Duan K et al (2021) A comprehensive survey on transfer learning. Proc IEEE 109(1):43–76. https://doi.org/10.1109/JPROC.2020.3004555