Survey Article Open Access

Thomas Cederborg*

Artificial Learners Adopting Normative Conventions from Human Teachers

Abstract: This survey provides an overview of implemented systems, theoretical work, as well as studies of biological systems relevant to the design of artificial learners trying to figure out what a human teacher would like them to do. Implementations of artificial learners are covered, with a focus on experiments trying to find better interpretations of human behavior, as well as algorithms that autonomously improve a model of the teacher. A distinction is made between learners trying to interpret teacher behavior in order to learn what the teacher would like the learner to do on the one hand, and learners whose explicit or implicit goal is to get something from the teacher on the other hand (for example rewards, or knowledge about how the world works). The survey covers the former type of systems. Human teachers are covered, focusing on studies that say something concrete about how one should interpret the behavior of a human teacher that is interacting with an artificial learner. Certain types of biological learners are interesting as inspiration for the types of artificial systems we are concerned with. The survey focus on studies of biological learners adopting normative conventions, as well as joint intentionality team efforts.

Keywords: Interactive Machine Learning, Interpreting Human Teachers, Social Learning, Autonomous Re-Interpretation of Human Teachers

***Corresponding Author: Thomas Cederborg:** Georgia Tech, E-mail: thomascederborgsemail@gmail.com

1 Introduction

In some cases, it is not possible at design time to specify what a learner should accomplish. This would for example be the case when a system is shipped to many different humans, each with a unique set of preferences, or when designers can not foresee all situations that the learner will encounter, or all tasks it will be expected to perform. It can also be difficult to specify even a known task, in a known environment, in terms of a reward function over inputs and outputs. Learners that are to operate without a pre specified goal will need some alternative way of figuring out what should be done. The present survey is concerned with the design of artificial learners that deal with this issue by figuring out what a human teacher would like the learner to do. Recent research from several separate disciplines are covered, with a focus on explaining how they relate to the general project of designing an artificial learner able to figure out what a human teacher would like the learner to do.

This survey covers several different types of implemented systems, but also a range of other types of work. Work that provides formalisms for implemented systems, define the problem, and in general creates stronger theoretical foundations are covered. Theoretical work is also an open research frontier, and possible avenues for progress are discussed. Studies of human teachers are also covered. Since the learner figures out what to do by interpreting the behavior of a human teacher, these studies are essential. Studies of biological learners, human or otherwise, are covered because a strategy employed by any type of learner can be used as inspiration for algorithms.

An important concept throughout the survey is viewing algorithms as built on top of interpretations of a human teacher. The focus of the survey is indeed on research that in some way contributes to better such interpretations. Implemented systems are viewed as built on top of interpretations. Studies of human teachers are viewed as providing information about the thing being interpreted. The interpretations of human teachers employed by certain types of biological learners can be useful as inspiration. Finally, formalisms and other theoretical results are needed to define more precisely what is meant by “a correct interpretation” (which is not always straightforward, especially when dealing with various types of flawed/limited/misinformed teachers). Because of the centrality of the concept throughout the paper, a proper description of how learning algorithms can be seen as interpretations is provided in the next section.

2 Learning algorithms as interpretations

A given algorithm that obtains a policy by interacting with a teacher can be viewed as built on top of an implicit or explicit interpretation of teacher behavior. An algorithm that imitates the actions of a teacher corresponds to some set of assumptions regarding how teacher actions relates to what the teacher would like the learner to do. In a similar way, a learner that is maximizing the number of times a teacher presses a positive evaluation button, minus the number of times the teacher presses a negative evaluation button, is also making a set of assumptions regarding how to interpret what the teacher means by pressing these buttons. It is often possible to specify several different interpretations for the same behavior, and it is possible to compare the accuracy of such interpretations. This comparison becomes well defined due to the fact that we know the purpose of the system; to do what the human teacher would like the learner to do.

The question of the accuracy of teacher models used by systems trying to achieve some other goal (for example competing with, or extracting something from, a human) fall outside the scope of this survey. For systems with the specific goal of doing what a human teacher would like it to do, an interpretation of the behavior of that teacher can be bad/unsuitable, in the sense of being bad/unsuitable for the purpose of achieving that specific goal. In the context of the types of systems covered by this survey, a given interpretation, underlying some algorithm, can therefore be classified as bad/unsuitable, according to a well defined criteria. The present survey will focus on recent research relevant for improving these underlying foundational interpretations that learning algorithms are built on top of.

While a perfect interpretation is not a realistic goal (no fully accurate interpretation will be representable in any implementable parameter space), this does not mean that all incorrect interpretations are equal. All interpretations are wrong, but some interpretations are more wrong than others. This is just one more instance where an accurate model of the world is impossible, but where some world models are better than others, even though we accept the fact that all implementable models of the real world are wrong. If we put a robot in any unstructured environment, for example a forrest or a city, a fully accurate model of the world will not be representable in any implementable parameter space, but some models are still better than others. The same principle holds for interpretations of a human teacher.

2.1 Finding good foundational interpretations is difficult and important

Before going into details about formalisms, theory, and delineating the boundaries of the field more precisely in the next section, this section will explain why focused investigation into foundational interpretations is needed. It is easy to see that some types of information sources will require some effort when it comes to building interpretations, for example learning from EEG readings, simply because in their raw form they make no sense to a human. Even though humans instinctively comprehend facial expressions, it is probably clear to researchers in any AI related field that some sophisticated processing will be required here. For some types of behavior however, it is easy to come up with short descriptions of interpretations of common human behaviors, for example: “imitate the teacher‘s actions”, “reproduce the effects of the teacher‘s actions”, “maximise the positive minus the negative feedback received”. It is however important to find out if these are actually good interpretations, or if they are just reasonably sounding simple sentences.

Imagine for example a teacher eating candy in front of the tv. This teacher might not want the learner to do the same, and might not like the damaged teeth that is one well understood and predictable consequence of the behavior. This is not simply a matter of the teacher being inefficient or “failing to achieve a goal”. Modelling the teacher as pursuing a clearly defined goal, and employing an optimal policy perturbed by Gaussian noise would probably not end well. A learner finding a clever new way of more effectively damaging teeth has misunderstood something foundational. This difficulty in interpreting a teacher action must be dealt with separately, in addition to other common machine learning problems, such as finding good state and action representations (there are interpretation issues that can’t be solved by trying to find out in exactly which situation the learner should watch tv, or by trying to find out what specific type of tooth damage is desired). Teacher actions are of course a valuable source of information, the point made here is simply that finding interpretations of teacher actions that are actually reasonably accurate (as opposed to just sounding reasonable) is a non trivial challenge. Evaluative feedback is also tricky when examined closer, and will serve as an example of how important it is to be careful, even when interpreting explicit teacher approval/disapproval.

2.2 Maximising human generated rewards

If we build a learner that is to do what a human teacher wants it to do, and that learner is maximising the number of times the teacher pushes a positive evaluation button minus the times the teacher pushes a negative evaluation button, then we are making assumptions regarding the interpretation of the human behavior. If we instead build a learner whose purpose is simply to get good evaluations, then we are not necessarily making the same set of assumptions, because any unexpected way of getting good evaluations or avoiding bad evaluations would simply be a clever and appreciated trick.

Consider the example of a human teacher that during learning will give more positive than negative evaluations, and that will stop giving evaluations when the learner "is done learning". Consider now the strategy of deliberately slowing down learning, resulting in a larger sum of rewards. The response that a designer has towards this clever way of maximising rewards is useful for figuring out what type of system is being designed. If the strategy is seen as a success, then the designer could be trying to get the learner to maximise rewards. If the strategy is seen as a failure, then the designer is probably not simply trying to build a learner that receives good evaluations, and it is likely that the designer has made some incorrect assumptions. Now, let's consider another teacher, this time one that gives more negative than positive rewards on average. A learner could respond with some creative strategy that makes the teacher want to avoid the learner and stop the interaction. If this "making the interaction itself undesirable" strategy would also be counted as a success, then the goal of the designer might really be to maximise rewards. Similarly, we might consider the case where concealing a mistake will avoid negative evaluations, or where secretly creating a mess provides the opportunity for positive evaluations when cleaning it up.

If the designer is really just trying to maximise positive evaluations, and approves of all these strategies to do so, then there might not be any incorrect interpretations involved, explicit or implicit. The present survey however deals with systems whose purpose is to do what a human teacher would want them to do. For such a learner, maximising positive minus negative evaluations corresponds to an interpretation or a set of assumptions whose accuracy can be questioned, as has been done in [1–4].

2.3 Alternative interpretation: evaluations refer to action choice

If we instead make the interpretation that teacher evaluations refer to the action choice of a learner, then this alternate interpretation has consequences for algorithm design. The Policy Shaping algorithm [5, 6] is built on this action evaluation interpretation, and has been shown to work well with human teachers. It is based on increasing/decreasing the probability of taking some action in some state, as a response to teacher approval/disapproval of taking that action in that state (directly shaping a policy as a response to evaluations). For this interpretation there is for example no reason to find clever ways to avoid interactions with a teacher that is on average negative. A very negative teacher is just providing ample information of a particular kind, specifically the teacher is giving a lot of information regarding actions that should not be taken.

It is possible to construct domains and situations, specifically designed to experimentally check how human teachers use evaluative buttons. Experiments have shown that human teachers do not seem to assume reward maximising learners [4], and for example routinely give more rewards for solving a problem than the punishment given for creating the problem. In the case explored in [4], the agent moved to a bad location, and then took a series of good actions that resulted in going back to the starting point. Human teachers tended to give more positive than negative feedback during this cycle. In light of these experimental findings, there is little reason to assume that human feedback will be potential based, even if they only care about the end state. The experiments further showed that neither model tested seemed to fully encapsulate all the aspects of how the teachers used evaluative buttons (which is hardly surprising given the complexity of the thing being modelled). These experiments illustrated that finding very accurate models is a challenging problem, but at the same time showed that carefully designed experiments can improve our understanding of how to interpret humans. The question of interpretation can also be investigated by looking at human learners, as in [7], which argues that human learners do not interpret social feedback from a human teacher as something that should be maximized.

All algorithms with the goal of the present survey rely on interpretations, either explicitly or implicitly. Even with behaviors such as explicit approval or disapproval, using dedicated evaluative buttons, we need careful consideration and human subject experiments to find reasonably accurate interpretations. It thus seems likely that a learner interpreting more complex behaviors, such as facial expressions, will also need to deal with similar foundational interpretation problems, in addition to all the other technical challenges involved. When interpreting the basic meanings of things such as for example smiles, nods, concerned surprise, and frowns, it is possible to draw inspiration from human button pushing behaviors. Based on the realisation that maximising positive minus negative feedback buttons constitutes a serious misinterpretation of how human teachers use feedback buttons, we might for example question the implicit interpretation assumptions that an algorithm that is maximizing smiles minus frowns would be built on top of. There is however no reason to be certain that things such as smiles, nods, concerned surprise, frowns, fear, disgust, etc will correspond well to how humans use positive and negative feedback buttons, and separate investigations would need to be carried out for these types of social signals. These foundational interpretation issues could of course be solved concurrently with the problem of recognising affect or other social cues (for example building a social signal recognition system focusing on signals that are easier to learn from because we have a better understanding of what they mean, or because they are more reliable). The survey will cover affect recognition/Social Signal Processing in a later section, here the point is simply to highlight the the similarity of the foundational interpretation aspects one faces when dealing with information sources where there is no signal processing involved, such as a human teacher pushing buttons.

While it's not possible to find perfect interpretations, it is possible to formulate, and experimentally validate, better interpretations. This matters to designers of systems because new foundational interpretations imply new types of algorithms.

3 Staking out the boundaries of a field

The scope of the survey covers the full spectrum of information sources, from those whose meaning might at first glance seem obvious to humans (such as evaluative buttons) on one end, and information sources that are obviously in need of carefully designed interpretations (such as EEG signals) on the other end of the spectrum. The different parts of the spectrum present different types of challenges. Some information sources might be prone to yield "obvious" but incorrect interpretations, for example when dealing with an information source humans are good at dealing with (but maybe less good at introspecting on). For information sources where the detection and representation issues are more complex, the foundational interpretation questions might get lost amongst all the other technical challenges.

Thought experiments and theoretical arguments can illuminate hidden assumptions and make some interpretations look more likely than others. New interpretations can be used to construct new algorithms, and it is possible to directly compare interpretations with carefully designed human experiments. Direct investigations of human teachers can inform us about appropriate interpretations, and certain types of strategies used by biological systems can be used as inspiration for artificial systems.

Absent an obviously correct way of mathematically formalising the problem, it is difficult to define strict boundaries of the field, and theoretical work becomes an important open research frontier. We will start with various types of formalisms, and then move on to other types of theoretical work.

For imitation learning/learning from demonstrations there exists multiple formalisms, and they have generally taken two main forms: classification of tasks and mathematical formalisms with success criteria.

In the classical work of [8], an algebraic framework is specified and a success criterion is defined for imitation learning. The set of imitator/learner and environment states is referred to as X, consisting of the states at each instance of a time series (where each time instance contain for example: states internal to the imitator/learner, the fact that the learner is holding an apple, states in the environment, etc). The set of demonstrator and environment states is referred to as Y. Both X and Y are contained in the state set Z. Finally, success is defined as minimizing a distance metric $d : Z \times Z \to \mathfrak{R}$ (where 0 is optimal imitation).

There are three issues with applying this formalism to the problem of learners trying to figure out what a human wants them to do; (i) it is not clear how such a distance metric should be obtained, (ii) the formalism in [8] does not include the possibility of incorporating other types of information sources besides demonstrations, and (iii) the formalism in [8] cannot formalize the situation of non-optimal demonstrations. Even with a correct framing and a perfect distance metric between imitator behavior and a demonstration, the situation where a demonstrator simply failed at achieving the task perfectly cannot be handled properly. Imagine, for example, a demonstrator trying to shoot a basketball at a hoop and failing most of the time. This is a situation an imitator might be able to infer a goal from, especially given complementary information sources, such as facial expressions. Even if the imitator has identical embodiment, and is in an identical situation as what the demonstrator was in during a demonstration, the learner should not miss on purpose if it knows what the goal was and is able to achieve it, even if the demonstrator did miss in the same situation. But according to the formalism in [8], missing the shot in this situation is always an optimal action. Experiments where an imitator outperforms the demonstrator, for example in [9], or in [10], where an imitator flies a helicopter better than a human, should ideally fit in a formalism.

The summary provided in [11] also offers a formalism for learning from demonstration. The demonstrations are seen as generated from a function that maps inputs to outputs, and the goal of the learner is defined as approximating that function. This leads us back to the question of how to do better than the demonstrator. The question is discussed in [11], and one solution offered is to either filter out bad demonstrations or smooth them over with regression techniques. This does sound like it can achieve the goal of having an imitator outperform a demonstrator, but not according to the definition of success described in [11], where the learner is to approximate the function that generates the demonstrations. It also does not deal with the case where the teacher is never able to achieve optimal performance (for example attempting to teach a robot how to throw a ball as far as possible, where the robot could in principle throw the ball much further than the teacher). The other approach proposed in [11] is to seek feedback. This strategy seems like a good idea (the goal of the present survey is in fact to cover this exact type of learner), but this approach also falls outside of the formalism as presented in [11].

In [12] an attempt to categorize imitation learning is made, with a focus on classifying various imitation learning tasks in terms of what type of goal the demonstrator would like the imitator to perform, for example replicating the exact movement, or replicating the end state of a manipulated object. This is a very useful description of the landscape of different types of tasks, but it does not address the issue of trying to specify what constitutes success for a learner trying to figure out what a teacher would like it to do.

In [13], different types of social learning are classified based on the degree to which a learner is trying to do what a teacher wants it to do. A distinction is also made between higher level teacher goals and low level teacher actions, forming a triangle. Learners that do not care at all about what the teacher wants the learner to do is in one corner of the triangle (for example learning how an object functions by observing a teacher, but without any consideration of that teacher's goals). Learners that care only about high level teacher goals are in another corner, and learners caring about only low level actions are in the third corner. Learners caring about multiple things are somewhere in the middle. This is very useful for describing different types of learner motivations, but as with [12], it does not address the issue of trying to specify what constitutes success for a learner trying to figure out what a teacher would like it to do. In the language of the framework presented in [13], the present survey deals with learners that are somewhere on the side of the triangle opposite the "don't care about the teacher corner". Where the learner is on this side of the triangle depends on what the teacher wants it to do. If the teacher is for example showing the learner how to dance (or perform some other task where only the actions matter, and the end state is always the same) it would be at one end of this side, while if the teacher wants to achieve some specific world state by any means (and there are no constraints regarding the types of actions that are acceptable), then the learner would be at the opposite end of this side of the triangle.

In [14], an agent based view of imitation learning is presented and five central questions are put forward. The imitator must decide who, when, what, and how to imitate, and someone must address the question of what constitutes success (for example formalized by an experimenter or the creator of an artificial imitator/learner). If the learner has access to several different types of interaction, the question of when to imitate can be cast as a question of which type of interaction is most efficient for figuring out what the teacher wants the learner to do (an alternative to requesting a demonstration could for example be to perform an action and then try to interpret the teachers response to that action). In [14], the question of what the imitator is trying to achieve is left open (and depends on what type of animal the imitator is, or what the designer of an artificial imitator is trying to achieve). If the imitator has a specific goal whose progress it can measure and that is unrelated to the teacher/demonstrator, then the "who to imitate" and "when to imitate" questions can be answered with respect to that goal. Consider a robot that is trying to maximise the amount of states it can reach in a measurable outcome space, or an animal that is trying to maximise food (for example given by a human as rewards for imitation). Who to imitate is then a question of who gives the most informative demonstrations relative to current skill levels (producing reproducible actions with outcomes it could not previously achieve) or a question of who gives it the most food. When to imitate depends on how effective this is for learning relative to other learning strategies or a question of whether imitation is the most effective way of getting food. The "how" and "what" questions also becomes well formalized (for example imitating the parts of the demonstration that results in food), meaning that the simultaneous exploration of all four questions is a well formed problem. In [15], all four questions are simultaneously addressed by an artificial imitator that is trying to expand the amount of outcomes it can reach through strategic and active learning. The learner in [15] leverages the fact that it has a well defined goal, with a success criteria that is both separate from the teacher, and defined so that the level of success is directly observable (expanding the set of achievable outcomes). It is less clear how to deal with all four questions simultaneously when the goal comes from the teacher. We can have some intuition that a teacher eating candy in front of tv should probably not be imitated (at least not at that specific time), and that the learner should probably find some other way of gaining information (for example relying on teacher evaluations of learner actions). But it is not easy to formalise the case when the goal comes entirely from the teacher, and is not directly visible to the learner.

In [16], a formalism is presented for tele-operated robots learning from demonstration, but it is not clear how to extend it to be suitable for the general case where a learner interacts with a possibly flawed teacher, and learning from different information sources in addition to demonstrations (eye gaze, an evaluation button, facial expressions, speech comments, EEG readings, etc).

In [17] a general formalism, seeking to cover any interaction mode and information source, is introduced for artificial learners autonomously improving their interpretation of a human teacher. The information sources described in this survey should all be covered by the framework in [17]. The information spaces of [17] are similar to [16], and parts of the formalism presented in [17] can be seen as building on ideas from [16].

In the formalism of [17], there is a shift in how interpretation is viewed, going from the situation with a static interpretation of a teacher's behavior, to a situation with a parameterized hypothesis space, that is updated based on observations. This can be seen as similar to the shift from a focus on solving MDPs to a focus on POMDPs. One can draw a parallel with moving from planning in a static world model, to re-estimating world dynamics based on observations (a static and known world model is replaced by an estimate that can be updated. In the same way, a static and known interpretation of a teacher is replaced by an estimate that can be updated). This transition was important to the field, and has for example been described in [18] and [19], where a family of formalisms are proposed for dealing with unknown world dynamics. While the formalisms described in [18] and [19] all concern different ways of describing how to model worlds with partially known dynamics, the formalism in [17] instead describes how to model a very specific situation involving a human teacher (focusing on what should be done, not how to do it). Moving from a static to an autonomously estimated interpretation of a human teacher's behaviors can be seen as analogous to moving from static to autonomously estimated world dynamics. Let's look at an intermediate step where the reward is a hidden state, modifying observations in a way that is not completely known. An extra hidden state H_r (hidden reward) would need to be added, and the reward signal removed from the observable space. The state H_r is now dependent on states and actions, and it modifies observable states. The learner has a prior over how states impact H_r, and a prior over how H_r impacts observable states. The goal of the learner is now to maximise H_r. Even though the learner might never be able to observe H_r perfectly, it can still make probabilistic updates regarding how H_r is modified (both by other hidden states, and by observable states), and how H_r modifies observable states. It is also possible to take information gathering actions that are specifically chosen because, when those actions are taken, different hypotheses regarding how H_r interacts with other states implies different predictions for observable states.

It has also been proposed that learning context-dependent skills could be achieved through Inverse Reinforcement Learning (IRL) [10, 20–22]. Demonstrations are assumed to be optimising a reward function, and the first step in IRL is to find this function. One advantage of this approach is that it allows potentially better generalization by letting the robot self-explore alternative strategies to achieve the goal that may be more efficient or more robust than those used by the demonstrator (as for example in [10]).

A learner can also be formalised as a team member, where a human and an artificial system jointly tries to achieve a goal. Teams of humans exists in abundance, and one can draw inspiration from these, including teams where team members have different roles (in our teacher-learner team, only the human have access to the team goal, resulting in very distinct participant roles). Humans have a long and diverse history of team collaborations in situations where team members have minds of very different kinds, and where each team member has a very distinct role. These can be useful as inspiration, and offer a wide range of team types, for example hunting dogs and hunters, horses and hunters/riders/cavalry, sheep dogs and shepherds, falcons and falconers, etc. Many challenges are of course unique to the the teacher-learner team, where a designed mind is to achieve the goals of a messy, evolved biological system. Nevertheless, there exists a very rich and diverse set of teams that function well, despite team members having very different types of minds, that one can be inspired by.

In [23], it is suggested that artificial systems should be conceptualised as part of a genuine collaborative activity, as opposed to setups where the robot is conceptualised as a tool that a human is using. Conceptualising the robot as an ally that shares intentionality with a human, instead of as a fancy hammer, makes it natural to start thinking about how the learner might solve some of the subtasks that socially collaborative humans routinely deal with. Especially relevant to the types of learners covered in the present survey would be: i) autonomously learning how the social signals of some particular human should be interpreted, ii) evaluating the competence of a collaborator (how likely is it that demonstrated actions are good actions?) iii) learning how to get more useful information from collaborators (postponing an action until the teacher is looking, displaying confusion with a facial expression or confirming understanding by nodding, asking good questions, etc). The idea of seeing the interaction between robot and human as a one directional transfer has also been challenged in [3] as well as in [24], where experiments show how a robot can use eye gaze to modify the behavior of human teachers in a way that helps the interaction.

In [25], a collaborative setup is explored, similar to the team member setup suggested in [23, 24]. A teacher/architect has a structure in mind, and tries to communicate this with a learner/builder, that builds the structure. One team member is able to act on the world, and the other has access to the joint goal. A central feature of this setup, besides the collaborative aspect, is that there is no goal conflict. The team goal is to accomplish the goal of the teacher/architect, and the difficulties come from trying to coordinate this project.

A formalism for a situation that is very similar to the team setup from [23–25] is presented in [26], where Cooperation is incorporated into the Inverse Reinforcement Learning framework (resulting in Cooperative Inverse Reinforcement Learning: CIRL). A robot is maximising, but not adopting, the reward function of a human. The human and the robot is collaboratively trying to maximise the reward function of the human. The reward function is not visible to the robot, and a collaborative efforts is aimed at improving the robots estimate of the reward function. There are many problems to deal with when it comes to this type of cooperation, especially in the case with a real, non idealised, human teacher in an unstructured environment. But the collaborative setting removes the added difficulties of competing interests, and conceptualising the interaction as a team effort opens many new avenues of approaching the problem.

There are several theoretical challenges one faces when trying to infer a utility function from the actions of a human teacher. One issue is that for any set of actions of an agent, there is an uncountable infinity of utility functions for which this set of agent actions is optimal, including some that are clearly bad descriptions of what a human teacher values (all functions that assign the same value to all actions in all states is for example a fit for any observed set of actions). Another issue when it comes to human teachers is the many different ways in which humans are flawed or limited. A human might miss a basketball shot, despite not wanting to miss. If inferring a utility function from behavior, it's good to take into account the fact that humans do not always act optimally. For a given set of actions, it is possible that there are no good descriptions of what the human is trying to do amongst the set of functions for which these actions are optimal. Because the human is not acting optimally. In a similar sense, it can be useful for an artificial learner to take into account the fact that humans choose actions given a certain set of cognitive limitations, and that humans do not have access to a well defined function that describe everything that they care about.

There is a question of how to theoretically define optimality in the case of teachers that are limited in various ways. For certain types of agents, the best way of achieving optimality according to one function is to maximise some other function [27, 28]. In [27, 28], the example of evolutionary systems is given, where the concept of an optimal reward function is defined as a function that results in agent actions that maximises a fitness function (conditioned on some specific set of environments and some specific type of limited agent). The evolutionary search for a good reward function relative to a given fitness function is distinct from an agent searching for a good value function relative to a given reward function. These considerations further complicates the question of how an artificial learner is supposed to figure out what function it is supposed to optimise based on observing a set of actions.

There is also a certain sense in which actions can be optimal when conditioned on the cognitive limitations of a human. See [29, 30] for a discussion of optimality given cognitive limitations in the context of modelling human minds, with a focus on limits to computational power.

Humans always make decisions in a context of limitations. A subset of these limitation consist of a very diverse set of cognitive limitations. Writing down passwords on post-its next to the computer might be optimal in a certain sense, since the option to just remember all passwords might not be available. Skipping a test as a result of not having studied might be optimal. The option of taking the test and giving the correct answers might not be available. Buying the month card at the gym, instead of the year card, due to uncertainty about motivation could be a good idea if "going to the gym regularly for a year" is not an action that can be reliably taken. Watching the first episode of a tv series might be a bad idea as it might be difficult to stop. The option to watch just one episode might not be available. Not driving at all might be a good idea if drunk, because the option to drive well is currently unavailable. Betting money on a chess game against a superior chess computer could be a bad idea because the option to take the bet and beat the computer is not available.

There is a diverse set of limitations that human actions are chosen with respect to. A human interpreting the actions of another human instinctively understand them within this context, and for example see that a human could write passwords on post-its, even if there is no intent of making this information available to others. It is obvious that the actions might have been chosen within the context of a cognitive reality where "just remember all passwords" is simply not an option. This is however not automatically obvious to an artificial learner interpreting human actions, and unless some care is taken to actually account for this, there might be misunderstandings (things that are obvious to humans are always at risk of being overlooked in implementations). In other words: a set of actions plus a set of limitations can lead an artificial learner to infer a different set of preferences than what would be inferred from just the set of actions.

A human can make choices that are simply bad, including making choices that unwisely ignore cognitive limitations. In other situations, such as those described in [29, 30], some actions can be optimal in the specific sense of being optimal with respect to a given set of cognitive limitations. Determining if a given teacher action is a mistake, or if it's good within the context of some set of limitations, seems like a non trivial problem, especially before the learner has figured out what the teacher is trying to do. In principle, both things could be true at the same time; a teacher that is unable to remember passwords might still make a mistake when writing down all passwords on post-its (even though the action of simply remembering all passwords is not available, there might be some other option that is available and that is better than the post-its). It is in general not straightforward to define what success means for a learner when teachers are flawed, limited, mistaken, misinformed, and sometimes simply fail. One can approach these issues with the aim of defining success in new and more complex situations. One can also approach them from the point of view of experimental setup design, deliberately avoiding ambiguities that results in theoretical questions that one is not able to answer yet (for example avoiding setups that lead to demonstrations that are only good in the context of certain limitations). See [17] for an attempt to address some of these issues.

In principle, it would also be possible for a learner to proactively avoid ambiguities that it is not able to deal with. Consider a teacher that inspects an apartment and gives an evaluation of a cleaning robot based on how clean the apartment looks, and a cleaning robot that is trying to figure out if sweeping dust under the rug is acceptable. A positive evaluation from a teacher that has only observed the end result would be difficult to deal with. A simple coping mechanism would be to sweep the dust under the rug when the teacher happens to be watching. Even if it is not clear how to deal with all the tricky theoretical issues, one can still use clever coping strategies that would be hard to find if we simply defined away all the complexity (for example by defining the problem as trying to get positive evaluations, as opposed to trying to do what the teacher would like the learner to do).

As with modelling other aspects of a human teacher, it is unrealistic to expect a fully exhaustive and completely accurate model of the cognitive limitations of a given human. It is however possible to do better than the default case (of not taking such limitations into account at all), even without a perfect model.

An alternate way of describing the scope of the present survey is by using the terminology suggested in [31] for value learning. In the terminology of [31], the present survey deals with research relevant to a specific type of value learning, specifically the kind that involve interactions with one human teacher. Finding a more accurate interpretation of a teacher could be described in the terminology of [31] as: “finding a better mapping from world models to goal distributions, through interactions with a human teacher” (interacting with a human can lead to better interpretations, and a better interpretation is the same thing as a better mapping).

Due to the complexity of the thing modelled, all implementable interpretations will be wrong, and this results in some complications. While some simulated teachers, acting in a simulated world, could be perfectly modelled, a learner that is to interact with a real human in an unstructured environment will always act on incorrect interpretations (just as any artificial learner operating in the real world will always have to act based on an incorrect world model). The question is how to improve interpretations, or how to compare the relative accuracy of two competing interpretations. Any proof or other theoretical work needs to keep in mind that the set of representable interpretations will not contain any fully accurate interpretation. All possible hypotheses in the search space will be wrong, and one should keep this in mind when considering things such as the optimality of different types of probabilistic updates and related overfitting issues. Formalising a situation where the learner’s success estimates are guaranteed to be wrong is not straightforward, especially with regards to overfitting (solutions to overfitting that rely on the correct answer being representable in an implementable parameter space is not applicable here). A thought experiment adapted from the section in [31] on value learning is useful for illustrating a class of solutions related to overfitting. The idea is roughly as follows: “a human writes a text describing a set of values, and then seals this text within an envelope. The agent tries to maximise the values described by the text, but is unable to read the text, and can only make basic guesses regarding what values a human would be more likely to describe.” There is clearly a lot of uncertainty in this scenario, but not all possible texts or value systems have the same probability of being written on a piece of paper, and the agent might be able to act in a fairly reasonable way. It seems like something along these lines could, at least in principle, avoid the problems of overconfidence that can arise if learning from some highly specific, but not fully correct, interpretation of a teacher.

It has been shown that the inability of a learner’s model to fully describe a human teacher can complicate the problem of deciding when an artificial learner should follow orders [32]. The basic problem is that a human teacher might give bad orders that a learner should not follow, due to various types of teacher flaws/limitations. If the learner is adjusting the parameters of a model of teacher preferences based on observations, but has no idea how that model was constructed, it might be difficult for the learner to know how appropriate the model is. If the learner’s model of the teacher’s preferences is missing important features, then it has been shown that this can lead the learner to systematically err on the side of disregarding teacher orders too much [32]. Basically, if there exist some thing that is important to the human teacher, but that can not be represented in the learner’s model space, then there is no way for the learner to build a complete model of what the teacher wants, and the teacher might seem irrational to the learner. If the teacher is not consistently pursuing any goal representable in the learner’s model space of possible teacher goals, then one possibility is that the teacher is consistently pursuing a goal that is not representable in the learner’s model space.

Given the complexity of the thing being modelled, it seems unavoidable that the learner’s model of the teacher’s preferences will lack certain features. Since humans are in general not acting perfectly rationally, it seems unavoidable that a learner will encounter a human teacher that is imperfectly pursuing a goal that is not fully representable in the learner’s model space (it is indeed difficult to imagine a situation where everything that a human cares about is fully representable in a learner’s model space, just as it is difficult to imagine a situation where a human is acting completely optimally in pursuit of a well specified objective). Disentangling such a situation completely seems difficult, but one does not have to have a perfect solution in order to beat naive solutions (such as simply assuming a perfect model space, and/or a perfectly rational teacher). Given that this situation seems unavoidable if learners are to interact with real human teachers in unstructured environments, then it seems worthwhile to look for clever coping strategies, even if one can not find a perfect solution to the problem. Situations where a learner can detect model misspecification is described in [32], and the coping mechanism of obeying the teacher’s orders more when detecting model misspecification is suggested (if the learner estimates that the teacher probably cares about things that the learner is unable to see/model, then it seems reasonable that the learner should put less confidence in the apparent irrationality of the teacher).

Research investigating how to detect preference model misspecification, and efforts to implement coping mechanisms to mitigate the associated problems, is still in its early stages, and there are many interesting avenues for future research. One open avenue of future research is related to the fact that the inability of a learner's model to fully capture a teacher's preferences could become entangled with the issue of interpretation. As discussed throughout the survey, a complete and fully accurate interpretation of a human teacher in an unstructured environment will not be representable in any implementable parameter space, meaning that the orders that a learner will encounter when interacting with a human teacher in an unstructured environment will be imperfectly understood. The learner then has to deal concurrently with an imperfectly understood order, which is part of an imperfect strategy, in pursuit of a goal that the learner is incapable of perfectly representing. If the learner notices that there is a problem with its interpretation of the teacher's order, then it might be a good idea to reduce the degree to witch the learner follows strange orders. Therefore, an incorrectly specified preference model and an incorrectly specified interpretation model might point in opposite directions. In other words; the general realisation that there is some sort of misunderstanding might not always be easy to deal with, because blindly following orders might not be such a good idea if the orders could be misinterpreted, and disregarding orders because the teacher seems irrational might not be such a good idea if the learner is not capable of understanding what the teacher is trying to achieve. While this problem looks messy, simply defining away the complexity involved would not solve the problem, it would just make it more difficult to predict/diagnose failure, as well as interfere with efforts to find practically feasible coping mechanisms (for example based on working with the human teacher, jointly attempting to identify the specific source of a given misunderstanding, perhaps leveraging methods used by humans to sort out misunderstandings in human-human interactions).

An issue related to the problems of flawed interpretations and flawed models of a teacher is that the difference between teacher and learner concepts might increase over time [33]. This is a special case of the problems arising when generalising far from observed data. As with any generalisation problem, the issue becomes more serious when the agent operates far from where it has data, which could be due to exploring new environments, but in this case could also be due to learning new things about the world (thus possibly causing concepts to change, in turn causing a mismatch between teacher and learner). Consider a robot that is to operate autonomously in an urban environment, for example crossing the street while running errands. Consider the ideal of "being civilised and polite while moving around in traffic, including not getting in the way of cars", as a rule that is supposed to be important for its own sake (not simply for the sake of avoiding getting hit by cars, or avoiding sanction from people). If the learner has a concept of "car" that includes an internal combustion engine as a defining feature (for example due to the fact that this is what allows every car seen during learning to function), this rule might work well while operating in certain environments. There are many ways in which a well behaved learner could suddenly stop behaving well in traffic; it could learn that some of the cars already in its environment are actually electric, or it could move to a new environment with electric cars, or electric cars could eventually be introduced to the current environment (in other words: epistemological, physical, or temporal distance between data and environment could lead to generalisation failure). A human that never considered the possibility of an electric car could easily adapt (for example by changing the concept of "car" or, if very stubborn, by changing the rule to "don't get in the way of cars, or those new electric things"). A learner operating in an unstructured environment would need some form of coping mechanism (for example by modifying concepts or rules the way a human would, by avoiding new environments or knowledge that might cause these types of ontological issues, by being cautious and checking back with a human teacher when things change, etc). In [33], the author proposes to build an AI in such a way that verification of concept match with a human becomes easier.

3.1 Summary

Despite the large amount of relevant theoretical work that exists, creating solid theoretical foundations is still an open research frontier. Some work is directly trying to introduce notations and success criteria, such as: [8, 11, 16–20, 31]. Other work have contributed by describing/classifying different types of systems/experiments [12–14], for example depending on what type of task is being tackled, or what question the research is trying to address. Another type of contribution relates to proposing a way of viewing the learning context that focus on learning as a joint effort of teacher and learner, such as: [3, 23–26]. Yet another strand of contributions illustrates that there is still much work to do by addressing theoretically tricky issues, such as flawed/limited teachers [27–33]. For designers of systems, there is already a lot of useful theoretical work. There are also many open frontiers where incremental progress on theoretical problems seem possible.

4 Implementations and experiments

There are many different types of relevant interactions, and many different types of relevant information sources. Interaction types range from things such as feedback and demonstrations, to things such as written stories and speech comments. Within specific interaction types there can be a range of possible types of information sources, for example ranging from a specifically designated feedback button, pressed deliberately by a teacher as a response to learner actions, to an estimate of affect based on the facial expressions of a teacher, or EEG readings. The focus of this section will be on recent papers, or important implemented systems that in some way push the boundaries of the field.

Being able to interpret the types of social signals that a human teacher naturally display during interactions would be a step towards long term, natural, learning interactions in an unstructured environment, such as a home or an office. Learning from the rich set of social signals could also lead to a potential benefit related to foundational interpretation issues. It might be easier to find reasonably good foundational interpretations if one has access to the richer information that social signals offer. It might for example be easier to deal with a sympathetic facial expression meaning roughly “don’t worry, it’s ok that you failed, don’t be sad” than a positive evaluation button that is pushed in an attempt to convey the same message (the button input would look identical in this case and in the case when the button is used to say that an action was good, but it might be possible to separate different facial expressions). It is also possible to choose which social signals to learn from based on how good models of foundational interpretations are available.

Social signal processing [34–39] is an active field of research. To learn what a teacher would like a learner to do based on social signals, it is necessary to solve two distinct challenges, (i): the technical side of detecting affective states and other cues, and (ii): finding reasonably good interpretations of the foundational meaning on the other hand (for example asking what implicit assumptions we would be making if we maximise smiles minus frowns, and how to test the accuracy of those assumptions in the environments that the system will be operating in).

See [34, 35] for two surveys of Social Signal Processing, covering different methods for extracting information sources from human behavior. In line with the idea of viewing the learner as a team member mentioned above, in [35] both recognition and production of social signals is covered. [35] also advocate comprehensive solutions to the problem of multimodal interaction, as opposed to individual detached solutions to individual parts of the problem. See also [36] for a survey focused on sensor fusion in the domain of social signals.

Detecting affect in facial expressions is still a difficult technical problem, and in [38] we see that there are a number of challenges, ranging from uncertainty regarding suitable representations, to lighting variations, to the tendency for shifts in affect to be accompanied by shifts in head pose (possibly complicating recognition). The full spectrum of the technical, the human, and the environmental is already involved in detecting affect, even before we introduce the foundational interpretation question of; “what do these social signals mean from the point of view of a learner trying to use them to figure out what should be done?”, which would need to be addressed if they are to be used by a learner.

Besides having a large number of possible inputs to choose from, social signal processing also have a large set of possible lessons that could be learnt from affective states. In [40] for example, the robot Leo interacts with a human, and tries to estimate the affective response that the human has towards a mutually attended object. For estimated positive affect, Leo tends to explore and interact with the object, while avoiding it if there is a fear response, and rejecting it if there is a disgust response.

To understand every aspect of affect in terms of what a learner should infer about the task at hand when the teacher displays various affective states would be very difficult. A complete model of this kind would however not actually be needed for the purpose of building a good learning system; it is merely necessary to find some subset of affective states that can be reliably detected and that can be given reasonably accurate, and useful, interpretations. For this purpose, the high dimensionality of the space of possible inputs and meanings/lessons is an advantage.

We can not be certain, a priori, whether some specific social signal or spontaneously displayed affective state will have foundational meanings that are basically the same for all teachers, or if there will be significant variation. Given a system that is learning meanings/interpretations from data, it could be the case that for one teacher, filtering out head pose variations will lead to more reliable data because these variations interfere with affect recognition. For another teacher, head pose variations could be less problematic, and at the same time could provide information on affective state.

How much the foundational meaning of a given social signal wary between teachers would also need to be experimentally determined. Let's look closer at the example of smiles. One possibility is that the type of lesson a learner should infer from a smile will be teacher dependent (for example that the resulting state of an action was good, or that the action choice was good, or both). For some teachers it could be that smiles of the type: (i) polite, (ii) spontaneous, (iii) a specific "training the robot" smile, all have different meanings. For other teachers they could all mean the same thing. Another possibility is that it will be possible to build fully teacher independent interpretations that describe most teachers reasonably well (for example that polite smiles basically always mean "correct action but bad outcome", spontaneous smiles always mean "good action and good outcome", and that the meaning of deliberate "training the robot" smiles can be modified by instructions in predictable ways). One could make educated guesses based on thought experiments, insights from cognitive science, and analogy with other information sources (such as the meanings of teachers using evaluative buttons), but it seems like sooner or later one would need to turn to dedicated experiments examining foundational interpretation issues of different types of smiles in a deliberate and systematic way.

The amount of useful information one can get out of smiles, and how smiles should be interpreted, could be dependent on how good the classifier is at for example separating sympathetic "you failed but it's ok, don't be sad" smiles, from spontaneous enthusiastic smiles. The two issues of what to learn from a signal, and how to detect that signal, might therefore be difficult to disentangle.

For autonomous interpretation, building individual interpretations for each teacher, one can draw inspiration from existing algorithms, for example learning to interpret EEG signals [41] (the issue of autonomous interpretation is treated extensively below).

If emotions are modelled continuously (see for example [37]), then there could be a more varied set of connections between foundational meanings and inputs, or teacher dependent correlations between reliability and certain vectors.

Social signals could also be used to deal with ambiguities that arise when learning from other sources. Take for example the problem of interpreting silence from a teacher pressing evaluative buttons, where one could try to use facial expressions and eye gaze to separate the cases where the teacher: (i) decides that the learner is hopeless and simply gives up, (ii) is satisfied with performance and therefore stops giving feedback, (iii) gets bored, or (iv) becomes distracted by something unrelated to the interaction.

In [42], EEG signals are collected in a "Wizard of Oz" type experiment, and is then used offline to differentiate between two separate cases: (i): the case where a human is initiating eye contact with the robot, and (ii) the case where a human is responding to a perceived robot effort to initiate eye contact. When a teacher and a learner is engaging in joint attention activity, the interpretations of the actions of a human (communicative or otherwise) could for example be dependent of whether or not the human is the one that initiated the eye contact. This work is an example of making progress in a novel way on an interesting research front that could be useful to building better learners in the long term.

The state-of-the-art is not close to a situation where a robot can get an understanding of what is going on inside the head of a human in the way that another human can, and there is no obvious way of constructing a roadmap to such a point. But the work of [42] does show that this is an area where it is possible to extend current abilities. In that sense it is a good example of a case where the present survey seeks to point to open research directions; areas where progress can be made with respect to the current state-of-the-art, as opposed to pointing to gaps in a well defined roadmap. If the outlines of a fully functional system is known, research opportunities can be described from the point of view of this outline. But if the outlines of a fully functional system is not known, the only point of view available is a set of frontiers along which the current state-of-the-art can be extended. Research projects are thus described as extending a boundary, as opposed to filling gaps.

For humans, normative rule adoption involve a diverse set of information sources and interaction types. Artificial learners are of course not restricted to this set (and can for example learn from EEG readings), but the set of normative rule adoption pathways that humans use provide important sources of inspiration that comes with a free, proof of concept, biological implementation. One natural form of interaction that plays a part in human normative rule transmission/adoption/creation is story telling. This avenue of research is still in the early stage of development, but proposals to learn values from written stories exists [43].

If we extend the idea to stories in general, then this path to normative rule adoption seems to have a very long history. In [44], it is suggested that story telling was an important factor in human evolution. In the account of [44], solving the need to explain something that is not going on right now, using methods such as pantomime and pointing, is intimately linked to the evolution of things such as joint intentionality, normative rule adoption, and language.

The general class of setups involving story telling include instances that are very far from the usual interaction environment covered in this survey, with a teacher trying actively to teach something to a learner, or an active team effort with back and fourth communication. To make a point about the range of information sources and setups that are possible, let's look an extreme edge case: consider an artificial learner trying to extract values from stories written in ancient Egypt, learning from a "teacher" that probably did not foresee this situation. Even this case would however fit perfectly in the present survey as an instance of a learner trying to figure out what a human teacher would like it to do, where instances involving a teacher that is unaware of what is happening is "only" a practical problem. Leaving the edge cases, and trying to find more practical versions, we could for example imagine a teacher selecting, modifying, or writing stories explicitly for the purpose of demonstrating some set of values to an agent. This could for example be done to avoid issues where we would not want the learner to adopt some fairly common story behaviors that might be present in randomly chosen sets of stories (for example the implicitly approved of behavior "employ drastic, dangerous, and unconventional strategies to avoid damage to yourself", employed by various types of protagonists dealing with various types of antagonists). This information source could of course also be combined with other information sources, and be part of a diverse set of back and fourth interactions involving clarification questions, demonstrations, feedback on actions, etc, etc. This research is still in a very early stage, and there are many open problems that needs to be addressed, but [43] helps to illustrate the diversity of the types of information sources available to an artificial mind adopting normative conventions. An important research frontier consists of extending the set of information sources that can be used for learning. The first system that learns from a given novel information source, such as [43] learning from stories, both extends the frontier of the larger project of building learners in general, and opens up a new frontier; dealing with all the technical issues that are unique to the specific new information source.

In [45], an artificial learner benefits from natural language explanations provided by a human teacher in the Mario game. Sentiment analysis helps the learner interpret the explanations, which provides action advice related to a type of object (enemies, coins, etc). If a learner knows what object, and what action related to that object, that a teacher is talking about, the learner still needs to figure out if the action is recommended or warned against ("it's not a good idea to fall into chasms" vs "you can miss out on points if you don't jump to catch coins"). The experiments showed that only doing sentiment analysis on the level of individual words is not enough, and that analysis of larger segments is needed. After sentiment analysis, the learner has a set of actions linked to objects that the teacher has either warned against or advised the agent to take. Linking actions to objects instead of states means that the data generalisation problem can become easier, and it changes the task of the non expert human teacher (who might be more used to talking about objects than states).

To further demonstrate the diversity of possible information sources, it has been suggested that a learner can treat the fact that a human teacher is trying to switch off the learner as an information source [46]. In [46], there is a focus on the off switch attempt as a safety measure, but the off switch attempt is treated as an information source to be interpreted, and this idea seems more general. A teacher saying "switch yourself off" (or trying to physically switch off/disable/constrain a learner) does seem like a relevant piece of information. As pointed out in [46], if the teacher is reasonably rational and well informed, then this can be interpreted as indicating that the learner has a bad model of what the teacher would like it to do. There are many possible reasons for why a teacher might try to switch off/disable/constrain a learner; maybe the teacher and learner disagree on how to achieve some specific goal, or maybe the learner has missed some side effect of its actions, or maybe the teacher has not noticed the problem that the learner was trying to solve, etc. There is a range of possible responses that the learner could take, such as trying to clear up the misunderstanding, re evaluate its model of what should be done, act more carefully until the issue is resolved, etc. In the more general case, it might not always be easy to know whether or not a teacher is trying to constrain/stop a learner; if the teacher takes a tool from a learner for example, this could be an attempt to stop whatever the learner is doing with the tool, but could also be due to the teacher simply wanting the tool, and it might be worthwhile to look for social signals that could help differentiate between the two situations.

This field of research is still in its early stages both practically and theoretically, and making incremental improvements over the current state of the art seems like a promising open research avenue. In particular, it should be possible to build systems that outperform default strategies such as ignoring, or treating as interference, the actions of a teacher that tries to switch off/disable/constrain a learner.

As mentioned above, if we interpret human evaluations as referring to action choice, then we can build new algorithms based on this new interpretation, such as Policy Shaping [5]. In [6] the algorithm is shown to work well with human teachers. The impact of different sets of instructions given to teachers is examined, and different ways of interpreting silence is explored.

How a learner might infer preferences based on actions that are good, but only within the context of certain limitations has been discussed in [47]. A learner might for example be faced with inferring the preferences of a human teacher that takes a detour to avoid walking past a donut shop. This could be an good action choice if a teacher limitation prevents the teacher from reliably being able to resist donuts (the action of taking the short route, and not stop at the donut shop, might be unavailable).

Efforts to find a good interpretation of a human teacher could also leverage the fact that humans can be flawed or misinformed in ways that can be predicted, and that can be taken into account during learning. If part of what happened is not visible to the teacher, but is visible to the learner, then the learner could in principle take this into account. In [48], a way of dealing with a certain type of flawed teacher is proposed. The learner builds an explicit model of what the human teacher thinks has happened, based on what is visible to the human teacher, and takes this model into account during interactions. The topic of learners building specific models of exactly how a particular human in a particular situation is flawed, and then using this in learning, would work well as a separate section in this survey, but it is a severely under explored topic.

If the adoption of linguistic rules is viewed as a special case of normative rule adoption, then we can treat language learning systems as part of this general research project. The concurrent learning of linguistic and non linguistic tasks is for example explored in [49]. The learner/imitator observes several interactions between two humans. There is one interactant that might perform some form of communicative behavior, and one demonstrator that shows what the appropriate response is to the current context (where the behavior of the interactant is treated as part of the context). The inputs and outputs are all continuous and the interactions are not labeled, meaning that the imitator does not know how many words are expressed in a given set of interactions (it could be two words spoken many times each, or several words spoken a few times each, and some interactions might contain no relevant linguistic information at all). A single imitation learning strategy is proposed for dealing with all tasks in this context, that does not need to be told which demonstrations are of linguistic tasks. Using a single imitation learning strategy that simply adopts normative conventions (of the linguistic, or the non linguistic, kind), and that does not rely on labels or symbolic input, also makes it possible to avoid/dissolve the “symbol grounding problem” (there is no symbolic language learning system whose symbols need to be grounded in a separate action learning system).

In [50], the problem of segmenting human behavior and recognising individual tasks is tackled. It also deals with the problem of transferring skills to a robot (the algorithms are implemented on an iCub and the system is fast enough for real time interactions with a human). The work focuses on estimating the goal of the human, and uses a semantic representation to encode this goal (as opposed to the reward function estimates of IRL). The goals are specified in a non fixed ontology, and an ontology extending algorithm is presented (allowing the learner/robot to infer goals that were not representable in the original ontology).

Learning from demonstrations is a large and diverse field of research, and has recently been covered in a comprehensive way [51], and as a result the present survey will not attempt an exhaustive analysis of implemented systems of this type.

4.1 Autonomously improving interpretations

An increasing amount of work has started to deal with the question of how to autonomously improve the interpretation of various teacher signals [41, 52–58].

In [52], a learner updates interpretation hypotheses of a teacher's comments, described as an extension of inverse reinforcement learning. Discrete symbols are used and the learner starts with a partial lexicon. The symbols refer to things such as: "go right", "yes", "no", "good robot", "go left". The learner starts with an incomplete model of how to interpret this feedback, knowing the meanings of some of the symbols, but not others. The learner updates a model of what the symbols mean, as well as a model of a reward function used to describe the task. Knowing the reward function makes interpreting the teacher easier, and knowing how to interpret the teacher makes finding a reward function easier, so both are updated concurrently during learning.

Some of the limitations of [52] were relaxed in [53], including the discrete labels and the partial a priori lexicon. In [53], the agent was simultaneously learning how to interpret teacher generated raw speech signals and learning to solve a sequential decision problem.

EEG is an example of an information source where the input looks incomprehensible to a human, and where individual teacher models are clearly needed. In [41, 54], agents autonomously estimate the meaning of these signals during learning. Due to the fact that raw EEG readings do not have obvious meanings to a human observer (no matter how they are represented), this is a good opportunity to see the world from the perspective of a learner and illustrate how important good interpretations are. Imagine trying to figure out what a completely alien type of mind would like you to do, based on some interpretation of EEG readings as a response to your actions. How could you tell if your interpretation was wrong? Wouldn't it be nice if whoever constructed the interpretation had taken some care to make sure it is actually reasonably accurate?

In [59], an agent learns from speech utterances of a human teacher. When the model of the teachers goal improves, this model can in turn be used to learn new words (it is easier to guess the meaning of a word if one knows what the speaker is trying to achieve).

Electromyograph (EMG) signals are used to control a prosthetic arm in [60]. The meaning of electrical signals are person dependent, and the tasks that an amputee might want to perform can change, making the ability to continuously re interpret the signal desirable. This is achieved by letting the human provide feedback interpreted as rewards. The interpretation of the feedback is static, but the learner and teacher are engaged in a collaborative effort to re interpret the EMG signals. The situation where a robotic learner is collaborating with a human teacher that is simultaneously generating both control and feedback signals is also explored in [61].

A series of experiments with human and simulated teachers is presented in [58], exploring the connection between teacher strategy and silence/implicit feedback. If we view human behavior as generated by a stochastic transform from things such as teacher goals and learner actions into things such as feedback or silence, then silence is easy to see as just another thing to interpret. Learners use the I-SABL algorithm, which estimate the teachers training/feedback strategy, and are thus able to utilise information efficiently for a number of different teacher strategies. There are many factors that might impact the type of training/feedback strategies a learner will encounter. It could for example be dependent on the specifics of the current (i): task, (ii): embodiment of the learner (Is the robot cute? Does it look sad when given negative feedback?), (iii): stage of learning (Does the teacher sometimes give up on a learner making slow progress? Or sometimes think "it is done learning"?), (iv) type of interaction setup (Does the teacher for example have access to a separate "motivate button" that could divert the impulse to "encourage" the learner away from the feedback button, as in [62]?), (v): instructions given to the teacher, as in [63] where a teacher is told what type of demonstrations will be useful to the learner. See also [6], where different instruction conditions were tested, (vi): learner behavior (see for example [3, 24] where learner behavior is shown to influence teacher behavior), (vii): teacher time constraints (Is the teacher giving feedback to a set of teacher selected learner actions, or while watching a series of learner actions? Is the interaction so slow that the teacher becomes bored and only occasionally pays attention, or so fast that the teacher only occasionally have time to press buttons?), (viii): teacher level of understanding (Is the teacher silent because it knows that it is missing a lot of information, or is the teacher silent despite knowing exactly what is going on?), etc.

In [64], a learner builds an interpretation of a teacher's facial expressions and uses that interpretation during learning of a grasping task. Adding the facial evaluation improved performance in an experimental condition where objects were not repeated during successive trials. The teacher had a single button that gave negative reward to the learner, and the interpretation of facial expressions is built as a predictor for button pushes. A smiling face for example means that completing an initiated grasp action will not result in the button being pushed. Using a single button with only negative rewards bypasses two different problems that can arise when interpreting evaluations as rewards. If a teacher provides more positive rewards for fixing a problem than for causing it, this can lead to positive cycles, where an agent accumulates rewards without actually making progress. Similarly, if a teacher will both provide on average more positive than negative evaluations, and will also stop giving evaluations when the learner is done, then dragging out learning can increase total rewards. The setup of [64], with only negative rewards, bypasses these two specific issues. But the question remains whether or not the implicit interpretation embedded in treating negative evaluations as negative rewards is actually a reasonably good approximation of what a human teacher means when pressing a minus button. Outside the contexts of scripted interactions, incorrect interpretations can lead to other complications. Any strategy that makes a human teacher avoid interacting with a learner would for example be seen as a success for a learner with this interpretation. Aside from attempts to map out specific failure paths, the fundamental question is if there exists a better interpretation when a human teacher pushes a minus button: for example that the action was wrong (an interpretation that does not imply that clever ways of making a teacher want to avoid the learner should be seen as a success).

The difficulty of the problem faced by learners autonomously reinterpreting teachers can be strongly influenced by how appropriate the parameter space is, or how well the priors over interpretation hypotheses fits with the actual behavior of the human teachers that the learner will encounter. This aspect can be improved by a systematic study of how humans actually teach artificial learners. Studies of human teachers (covered extensively below in section 5.1) are therefore not only relevant to designers of static interpretations, but also important to designers of systems that autonomously re-interpret information sources.

4.2 Improving the information content of interactions

It has been shown that human teachers will attempt to modify feedback behavior to suit a learner [3, 24], opening up an entire range of possibilities for clever learners. This is true for different types of learners. If the learner is trying to maximise rewards, then it could for example try to look confused or uncertain/questioning when it knows that it has done the right thing (so that the human gives feedback). It could also try to look sad if it thinks that it has failed, in an attempt to avoid negative feedback. Other types of strategies are useful for a learner of the type that we are concerned with in this survey. There is a whole range of behaviors that can help the learner figure out what a human teacher would like it to do, for example looking confused when it needs feedback to know whether or not it has done the right thing, or trying to avoid looking sad (in an attempt to avoid the problematic situation where positive feedback is used as encouragement).

Learners modifying teacher behavior is described in [65] which investigates active learning with the specific viewpoint of how a learner can act to maximise informative observations during interaction with a human teacher. As mentioned above, in [24] a robot learner influences the way a human teacher gives movement demonstrations, showing experimentally a learners ability to actively influence the interaction in a positive way (the conceptual basis in [24] is seeing the learner as a team member, pursuing a joint goal with the human teacher). In [66] a learner displays its current uncertainty in a way that helps the teacher give the type of demonstrations that are most suited to the current needs of the learner.

A related team strategy for facilitating understanding is for a learner to act in such a way as to make it easier to see what it is doing [67, 68]. For a learner that is trying to be legible while interacting with a human teacher in an unstructured environment, it might be a good idea to take into account the fact that the teacher might have a limited ability to observe what the learner is doing [69]. See also [70] for an algorithm that tries to adapt to a human during cooperation, actively changing its behavior to be more legible.

In general, improving how smoothly an interaction flows can be valuable. See [71] for a proposal to implement systems whose attentional behavior resembles that of humans, partly to facilitate natural social interaction of a type that humans are used to. If the robot attends to things in ways that are similar to how a human attends to things, human-robot interactions might be more natural. In addition to the practical advantage of using already established interaction protocols, this might increase the probability that the human participates in joint intentionality activities, or even the degree to which the human tends to conceptualise the robot as a person. For a survey of joint attention in the context of socially interactive robots, see [72]. The focus of the survey in [72] is less on the learning of joint attention skills, and more on joint attention skills as a basic building block that is essential for more complex social skills. Joint intentionality human-robot team efforts would be more difficult if the robot is not displaying the types of behaviors (or possess the types of social skills) that the human expect from an interlocutor, such as acknowledging joint attention.

Another example of an agent acting to improve the information it gets from interaction is described in [73]. An agent modifies its' speed based on how uncertain it is regarding what it should do. Static speed agents needed either more time, or more actions, to learn the task. In [73], human behavior is treated as valuable information to be used for learning, not as a value to be maximised, and therefore it makes sense to attempt to elicit feedback in precisely those situations where the learner does not know what to do (from a "positive evaluation maximisation" perspective on the other hand, it would make more sense to elicit feedback in situations where the correct action was confidently known, and avoid feedback in cases where the agent might make mistakes). In [73] slowing down action execution in times of high uncertainty lead to increased performance. There are many possible explanations for this, for example: giving the teacher more time to give more or better feedback in cases where it matters most, the teacher interpreting the delay as hesitation, making them more likely to give feedback, or even interpreting the slowdown as a request for feedback. The fact that teachers preferred to interact with static speed agents raises interesting questions about priorities during human robot interactions. These priorities might be task dependent, with some considerations dominating in day to day interactions with a domestic robot, and other considerations dominating in interactions involving a bomb disposal robot.

In [74], an algorithm for choosing actions that allow a learner to estimate the internal state of a teacher is proposed, and tested on a driving task where the learner estimates the driving style of a human. The specific implementation in [74] is a driving task, but the algorithm is general in terms of the type of domain as well as in terms of the type of internal state. In the general formulation, something about a human is not directly observable, but does have an impact on other things that are observable. This internal state can for example be a driving style or a goal. In some situations, different possible internal states predict different observations. A learner can deliberately create such situations (that is, it can take actions that tells it something about an internal state of a human). This internal state is approximated as being static, in particular, it is not modified by the learner's actions. In the driving task, it would for example be possible that a human becomes annoyed at the information gathering actions of a learner, and change driving style as a result. As suggested in [74], it would be interesting to relax the assumption about a static internal state. For the types of learners covered in the present survey, what to do is dependent on an internal state of a human teacher, and this state is not directly observable. When what to do is dependent on internal states that are not directly observable, and that are potentially modified by the information gathering actions of a learner, non trivial theoretical, as well as practical, problems arise. This seems like a promising open research frontier, both theoretically and practically, because it is possible to make progress even without a perfect model of how information gathering actions will impact internal states (to improve on the current state of the art, one only needs to beat the default "zero impact model").

4.3 Summary

What type of information sources a learner should utilise, or how the information should be collected, is still an open research question. Social Signal Processing is an active research field [34–40, 64], and offer a large set of potential information sources. There is no clear answer to the question of which information sources are most suitable for a learner, and research is ongoing on a range of different types of systems, such as EEG [41, 42], stories [43], explanations [45], evaluative feedback [5, 6, 46, 58, 61, 62], demonstrations [9, 10, 15, 20–22, 26, 47–51], and facial expressions [64].

Recently, an interesting research field has opened up, investigating different ways in which a learner can autonomously re-interpret a human teacher [41, 52–60].

Finally, a strand of research is focusing on how the quality of the information that the learner has access to can be increased, for example through active information gathering actions, or through making the interaction between teacher and learner run more smoothly and thus generate more useful information [3, 24, 65–74].

5 Studies of biological systems

Studies of biological systems can be interesting for two distinct reasons. All learning algorithms of the type explored in this survey are built on top of interpretations of human teachers, either explicitly or implicitly. Investigating the behavior of human teachers as they interact with an artificial learner is therefore essentially a matter of trying to understand the thing that is being modelled.

Secondly, we can also be inspired by the strategies being employed by biological learners, human or otherwise. Some biological learners are more relevant than others, and an effort will be made to describe and distinguish between the different types of learners. Humans are uniquely prone to engaging in the normative rule adoption type of learning that we are concerned with, and will be contrasted with other great ape learners, especially chimpanzees. Other types of minds, such as the parrot Alex [75], share an interesting epistemological position with our artificial learners in the sense of trying to understand a fundamentally different type of mind, and not being able to gain much information from asking: "if I did that, what would it mean?".

5.1 Human teachers

Studies of human teachers can be done with or without an implemented artificial learner that the human teacher interact with. The reaction or performance of a specific learning system when interacting with human teachers can tell us things about those teachers. The back and fourth interaction between system and human can tell us much about humans that are hard to learn without an implemented system. Therefore this section will cover research that use implemented systems, but our focus here is on what these experiments tell us about the humans, rather than on covering details about the implemented systems themselves.

Three complementary research directions in the study of human teachers are especially relevant for our purposes. First is the experimental paradigm trying to figure out how humans actually behave, leading to better models if assumptions are static and better initial interpretation hypotheses/better parameter spaces if assumptions can be updated. Second is trying to figure out how various learner actions influence a teacher in a social situation, something that can lead to better feedback (for example displaying confusion or understanding in order to help people give better feedback). The third is how the behavior of teachers can be modified by researchers so that they give more useful feedback (for example explaining to teachers what types of demonstrations will be useful [63], or giving evaluators a dedicated button for motivating/encouraging a learner, in order to make the teacher stop using a feedback button for this purpose [62]).

It has been repeatedly shown that interpreting human behavior can be problematic, for example because humans make errors, and their behavior violates many assumptions of common machine learning algorithms in general, and reinforcement learning algorithms in particular [1–3, 62].

In several studies [3, 40, 62, 76–78], human teacher intentions were investigated, showing, for example, that teachers attempted to include multiple communicative intents in a single channel, that negative feedback had a different interpretation than positive feedback, and that teachers sometimes gives positive feedback attempting to guide future actions. Humans also tend to give more positive than negative feedback, even in the very beginning of a learning episode, before performance is good (showing that the bias towards positive feedback is not caused by high learner performance). Various ways of modifying the algorithms to better fit the actual intent of human teachers were shown to improve learning.

Humans might use positive and negative feedback in qualitatively different ways [76], which means that an artificial learner can learn different types of lessons from these two different types of feedback. In [76], the learner reversed an action as a response to negative feedback, starting over from a previous state, for example returning an object to where it was before. This particular strategy requires that the learner is able to reverse states, but the basic idea that a learner does not have to respond symmetrically to positive/negative feedback seems more general.

In the work of [77], teachers frequently gave a reward designed to guide future actions, despite the fact that the signal was used as a standard reward. Human teachers will also, for example, use a mechanism meant as a channel for evaluating actions to try to motivate the learner. One can improve performance by including a button in the setup, dedicated to motivational communication [62]. The "motivate button" reduced the tendency of humans to use the evaluation channel for motivation purposes, and thus brought the actual human behavior closer to the assumptions of the learning algorithm used (the learning algorithm did not model the evaluation button as encouragement). If a learner has access to a policy update algorithm with a built in assumption that the teacher is all knowing, then one way to reduce the gap between model and reality would be to delay an action until the teacher is paying attention.

Humans might engage in communicative/social behavior that does not say much about what they want, for example pointing out some unusual object without any thought regarding how this object could fit with the achievement of any particular goal. Interpreting these types of behaviours correctly is important in order to avoid misunderstandings. This is similar to a teacher feeling sorry for an unsuccessful learner and trying to cheer it up with the positve evaluation button. It's important to interpret even unhelpful behavior correctly in order to avoid misunderstandings/reduce noise.

On a more abstract level [3, 40, 62, 76–78] shows that specifying an interpretation of a human teacher is in general very difficult, and that better interpretations can lead to better learning. These studies show the need to pay close attention to building better interpretations, since flawed interpretations are likely, and improving them is both possible and useful.

See [79] for a survey of work using embodied artificial agents in order to learn lessons about human social interaction in general, and about how humans perceive and interact with robots in particular. The robot can act either as a model of a human, or as an interaction partner to the human, creating a situation for the purposes of examining the response of the human. Interactions with humanoid robots can leverage the fact that the behaviour of one of the interactants is possible to vary in a controlled way, and use this to learn something about human-human interactions. When examining human-human interactions, then there is a tradeoff between ecological validity and the ability to construct controlled experiments. When the primary focus is instead on examining human-robot interactions, the robot does not pose a problem of ecological validity. The survey in [79] has a neuroscience focus, and for example cover experiments involving methods such as EEG, fMRI, or PET scans to analyse humans engaged in various types of social interactions.

For some tasks it is possible to define an optimal teaching sequence, allowing a comparison with actual human behavior, as in [80], where humans were found to not behave according to those optimal teaching strategies. In [81], different interaction protocols were tested on non-experts to see how they perform in actual situations. An experiment presented in [45] showed that performance can be improved by designing an interaction so that the cognitive load of the human teacher is reduced. In terms of modifying the behavior of teachers, [63] shows that researchers explaining what types of demonstrations will be useful to the learner helps a teacher produce useful demonstrations.

As mentioned above, experiments have shown that human teachers tend to give more rewards for solving a problem than the punishment given for creating the problem [4]. The meaning of this behavior under the interpretations underlying an algorithm that maximise positive minus negative feedback would be that the human teachers wanted the learner to go around in a circle, which was far from the real intentions of the subjects in [4]. The observed behavior was found to be a better fit with the interpretation that feedback was a noisy evaluation of action choice.

In [82], several ways in which human teachers break implicit assumptions of many learning algorithms are discussed, further underscoring the usefulness of improving the interpretations of human behavior. In [83], a modification to the TAMER framework [84] is presented, designed to allow a learner to take advantage of feedback on future actions. This could allow the learner to take advantage of "no don't do that!" type feedback, if future intent can be displayed by the learner and understood by the teacher. Even when it is impossible to find out what specific imagined future learner action was evaluated by the teacher, determining that a particular evaluation was not an evaluation of any of the previously performed actions can still reduce noise. Building on these findings, [85] presents a study observing teaching behaviors in five different navigation tasks. In [57], studies of human teachers are used to build a parameterised model of how a teacher gives feedback. The values of the parameters are learnt autonomously during interactions with individual teachers, showing the strong connection between studies of human teachers, and the design of algorithms that update individual teacher models/interpretations based on observations.

In [86], an interpretation of human feedback as dependent on learner policy was suggested, leading to the COACH algorithm. An experiment with human teachers showed that the response to a given state action pair was dependent on the estimated skill level of the learner. The human teachers were both given a description of learner competence, and observed learner actions, prior to giving feedback. This correlation between feedback and estimated learner competence was shown to be distinct from tendencies to simply give less feedback over time. The interpretation/algorithm was also validated in a separate experiment involving a robotic learner.

The tricky question of how demonstrations should be interpreted when a human teacher has an incorrect model of their own behavior has been investigated in [87], in a driving task. Some human teachers believe that they have a defensive driving style, while in reality they actually have an aggressive driving style, and they would like an artificial learner to drive defensively. Such a teacher believe that they would like the learner to drive the way that they drive, but in reality they would actually like the learner to drive more defensively than they drive. Another way to look at this is that the teacher would like the learner to drive in a way that corresponds to how the teacher believes that the teacher drives. The study in [87] is an example of how one can disentangle tricky issues involving teachers with a limited ability to model their own behavior. The question: “does the teacher want the learner to drive as the teacher drives?” is tricky, but the question of how the teacher would like the learner to drive is actually more straightforward (they want the learner to drive defensively). The learner only needs to know how to drive, and in this case it is possible to figure that out without fully working out the theoretical questions about what limited teachers actually want.

One way to deal with the fact that some human teachers would like the learner to drive the way the teacher thinks that the teacher drives (but not the way the teacher actually drives) is to sidestep the problem by asking for comparisons between choices. In [88], an algorithm is presented that chooses a set of actions from a continuous action space, which are then presented to a human teacher who chooses one preferred action. What actions the human is presented with is based on how much information the answer will provide. This setup sidesteps the tricky question of whether or not such a teacher would actually like the learner to imitate them. More generally, it can be used in situations where teachers are having problems performing the types of actions that they would like the learner to perform. This inability could be due to the fact that the teacher incorrectly believes that it wants to be imitated, but it could also be due to any number of other reasons, including that the task is dangerous, or that the teacher and learner have very different types of embodiments.

In practical terms, the setup in [88] requires that the learner has the ability to generate and display at least two action candidates. As mentioned throughout the survey, there are numerous difficulties involved in interpreting evaluations from human teachers. Some of these difficulties involve trying to determine what action a given favoured/unfavored action is compared to. In the case of teacher feedback on a single learner action, the evaluated action could for example be compared to: the learner’s previous actions during the current interaction, or the action that the teacher would expect from an artificial system, or the best possible action, or the action that the teacher believes to be best, or the action that the teacher would have taken, or the action that the teacher thinks that the teacher would have taken, etc, etc. If a teacher chooses an action from a known set, this interpretation problem is mitigated. More generally, asking for a choice between specific alternatives can reduce some types of ambiguity that arise when a human teacher is unaware of its own limitations. Experiments with an implemented learner also showed that the active learning algorithm presented in [88] works well in practice when interacting with human teachers.

The type of data generated in a setup like [88] is in some ways similar to the type of data one would get by asking a teacher for Explicit Action Advice in an environment with a limited set of discrete actions that is known to both teacher and learner. In both cases the resulting data is in the form of the expressed preference of a teacher over a known set of discrete action choices.

5.2 Biological learners

When dealing with human teachers, it can be useful to understand unhelpful or counterproductive behavior, even if this understanding only results in the learner ignoring some signals (thereby reducing noise). In the case of biological learners, we are however only looking for inspiration, and should therefore focus on behaviors that are useful for our specific purposes. It would not be surprising if many behaviors of evolved systems turn out to be unsuitable for an artificial system that is supposed to do what another mind would like the system to do. This is a diverse, large, and old field of research, see for example behaviourism for some influential early work [89, 90] (basically the idea that for some biological systems, positive/negative feedback can increase/decrease the frequency of certain behaviours). An exhaustive overview of every relevant type of learning strategy, employed by some biological system, that is potentially interesting as inspiration for an artificial system, is beyond the scope of this survey. This subsection will instead focus on answering the question of what types of biological systems are most relevant as inspiration for our particular type of learner.

Normative rule adoption in biological systems is directly relevant as inspiration for a learner trying to figure out what a human teacher would like it to do. Language learning can be seen as a special case of normative rule adoption, meaning that we can use strategies for language learning as inspiration for our artificial learners. Joint intentionality is also an important interaction type, especially when viewing the learner as a team member, trying to achieve a joint team goal, together with the teacher. Discovering and repairing misunderstandings in the context of a joint intentional activity is something that humans can do, and that would be very useful to a learner of the type we are concerned with.

Several pro-social mechanisms that reduce competition and enforce cooperation is less relevant, since there is no goal conflict to suppress for the type of learner we are concerned with. Strategies related to competition, and interactions where a biological system is learning to extract something from another interactant, are also less interesting. Strategies for more general purpose/goal independent social learning (for example discovering how an object works by observing someone manipulate it) are relevant for many types of learning systems. The focus of this subsection will not be on this type of learning, which is relevant regardless of what the system is trying to do. The focus will instead be on trying to explain what types of biological systems are relevant as inspiration, specifically for figuring out what a human teacher would like the learner to do.

Shared intentionality interactions [44, 91–96] are in humans closely related to normative rule adoption, including language acquisition [97–99]. When considering learner and teacher teams that are jointly trying to add the teachers goal to the common ground, then biological learners engaging in shared intentionality interactions can be a very useful source of inspiration.

Two or more individuals are engaged in a shared intentionality activity if the goal of each individual is that the group succeeds, and if this fact is common knowledge (the goal, and the fact that everyone share this goal, is part of the common ground). A simple example is the situation where two people are jointly lifting a sofa, and where it is obvious that they are both working towards the same goal. This situation is importantly different from two people, each holding one end of a sofa, and each having the same goal position of the sofa, but neither one having any idea why the other is holding the sofa, or what the other is trying to do. It is interesting to note that some acts can be seen as somewhere in between communicative and non communicative, for example someone starting to very lightly tilt the sofa. In the case of a shared intentionality activity, this can be interpreted by the other team member as “we should hold the sofa in another way”.

In [100], the strongly related concepts of “we-mode” versus “I-mode” and collective intentionality are discussed. See also [101] for an early discussion of different ways in which the meaning of a sentence is understood, for example due to an extensive common ground.

An evolutionary account of human morality and normative rule adoption is presented in [99]. Human normative rule adoption has co evolved with several other behaviors. Some of these are highly relevant as inspiration for artificial normative rule adopters, such as the joint intentionality described above. Others are less relevant, such as punishments for individuals that abandon a joint project (internalised and directed towards the self, or directed at an interaction partner).

In the account of [99], this process starts with joint intentionality, where two individuals engage in a collaborative team effort, for example hunting. The next evolutionary step is joint commitment, where each team member has a responsibility to the team to continue an effort. Enforcement mechanisms related to joint commitment are not necessary to implement when there is no goal conflict. At an even later evolutionary stage, Tomasello argues in [99] that human ancestors faced evolutionary pressures related to conflicts between tribes. Consider two tribes, each too large for it to be possible to know all tribe members personally, engaged in a type of conflict where there is an advantage to being able to coordinate and function together with all other tribe members. The tribe that adopt normative behavioral rules will end up behaving in similar ways, and thus have a way of recognising each other, and therefore coordinate. There is no incentive for an individual to “defect” by refusing to adopt normative rules, as this would risk being classified as “not part of the tribe” by fellow tribe members. The evolutionary pressure is towards doing things for the sole reason that “this is how things are done in this tribe”. This is a different evolutionary pressure than the pressure that comes from the benefits of learning purely functional strategies from imitation (doing something because it worked when someone else did it). The adaptation to this pressure, to adopt normative conventions simply because “this is how things should be done”, is indeed exactly what we would like the learner to do.

Understanding the evolutionary pressures behind various interesting adaptations (for example joint intentionality and normative rule adoption) can give us a better understanding of the adaptations themselves. But for system designers that are seeking inspiration from biological systems, these pressures are not strictly speaking necessary to understand for their own sake (in principle, a model of a biological system does not even have to be accurate to serve as inspiration for an artificial system).

When drawing inspiration from humans engaged in joint intentionality activities, it might not be necessary to know why they are maintaining a joint goal (which could for example be due to strategic reasons, or due to joint commitment). As long as humans are secure in joint intentionality (for whatever reason), their coordination strategies can be used as inspiration. The fact that there is no need for mechanisms of goal conflict resolution means that, in some ways, the type of interaction explored in the present survey is simpler than the problems that humans had to adapt to. The lack of goal conflict means that some complex mechanisms does not have to be implemented, such as internalised second-personal sanctions, which play a part in human interactions [99]. That the goal is not only shared, but also part of the common ground can be used by both team members when interpreting the other mind. Each team member knows that this is common knowledge, and behaviors of the other can be understood based on this knowledge (the fact that they are both working together, trying to give the learner a better idea of what should be done, is part of the common ground).

The tendency of humans to adopt normative conventions has also been investigated in the context of feedback [7]. Unlike most work on biological learners, [7] analyse learner feedback response in terms of what lessons we can learn about appropriate interpretations of human teacher feedback, and what this in turn can teach us about the design of artificial learners. Humans and many other animals can be modelled as building an estimate of the value of certain behaviors by observing environmental rewards over time, but [7] argue that humans respond differently to feedback in a social context. [7] argue that humans do not interpret feedback in a social setting as an environmental reward that should be maximised, and instead offer two alternative interpretations. First, it can be interpreted as stating that an action is instrumentally useful/harmful. A teacher might want to encourage/discourage an action that is useful/harmful, especially when figuring out this usefulness is difficult for the learner, for example because some outcome is not immediate, or not immediately observable, or not frequent (as with seat belts and car crashes), or where the causal relation between the action and the outcome might be difficult to detect (as with eating candy and damaged teeth), or because the action is only the first step on the way to something useful. The second interpretation suggested is that the evaluated action is intrinsically good/bad. As opposed to the interpretation of feedback as a reward to be maximised, neither of these two interpretations incentivise a learner to avoid negative feedback (for example by hiding mistakes/transgressions or avoiding a teacher that is more negative than positive), or to seek positive feedback (for example by deliberately creating a problem for the purposes of getting positive evaluations when solving it). These interpretations seem well adapted to the findings mentioned above regarding the way human teachers tend to give feedback [4] (we would in general expect reasonably good agreement between what human teachers mean, and what human learners understand, but scientifically the two questions are distinct, and each question can be addressed by its own dedicated investigations).

These two different interpretations of feedback, as indicating instrumentally useful actions or intrinsically good actions, is related to the distinction made throughout the survey between two different types of social interaction. On the one hand we have the type of social interaction where an agent tries to extract something from a teacher, such as knowledge about how the world works, or in other words: learning about instrumentally useful/harmful actions. And on the other hand we have the type of social interaction where an agent tries to learn normative conventions, or in other words: learning about intrinsically good/bad actions. This is a distinction regarding what the learner adopts, and not directly a distinction about teacher intentions. A learner can infer and then adopt normative conventions based on learning that some specific action is instrumentally useful (usefulness is conditioned on, and thus related to, goals). Learning about instrumentally useful actions can be useful for many different types of artificial learners, pursuing many different types of goals, including the kind of learner we are concerned with in this survey. Figuring out what normative conventions a teacher would like a learner to adopt is however especially useful for the types of learners covered in the present survey. In the terminology of learning from demonstration, this corresponds to the distinction mentioned above; learning how to do something, versus learning what should be done [14].

How human learners do respond, and how artificial learners should respond, to various types of teacher feedback is a promising research frontier, with plenty of subtle open questions to address. A learner could for example respond asymmetrically to positive and negative feedback, as well as draw on facial expression and tone of voice analysis to distinguish between feedback indicating instrumentally and intrinsically good actions. It would for example be possible to treat positive feedback that is persistent, consistent, but non enthusiastic, as indicating instrumentally useful actions, and at the same time treat immediate, and instinctively produced, negative feedback accompanied by a certain tone of voice and facial expression, as indicating that the action is intrinsically bad.

Because of this complexity, the question of how an artificial system should respond to evaluative feedback might become entangled with things such as social signal processing: if an artificial learner is bad at detecting tone of voice, then perhaps it should not respond to an evaluation in the same way as a human learner (who is better at detecting tone of voice). A different tone of voice detector could therefore change the appropriate interpretation of evaluative feedback (not just make it more accurate, but make a new type of interpretation/adaptation/response appropriate, because it is now possible to distinguish a particular type of feedback from a more general class of feedback).

The question of appropriate response/interpretation of evaluative feedback is yet another question that might become entangled with the question of teacher competence and various sorts of teacher limitations. If the teacher is known to be giving instrumental feedback, then a learner might conclude that the teacher is incompetent based on some specific evaluation. The same inference would not be appropriate if the teacher is known to be giving intrinsic feedback. Conversely, if the teacher is known to be very competent, then the learner might conclude that a particular evaluation is meant as intrinsic (a parent that is fully aware that stealing is profitable, and understands fully that a particular theft was indeed risk free, is more likely to mean that stealing is intrinsically bad). The same inference might not be appropriate if the teacher is not known to be competent. An artificial learner might also be designed to try to solicit a certain type of feedback based on teacher competence (an incompetent teacher can still tell an artificial learner about intrinsically good/bad actions, just as it is possible to infer a goal from a failed demonstration). Recognising incompetence might save an artificial learner from incorrectly concluding that some specific action is intrinsically bad (if the action is known to be very useful, and the teacher is incorrectly assumed to be competent, the learner might incorrectly conclude that a negatively evaluated action is intrinsically bad). Determining if some specific behavior should even be interpreted as feedback might also be challenging when operating in an unstructured environment: a robot attempting to wake a sleeping human due to some emergency might for example encounter everything from conventional types of negative feedback, to threats and bribes, to attempts to switch off the robot.

It is important to remember that investigations of biological systems are interesting indirectly, for example by saying something about how a teacher should be interpreted, which in turn can be useful when designing artificial systems. One must however remember that something that is good for a human is not necessarily appropriate for an artificial system. When implementing an artificial system that adopts normative conventions from demonstrations, it is the responsibility of the designer/learner to make sure that conventions are adopted in a way that is actually appropriate for an artificial system. We might for example want to avoid a situation where an artificial learner starts to intrinsically value doing certain activities, such as watching tv, as opposed to helping the teacher enjoy those activities (see for example the CIRL formalism [26] discussed above, where a robot seeks to learn, and then maximise, the reward function of a human, without adopting it). When implementing an artificial system that adopts normative conventions from feedback, some of this responsibility is transferred to the teacher (if an artificial learner is watching tv, then it is up to the teacher to decide if this should be encouraged. If a human teacher is watching tv, it is instead up to the learner to decide if this should be imitated).

Returning to imitation learning, chimpanzees and humans are the two most studied great ape species, and there is an important difference in the way they interact and learn linguistic skills through imitation. This difference can be used to further clarify the distinction between the two different categories of social learning discussed throughout this survey (adopting normative conventions versus extracting something from the teacher).

Chimpanzees have very sophisticated social cognition and understanding, but they do not engage in shared intentionality type activities or normative rule adoption. Chimpanzees are capable of estimating others perception, knowledge, and goals. For example, when a chimpanzee wants an object that is controlled by a human, they can differentiate between the human not trying to give the object, and the human trying but failing [102]. They also understand that others make inferences [103] but not that others can have false beliefs (see [104] for a comprehensive overview of how chimpanzees model other minds). Since chimpanzees do not engage in shared intentionality activities, they can have trouble understanding humans that are trying to be helpful. To illustrate one way in which the lack of shared intentionality hinders chimpanzee communicative abilities we can look at [105], where a chimpanzee that is looking for food and knows that the food is in one of three locations, will not favour the location pointed to by a human. The chimpanzee follows the pointing to the location but does not assume that the human is trying to help it, and thus the location that the human is indicating is not assumed to be more likely to contain the food than other locations. If the human appears to be looking for food and tries but fails to reach the location, then the chimpanzee understands the behavior and favours this location.

Since chimpanzees does “action x gets me y” type imitation learning (as opposed to normative rule adoption), their gestures are learnt in the form: Chimp1 notices that when Chimp2 raises its arm, then Chimp2 will initiate play (raising the arm is a preparation to play-hit), then Chimp2 notice that raising its arm induces Chimp1 to start playing (this example is modified from [98]). Now Chimp2 knows how to initiate play with Chimp1 using a gesture. This type of learning enables two chimpanzees to establish a communicative convention, but it does not seem to result in the propagation of a large set of diverse linguistic conventions/normative rules within a population or across generations (learning to follow a rule for the purposes of avoiding punishments will not work here since there is no established enforcement mechanism, such that not starting to play when observing a raised hand results in punishments, and no clear path to the establishment of such an enforcement mechanism). In contrast, a human child might observe an interaction between two others, and imitate the normative rules of the “raised arm” convention (and as it passes through generations its meaning could change and/or become more general, for example due to imperfections in the rule adoption strategies used).

As mentioned above, it is possible to view the adoption of linguistic conventions as a special case of adopting normative conventions, where the context and/or the action space includes some linguistic component. If we adopt this perspective, then we can draw inspiration from a wider set of normative rule adoption strategies found in biological systems. Strategies for learning the normative conventions of language deal with the same ambiguities as learning other types of normative conventions. To illustrate the similarity of the problems faced by an imitator adopting normative rules for how to respond to some non communicative context on the one hand, and the problem faced by a language learner on the other hand, we will look at the classic gavagai problem [106] from linguistics. Quine gave an example with a man pointing at a rabbit and saying “gavagai” to someone that does not speak the language. The “gavagai problem” is to find out what the word means from these types of interactions (it could mean rabbit, but might also mean “dinner”, “what is that?” or “catch it!”). A similar type of ambiguity can be said to exist when interpreting non verbal, but social, physical actions. In the words of [107]: “the exact same physical movement may be seen as giving an object, sharing it, loaning it, moving it, getting rid of it, returning it, trading it, selling it, and on and on depending on the goals and intentions of the actor.”.

In [108, 109], the links between action and perception is explored, based on the hypothesis that sensorimotor skills/actions, social interaction skills, and linguistic skills develop in parallel and have a strong impact on each other. As argued in [108], a review paper establishing a research roadmap, a central challenge is the understanding of how language and action-perception learning and representations are integrated. This establishes a line between systems that learn to act (in a non linguistic way) and systems that learn to act (in a linguistic way), even though these two types of systems are usually developed by separate communities. Team actions and team communications are not necessarily two completely separate things. A jointly held sofa can be tilted, a failed action can be followed by the teacher re doing it properly (simultaneously indicating that the action was wrong, and correcting the mistake).

To fully illustrate the connection, let’s look closer at the importance of normative rule adoption for the emergence of communication. A learner trying to figure out what it should be doing based on the teachers performed actions is fundamentally different from watching the teacher for the purpose of building a world model. It needs to notice “people follow rule x”, as opposed to discovering “the tribe has a set of established norms, and enforces them in a way such that following rule x is beneficial”. It is easy to imagine how this could be true in a human tribe when language is already established (for example since non linguistic individuals may have trouble forming social relationships and/or have practical difficulties in cooperating). If it is beneficial for an individual to learn language, this can be seen as an enforcement mechanism (the tribe will treat those that adopt normative linguistic conventions better than those that do not).

If an agent adopts normative rules without needing to see a benefit, it needs “only” figure out what rule the others are following. An agent that needs to see a benefit for it to adopt a normative rule has an extra inference to make before seeing the point in adopting normative linguistic conventions, and it needs to be born into a group that is already linguistic enough that there is a benefit to adopting these rules. For this reason, the latter type of learning is less conducive to language, both due to the extra inferences required, and due to the fact that language can only be learnt when there is an enforcement mechanism already established. An agent that has no predisposition to adopt normative rules could in principle discover that it is beneficial to adopt the local linguistic conventions by observation (if it is born into a group that is already linguistic) but it seems more likely that an agent that unquestioningly adopts normative rules will do so.

5.3 Summary

Given that the human teacher is the source of the learners goal, as well as a potentially valuable partner in a joint project of getting the learner to understand this goal, investigations of human teachers are central to the project described in this survey. This is an active and diverse research field [1–4, 40, 57, 62, 63, 76–88], and is in general more focused on trying to find good interpretations compared to work on biological learners. While algorithms can be evaluated in experiments according to various success metrics, the study of human teachers can evaluate how appropriate those metrics are. In other words; while an experiment can confirm that a given algorithm receives a lot of positive evaluations, investigations of human teachers can be used to evaluate what humans mean with evaluations, and decide how informative the sum of evaluations are with regards to how well the learner has succeeded in figuring out what the teacher would like it to do.

While there exists research that investigate biological learners with the explicit goal of finding more accurate interpretations of human teachers, for the purpose of building better artificial systems [7], most work on biological learners is relevant in a less direct way. This is a diverse and old field of research [75, 89, 90, 102–109], and we need to be more selective because biological learners are only relevant as inspiration. One strand of biological learner research that is especially relevant for our purposes deals with shared intentionality [44, 91–101]. If we see the teacher and learner as a team, jointly trying to achieve an objective, then this work is indeed an important aspect of the teacher-learner setup. Shared intentionality is also an important concept for language learning, which can be seen as a special case of normative rule adoption.

6 Summary

When dealing with a situation where it is not possible to define what should be done at programming time, in terms of an agents inputs and outputs, an artificial learner will need to somehow infer what should be done on its own. The survey has dealt with artificial systems that infer what should be done by interacting with a human teacher, interpreting teacher behaviors, and trying to figure out what that teacher would like the learner to do. There are several different strands of research that contribute in some way towards finding better foundational interpretations of these teachers.

To summarise these various strands of research, table 1 is organised according to five types of contribution. An individual cited paper can contribute in multiple ways, and therefor belong to more than one category. Each category corresponds to one specific way in which a paper can contribute towards the goal of creating artificial learners that do what a human teacher would like them to do. This categorisation does not always have to correspond to the goals of the researchers, but instead lists how the work is relevant to our present objective. It is for example possible to draw inspiration from how biological systems learn, even when those biological systems were originally studied for their own sake, without any consideration of how this might inspire artificial systems. There is no obviously correct way of categorising contribution types, but to make a table one has to draw the lines somewhere, and the five categories chosen for table 1 are:

1. Formalisms and theory: These papers contribute to formalising the field, describe an open theoretical problem, or suggests a possible avenue of future theoretical research. Since there is no single and generally accepted theoretical framework, and many open problems, theoretical work constitute one of the major open frontiers.
2. Implementations of learners: These papers represent implementations that learn from humans in interesting ways, and an effort has been made to cover a wide spectrum of different types of interactions and information sources. In general, these implementations have a focus that is more towards learning what should be done, than learning how to do it.

3. Social Signal Processing: Not all implemented systems covered represent complete learner implementations, or deal directly with finding new foundational interpretations of teachers, but are instead focused on detecting various social signals. It is not clear which information sources will be most useful to a learner, and research in SSP is expanding the set of possible choices. It might also be necessary to deal with SSP concurrently with foundational interpretation issues for some social signals, because issues of interpretation and detection might be difficult to disentangle.
4. Interpretation of human teachers: We have described learning algorithms as interpretations of human teachers. Studies that directly investigate these teachers are therefore perhaps the most direct type of research covered. Because a learners of the type covered in this survey does not start with a goal, it is difficult to assess the success of a learner without dealing with foundational interpretation issues of human teachers. These studies are therefor important in both practical terms, and in terms of providing more accurate evaluation methods.
5. Inspiration from biological learners: Biological learners, both human and non human, implement an array of solutions to various learning problems, and studies of them can therefore provide inspiration for artificial learners. Since much research of this type is not done for the purposes of inspiring artificial systems, and since many biological learner behaviors are unsuitable for artificial systems of the type this survey seeks to cover, an effort has been made to figure out which types of research is most useful.

.

Table 1. All cited papers, organised by contribution type:

Formalisms and theory	**[3, 8, 11–14, 16–20, 23–33, 106]**
Implementations of learners	**[1–6, 9–11, 15, 20–22, 26, 40, 41]** **[43, 45–71, 73, 74, 76, 77, 83–86, 88]**
Social Signal Processing	**[34–40, 42, 72]**
Interpretation of human teachers	**[1–7, 23, 24, 27–30, 32, 40–42]** **[45–64, 73, 74, 76–88, 99, 106]**
Inspiration from biological learners	**[7, 12, 44, 75, 89–109]**

7 Conclusions

Interactions do not on their own tell a learner what a human teacher would like the learner to do without some sort of interpretation, either explicit or implicit. It is difficult to find good interpretations of human behaviors, and it might be tempting to simply implement an algorithm without thinking much about what implicit assumptions it is built on top of. This does however not address the interpretation problem, it merely makes it difficult to diagnose any problem that might arise from incorrect interpretations.

Several different efforts to formalise the problem has been covered, and this remains an important frontier for future research. One major theoretical issue is the fact that the human teacher might be limited/flawed/uninformed/mistaken in serious and systematic ways. There are many different types of limitations, which means that defining success is tricky. It can for example be complicated to define what success means if a teacher that is evaluating a learner is misunderstanding the situation. The problem of dealing with limited teachers is distinct from the problem of interpreting a teacher, even though these issues can become entangled. It can also be difficult to define success if the teacher is providing flawed demonstrations, either due to simply failing, or due to acting in a way that is good, but only within the context of the various limitations that the teacher is operating within. These difficulties are separate from problems such as determining what aspect of the demonstrations are relevant. A related issue concerns situations when the information gathering actions of a learner might potentially modify internal states of a teacher. These are promising open frontiers, with plenty of room for incremental improvements over the current state of the art.

Creating theoretically solid descriptions of this learning situation as a joint team effort of a teacher-learner team is also an important frontier, especially dealing with communication/interpretation problems in situations where the teacher is limited in various ways and the learner has no direct access to the team goal.

An interesting open research frontier for implemented systems is foundational interpretation issues on modalities such as facial expressions or tone of voice. Such research is especially interesting in the sense that it needs to solve many technical difficulties simultaneously with trying to answer questions of the type: “should smiles be maximised?”. How to interpret smiles, and other social signals, could indeed be dependent in complex ways on how smiles are being detected, how many different categories detected smiles are categorised into, which types of smiles can be reliably distinguished from each other, etc. How a learner should respond to a smile can therefor be dependent on a number of entangled issues, such as estimates of: the competence of the teacher, the limitations of the teacher, how human teachers in general use social feedback, how much attention the teacher is paying to the learner, as well as the specifics of the social signal processing system used to detect/categorise smiles. Simultaneously dealing with all these issues might result in combined rules of the type: "a learner should treat certain types of smiles in the same way as human learners do, iff the learner can both reliably detect joint attention, and also reliably filter out smiles that are sympathetic, polite, sarcastic, nervous, etc".

In general, foundational interpretation aspects of social signals is an interesting avenue of experimental research. One could for example put a human teacher in front of an artificial learner that is executing some predetermined set of actions. If the task is designed to be easy for the human to understand, and the actions are designed to be easy to judge, one could gather data from things such as tone of voice, facial expressions, etc, and analyse how a human teacher expresses various sentiments. We would then have a sort of reverse situation from trying to infer meaning from known signals; we would know the meaning (at least approximately), and would be trying to analyse the signal. To make a subject engaged, one could for example give them an evaluative button, or ask them to verbally instruct or critique the learner (which would be needed anyway for tone of voice studies). To avoid frustrating human teachers, it might be useful to explain that the learner will not learn online, and that instructions will only be used after the session is over. Such a study could provide answers to questions of the type; "what does a human look like/sound like/etc when the learner is, for example; failing completely/succeeding comfortably/barely avoiding a disaster/managing to limit the damage of a self generated problem/being too passive and failing to achieve some bonus opportunity/etc.

This type of research could both lead to better interpretations of a specific signal, and can also be used to select what signal to learn from (being able to detect and interpret the signal is not enough for learning; the signal also needs to have an interpretation that is actually useful for learning). It is difficult to know precisely what this type of experiment might result in, but it seems likely that results will be more along the lines of opening new frontiers, than along the lines of filling known gaps. In general, novel interpretations imply new types of learning algorithms.

Another open avenue is creating new types of frontiers by learning from novel information sources. It is difficult to say which unexplored information sources are most suitable for an initial implemented system. In principle, any pathway that has no current implementation, but that humans use for learning what should be done, is a candidate. Given that artificial systems can also learn from things that humans do not learn from, such as EEG readings, this is a very large and diverse frontier.

While studying human teachers with the goal of finding better interpretations is an active research field, experiments on biological learners are however usually not focused on finding inspiration for artificial learners based on understanding how biological learners interpret teachers. This represents an open frontier, and it would be possible to for example set up experiments trying to figure out how various teacher behaviors is actually interpreted, specifically targeted at generating new ideas for implemented systems.

Such experiments, specifically designed to find out how biological learners tend to interpret human teachers, would hopefully lead to genuinely novel ideas for interpretations; ideas that are not being explored for the simple reason that no one has thought of them. Any such interpretation would need to be verified and tested, for example with studies of human teachers, or with implemented systems interacting with human teachers, because the interpretations of biological learners are not always perfect, and their responses are not always appropriate for an implemented system. As results from this type of experiments would be relevant in the sense of inspiring artificial systems, it seems likely that they would create new frontiers, rather than to fill in known gaps. Since one possible outcome of these studies would be to find unexpected interpretations, the set of possible impacts this might have on algorithm design is large, diverse, and not easy to summarise.

Real world applications for the types of systems discussed in this survey include situations where it is difficult or impossible to specify at programming time what should be done. This is a very diverse set of scenarios, ranging from things such as personalised music selection to a robot assisting during a surgery. Instead of trying to describe the entire set of possible applications, these two will serve as examples. In both those cases a learner could use a range of information sources, including demonstrations (manually selecting a song/teleoperating the surgical robot) and different types of feedback, and in both cases it is not obvious how to pre program a fixed numerical success measure that the robot can extract from sensor readings. Interpretation of human behaviour is not necessarily straightforward, negative feedback in a music selection app might for example mean that a specific song is the wrong genre for the specific situation, or is worse than the song the teacher assumed would follow the previous song, or worse than the song the teacher would have chosen, or might be a too abrupt change in mood compared to the preceding song, etc. Interpreting silence could be even more problematic as the teacher might be silent due to being pleased, or due to giving up, or due to being busy, etc.

A surgeon giving feedback to a learner could provide a numerical value after a surgery, but one would need to think about what this represents; an evaluation of the post-op condition of the patient?, an evaluation of the success of the surgery?, an evaluation of how helpful the robot was during the surgery?, an evaluation of how responsive the robot was to specific commands during the surgery?, etc. Some surgeons might generate this number quickly and without much thought regarding what is meant or how it will be used, while other surgeons might very carefully read a manual, try hard to understand how the system will use the number, and carefully try to estimate a set of different criteria. Let's take an example where a human and a robot is jointly performing a surgery, and the robot has certain advantages, such as speed and precision, but where a human is better at determining what to do, and when. The robot observes a complication before the human does, and starts addressing it before the human is aware of it, and in the process ignores a series of verbal commands. While it is happening, the episode is disconcerting to the human surgeon. But when reviewing what happened after the surgery is completed, the decision to address the complication immediately, without waiting for the human to figure out what is going on, seems reasonable. It seems likely that interpretation of information sources such as facial expressions, speech commands and tone of voice during the surgery, a numerical value provided shortly after the surgery, a different numerical value provided after the surgeon understands what happened, etc, will be a challenging issue. If robot surgeons become common, and information is pooled, data might exist from a large number of different situations. Data from the messy, unexpected, and complicated situations might be important, and manually throwing away "noisy data" might not be a good approach. Perfect interpretations are neither possible, nor necessary in order to outperform naive interpretations (just as perfect robotic surgeons are neither possible, nor necessary in order to outperform human surgeons).

When dealing with unstructured real world environments, the project of finding better interpretations of actual human teachers is not one that can be "completed" any more than it is possible to "complete" the project of finding better world models. But making incremental progress is certainly possible, and hopefully this survey have described the various strands of research currently contributing to this progress. Making progress towards better interpretations is important for the same reasons that it is important to improve the world model of any autonomous learning system (not because of some hope that it could eventually lead to a perfect model of the real world, but because better models tends to lead to better behavior).

If the problem at hand was to find a complete and flawless interpretation of all types of information sources (from speech comments, to facial expressions and eye gaze, to EEG readings), then the vast array of possible information sources, and the diversity of possible meanings, would be a problem. If we instead only want to find some subset of information sources for which it is possible to find robust, reliable, and useful foundational interpretations, then the size of the search space is an advantage. This is true for both researchers building static interpretations, as well as for learners autonomously building interpretations of individual teachers. If a given type of social signal is used differently by different people, designers of static interpretations can use another social signal. If some specific human uses a social signal in a way that is not well approximated by any model that a learner can find, then the learner can switch to trying to find an interpretation for some other type of social signal, or even switch to some other type of interaction. In the same way, the fact that there are so many different strands of research that are relevant to the problem might make it hard to get a grip on the field. But it also means that there is a wealth of inspiration sources and avenues of attack. The size of the problem and the diversity of relevant research efforts can thus also be seen as an advantage to researchers.

References

[1] Charles L. Isbell, Christian Shelton, Michael Kearns, Satinder Singh, and Peter Stone. A social reinforcement learning agent. In *Proceedings of the Fifth International Conference on Autonomous Agents*, 2001.

[2] Charles L. Isbell, Michael Kearns, Satinder Singh, Christian Shelton, Peter Stone, and Dave Kormann. Cobot in LambdaMOO: An adaptive social statistics agent. *Autonomous Agents and Multiagent Systems*, 13(3), November 2006.

[3] Andrea L. Thomaz, Guy Hoffman., and Cynthia Breazeal. Reinforcement learning with human teachers: Understanding how people want to teach robots. *Proceedings of the 15th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2006.

[4] M Ho, M Littman, F. Cushman, and J Austerweil. Teaching with rewards and punishments: Reinforcement or communication? In *Proceedings of the 36th Annual Meeting of the Cognitive Science Society*, 2015.

[5] Shane Griffith, Kaushik Subramanian, Jonathan Scholz, Charles L. Isbell, and Andrea L. Thomaz. Policy shaping: Integrating human feedback with reinforcement learning. In *Proceedings of the International Conference on Neural Information Processing Systems (NIPS)*, 2013.

[6] T. Cederborg, I. Grover, C. L. Isbell, and A. L. Thomaz. Policy shaping with human teachers. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[7] Mark K. Ho, James MacGlashan, Michael L. Littman, and Fiery Cushman. Social is special: A normative framework for teaching with and learning from evaluative feedback. In *Cognition*, 2017.

[8] C. Nehaniv and K. Dautenhahn. Of hummingbirds and helicopters: An algebraic framework for interdisciplinary studies of imitation and its applications. *Interdisciplinary approaches to robot learning.,World Scientific Press*, 24:136–161, 2000.

[9] M. Stolle and C.G. Atkeson. Knowledge transfer using local features. *Proceedings of the IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*, 2007.

[10] Pieter Abbeel., Adam Coates., and Andrew Y. Ng. Autonomous helicopter aerobatics through apprenticeship learning. *International Journal of Robotics Research*, 29(13):1608–1639, 2010.

[11] B.D. Argall, S. Chernova, M. Veloso, and B. Brett. A survey of robot learning from demonstration. *Robot. Auton. Syst*, 57(5):469–483, 2009.

[12] Chrystopher L. Nehaniv. Nine billion correspondence problems. In *C. L. Nehaniv and K. Dautenhahn (Eds.), Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions. Cambridge University Press*, 2007.

[13] Manuel Lopes, Francisco S. Melo, Ben Kenward, and Jose Santos-Victor. A computational model of social-learning mechanisms. *Adaptive Behaviour*, 2009.

[14] K. Dautenhahn and C. L. Nehaniv. The agent-based perspective on imitation. In *Imitation in animals and artifacts. MIT Press*, pages 1–40, 2002.

[15] S.M. Nguyen and P-Y. Oudeyer. Socially guided intrinsic motivation for robot learning of motor skills. *Autonomous Robots, Springer*, 36(3):273–294, 2014.

[16] E. A. Billing and T. Hellstrom. A formalism for learning from demonstration. *PALADYN Journal of Behavioral Robotics*, 1(1):1–13, 2010.

[17] Thomas Cederborg and Pierre-Yves Oudeyer. A social learning formalism for learners trying to figure out what a teacher wants them to do. *PALADYN Journal of Behavioral Robotics*, pages 64–99, 2014.

[18] Pierre Bessiere, Christian Laugier, and Roland Siegwart. Probabilistic reasoning and decision making in sensory motor systems. *Springer Tracts in Advanced Robotics*, 2008.

[19] Joao Filipe Ferreira and Jorge Dias. Probabilistic approaches to robotic perception. *Springer Tracts in Advanced Robotics*, 2014.

[20] A.Y. Ng and S. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pages 663–670, 2000.

[21] P Abbeel and A Ng. Apprenticeship learning via inverse reinforcement learning. *ICML*, 2004.

[22] G. Neu and C. Szepesvári. Training parsers by inverse reinforcement learning. *Machine learning*, 77(2):303–337, 2009.

[23] C. Breazeal, A. Brooks, J. Gray, G. Hoffman, J. Lieberman, H. Lee, A.L. Thomaz, and D. Mulanda. Tutelage and collaboration for humanoid robots. *International Journal of Humanoid Robotics*, 1(2), 2004.

[24] A.-L. Vollmer, M. Muhlig, J. J. Steil, K. Pitsch, J. Fritsch, K. J. Rohlfing, and B. Wrede. Robots show us how to teach them: Feedback from robots shapes tutoring behavior during action learning. *PloS one*, 9(3):39–58, 2014.

[25] Anna-Lisa Vollmer, Jonathan Grizou, Manuel Lopes, Katharina Rohlfing, and Pierre-Yves Oudeyer. Studying the co-construction of interaction protocols in collaborative tasks with humans. In *The Fourth Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics*, 2014.

[26] Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell. Cooperative inverse reinforcement learning. *Neural Information Processing Systems (NIPS)*, 2016.

[27] Satinder Singh, Richard L. Lewis, and Andrew G. Barto. Where do rewards come from? *Proceedings of the Annual Conference of the Cognitive Science Society (CogSci)*, 2009.

[28] Satinder Singh, Richard L. Lewis, Andrew G. Barto, and Jonathan Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *Transactions on Autonomous Mental Development*, 2(2), 2010.

[29] Richard L. Lewis, Andrew Howes, and Satinder Singh. Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 2014.

[30] Andrew Howes, Richard L. Lewis, and Satinder Singh. Utility maximization and bounds on human information processing. *Topics in Cognitive Science*, 2014.

[31] Nick Bostrom. Superintelligence: Paths, dangers, strategies. In *Oxford University Press*, 2014.

[32] Smitha Milli, Dylan Hadfield-Menell, Anca D. Dragan, and Stuart Russell. Should robots be obedient? *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[33] Kaj Sotala. Concept learning for safe autonomous ai. In *AAAI Ethics and Artificial Intelligence Workshop*, 2015.

[34] A.Vinciarelli, M.Pantic, D.Heylen, C.Pelachaud, I.Poggi, F.D'Errico, and M.Schroeder. Bridging the gap between social animal and unsocial machine: A survey of social signal processing. *IEEE Transactions on Affective Computing*, 3(1):69–87, 2012.

[35] A.Vinciarelli and M.Pantic. Social signal processing. In *R.Calvo and S.D?Mello and J.Gratch and A.Kappas (eds.), Oxford Handbook of Affective Computing. Oxford University Press*, 2013.

[36] Z. Zeng, M. Pantic, G. Roisman, and T. Huang. A survey of affect recognition methods: Audio, visual and spontaneous expressions. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 31(1):39–58, 2009.

[37] Hatice Gunes and Bjorn Schuller. Categorical and dimensional affect analysis in continuous input: Current trends and future directions. *Image and Vision Computing*, 31(2):120–136, 2013.

[38] Evangelos Sariyanidi, Hatice Gunes, and Andrea Cavallaro. Automatic analysis of facial affect: A survey of registration, representation, and recognition. *IEEE Transactions on Pattern Recognitio and Machine Intelligence*, 37(6):1113–1133, 2015.

[39] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic. Gaussian process domain experts for model adaptation in facial behavior analysis. In *International Conference on Computer Vision and Pattern Recognition (CVPRW16). 4th Workshop on Context Based Affect Recognition*, 2016.

[40] A.L. Thomaz, M. Berlin., and C. Breazeal. Robot science meets social science: An embodied computational model of social referencing. *Workshop toward social mechanisms of android science (CogSci)*, pages 7–17, 2005.

[41] Jonathan Grizou, Iñaki Iturrate, Luis Montesano, Pierre-Yves Oudeyer, and Manuel Lopes. Interactive learning from unlabeled instructions. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence*, 2014.

[42] S Ehrlich, A Wykowska, K Ramirez-Amaro, and G Cheng. When to engage in interaction, and how: Eeg-based enhancement of robot's ability to sense social signals. In *International Conference on Humanoid Robots*, pages 1104–1109, 2014.

[43] Mark O. Riedl and Brent Harrison. Using stories to teach human values to artificial agents. In *Proceedings of the 2nd International Workshop on AI, Ethics and Society*, July 2016.

[44] M Tomasello. A natural history of human thinking. In *Harvard University Press*, 2014.

[45] Samantha Krening, Brent Harrison, Karen Feigh, Charles Isbell, Mark Riedl, and Andrea Thomaz. Learning from explanations using sentiment and advice in rl. *IEEE Transactions on Cognitive and Developmental Systems*, 2016.

[46] Dylan Hadfield-Menell, Anca D. Dragan, Pieter Abbeel, and Stuart Russell. The off-switch game. *International Joint Conference on Artificial Intelligence (IJCAI)*, 2017.

[47] Owain Evans, Andreas Stuhlmüller, and Noah D. Goodman. Learning the preferences of ignorant, inconsistent agents. In *AAAI*, 2016.

[48] Cynthia Breazeal, Jesse Gray, and Matt Berlin. An embodied cognition approach to mindreading skills for socially intelligent robots. *I. J. Robotic Res*, 28:656–680, 2009.

[49] Thomas Cederborg and Piere-Yves Oudeyer. From language to motor gavagai: Unified imitation learning of multiple linguistic and non-linguistic sensorimotor skills. *IEEE Transactions on Autonomous Mental Development*, 2013.

[50] K Ramirez-Amaro, M Beetz, and G Cheng. Transferring skills to humanoid robots by extracting semantic representations from observations of human activities. *Artificial Intelligence*, 2015.

[51] Sonia Chernova and Andrea L. Thomaz. Robot learning from human teachers. In *Synthesis Lectures on Artificial Intelligence and Machine Learning, Morgan Claypool Publishers*, 2014.

[52] Manuel Lopes, Thomas Cederborg, and Pierre-Yves Oudeyer. Simultaneous acquisition of task and feedback models. In *International Conference on Development and Learning (ICDL)*, 2011.

[53] Grizou J, Lopes M, and Oudeyer P-Y. Robot learning simultaneously a task and how to interpret human instructions. In *Joint IEEE International Conference on Development and Learning and on Epigenetic Robotics (ICDL-EpiRob)*, 2013.

[54] Jonathan Grizou, Iñaki Iturrate, Luis Montesano, Pierre-Yves Oudeyer, and Manuel Lopes. Calibration-free bci based control. In *AAAI Conference on Artificial Intelligence*, 2014.

[55] Jonathan Grizou. Learning from unlabeled interaction frames. *Ph.D Thesis*, 2014.

[56] I. Iturrate, J. Grizou, J. Omedes, P-Y. Oudeyer, M. Lopes, and L. Montesano. Exploiting task constraints for self-calibrated brain-machine interface control using error-related potentials. *Plos One*, 2015.

[57] Robert Loftin, Bei Peng, James MacGlashan, Michael L. Littman, Matthew E. Taylor, Jeff Huang, and David L. Roberts. Learning something from nothing: Leveraging implicit human feedback strategies. In *Proceedings of the Twenty-Third IEEE International Symposium on Robot and Human Communication (ROMAN)*, 2014.

[58] Robert Loftin, Bei Peng, James MacGlashan, Michael L. Littman, Matthew E. Taylor, Jeff Huang, and David L. Roberts. Learning behaviors via human-delivered discrete feedback: modeling implicit feedback strategies to speed up learning. In *Journal of autonomous agents and multi-agent systems*, 2015.

[59] Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. Learning to interpret natural language commands through human-robot dialog. In *International Joint Conference on Artificial Intelligence (IJCAI)*, 2015.

[60] P.M. Pilarski, M.R. Dawson, T. Degris, F. Fahimi, J.P. Carey, and R.S. Sutton. Online human training of a myoelectric prosthesis controller via actor-critic reinforcement learning. In *IEEE International Conference on Rehabilitation Robotics (ICORR)*, 2011.

[61] K. W. Mathewson and P. M. Pilarski. Simultaneous control and human feedback in the training of a robotic agent with actor-critic reinforcement learning. In *IJCAI Workshop on Interactive Machine Learning*, July 2016.

[62] Andrea L. Thomaz and Cynthia Breazeal. Teachable robots: Understanding human teaching behavior to build more effective robot learners. *Artificial Intelligence Journal*, pages 716–737, 2008.

[63] Maya Cakmak and Manuel Lopes. Algorithmic and human teaching of sequential decision tasks. In *AAAI Conference on Artificial Intelligence*, 2012.

[64] V. Veeriah, P. M. Pilarski, and R. S. Sutton. Face valuing: Training user interfaces with facial expressions and reinforcement learning. In *IJCAI Workshop on Interactive Machine Learning*, July 2016.

[65] M. Cakmak and A.L. Thomaz. Designing robot learners that ask good questions. In *International Conference on Human-Robot Interaction (HRI)*, 2012.

[66] Kaushik Subramanian, Charles L. Isbell, and Andrea L. Thomaz. Exploration from demonstration for interactive reinforcement learning. In *Proceedings of the 15th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2016.

[67] K. W. Strabala, M. K. Lee, A. D. Dragan, J. L. Forlizzi, S. Srinivasa, M. Cakmak, and V. Micelli. Towards seamless human-robot handovers. *Journal of Human-Robot Interaction*, 2(1):112–132, 2013.

[68] Anca D. Dragan and Siddhartha Srinivasa. Generating legible motion. *Robotics: Science and Systems*, 2013.

[69] Stefanos Nikolaidis, Anca D. Dragan, and Siddhartha Srinivasa. Viewpoint-based legibility optimization. *Robotics: Science and Systems (RSS)*, 2016.

[70] Freek Stulp, Jonathan Grizou, Baptiste Busch, and Manuel Lopes. Facilitating intention prediction for humans by optimizing robot motions. In *International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[71] P Lanillos, J. F Ferreira, and J Dias. Designing an artificial attention system for social robots. In *International Conference on Intelligent Robots and Systems (IROS)*, 2015.

[72] Joao Filipe Ferreira and Jorge Dias. Attentional mechanisms for socially interactive robots, a survey,. *Transactions on Autonomous Mental Development*, 6(2):110–125, 2014.

[73] Bei Peng, James MacGlashan, Robert Loftin, Michael L. Littman, David L. Roberts, and Matthew E. Taylor. A need for speed: Adapting agent action speed to improve task learning from non-expert humans. In *Proceedings of the International Conference on Autonomous Agents and Multiagent Systems*, 2016.

[74] Dorsa Sadigh, Shankar Sastry, Sanjit A. Seshia, and Anca D. Dragan. Information gathering actions over human internal state. *International Conference on Intelligent Robots and Systems (IROS)*, 2016.

[75] Irene M. Pepperberg and Diane V. Sherman. Training behavior by imitation: from parrots to people ... to robots? In *C. L. Nehaniv and K. Dautenhahn (Eds.), Imitation and Social Learning in Robots, Humans and Animals: Behavioural, Social and Communicative Dimensions. Cambridge University Press*, pages 383–406, 2007.

[76] Andrea L. Thomaz and Cynthia Breazeal. Asymmetric interpretations of positive and negative human feedback for a social learning agent. In *Proceedings of the 16th IEEE International Symposium on Robot and Human Interactive Communication (RO-MAN)*, 2007.

[77] Andrea L. Thomaz and Cynthia Breazeal. Reinforcement learning with human teachers: Evidence of feedback and guidance with implications for learning performance. *AAAI*, 2006.

[78] Andrea L. Thomaz and Cynthia Breazeal. Experiments in socially guided exploration: Lessons learned in building robots that learn with and without human teachers. *Connection Science, Special Issue on Social Learning in Embodied Agents*, pages 91–110, 2008.

[79] Wykowska A, Chaminade T, and Cheng G. Embodied artificial agents for understanding human social cognition. *Phil. Trans. R. Soc. B*, 2016.

[80] M. Cakmak, C. Chao, and A.L. Thomaz. Designing interactions for robot active learners. *IEEE Transactions on Autonomous Mental Development*, 2(3):108–118, 2010.

[81] B. Akgun, M. Cakmak, J. Yoo, and A.L. Thomaz. Trajectories and keyframes for kinesthetic teaching: A human-robot interaction perspective. *International Conference on Human-Robot Interaction*, 2012.

[82] W. Bradley Knox, Brian D. Glass, Bradley C. Love, W. Todd Maddox, and Peter Stone. How humans teach agents: A new experimental perspective. In *International Journal of Social Robotics*, 2012.

[83] W. Bradley Knox, Cynthia Breazeal, and Peter Stone. Learning from feedback on actions past and intended. In *Proceedings of 7th ACM/IEEE International Conference on Human-Robot Interaction (HRI)*, March 2012.

[84] W. Bradley Knox and Peter Stone. Interactively shaping agents via human reinforcement: The tamer framework. In *The 5th International Conference on Knowledge Capture*, 2009.

[85] W. Bradley Knox, Cynthia Breazeal, and Peter Stone. Training a robot via human feedback: A case study. In *Proceedings of the International Conference on Social Robotics (ICSR)*, 2013.

[86] James MacGlashan, Mark K. Ho, Robert Loftin, Bei Peng, Guan Wang, David L. Roberts, Matthew E. Taylor, and Michael L. Littman. Interactive learning from policy-dependent human feedback. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, 2017.

[87] Chandrayee Basu, Qian Yang, David Hungerman, Mukesh Singhal, and Anca D. Dragan. Do you want your autonomous car to drive like you? *International Conference on Human-Robot Interaction (HRI)*, 2017.

[88] Dorsa Sadigh, Anca D. Dragan, Shankar Sastry, and Sanjit A. Seshia. Active preference-based learning of reward functions. *Robotics: Science and Systems (RSS)*, 2017.

[89] B. F. Skinner. The behavior of organisms: An experimental analysis. In *New York: Appleton-Century*, 1938.

[90] B. F. Skinner. Science and human behavior. In *New York: Macmillan*, 1953.

[91] M Gilbert. On social facts. In *London: Routledge*, 1989.

[92] M Bratman. Shared cooperative activity. *Philosophical Review*, 101(2):327–340, 1992.

[93] J Searle. The construction of social reality. In *New York: Free Press*, 1995.

[94] R Tuomela. The philosophy of sociality: The shared point of view. In *Oxford: Oxford University Press*, 2007.

[95] M Bratman. Shared agency: A planning theory of acting together. In *Oxford University Press*, 2014.

[96] M Gilbert. Joint commitment: How we make the social world. In *Oxford University Press*, 2014.

[97] M Tomasello. Why we cooperate. In *MIT press*, 2009.

[98] M Tomasello. Origins of human communication. In *MIT press*, 2008.

[99] M Tomasello. A natural history of human morality. In *Harvard University Press*, 2016.

[100] R Tuomela. The philosophy of sociality: The shared point of view. In *Oxford University Press*, 2007.

[101] H. P. Grice. Logic and conversation. In *Cole, P. and Morgan, J. (eds.) Syntax and semantics. New York: Academic Press*, 1975.

[102] J Call, B Hare, M Carpenter, and M Tomasello. Unwilling versus unable: chimpanzees' understanding of human intentional action. *Developmental Science*, 7(4):488–498, 2004.

[103] M Schmelz, J Call, and M Tomasello. Chimpanzees know that others make inferences. In *Proceedings of the National Academy of Sciences*, 2011.

[104] J. Call and M Tomasello. Does the chimpanzee have a theory of mind? 30 years later. *Trends in Cognitive Science*, 2008.

[105] M Tomasello. Why don't apes point? In *N. Enfield and S.C. Levinson (Eds.), Roots of human sociality: Culture, cognition and interaction*, pages 506–524, 2006.

[106] W. V. O. Quine. Word and object. In *MIT press*, 1960.

[107] M Tomasello, M Carpenter, J Call, T Behne, and H Moll. Understanding and sharing intentions: The origins of cultural cognition. *Behavioral and Brain Sciences*, 2005.

[108] Angelo Cangelosi, Giorgio Metta, Gerhard Sagerer, Stefano Nolfi, Chrystopher Nehaniv, Kerstin Fischer, Jun Tani, Tony Belpaeme, Giulio Sandini, Luciano Fadiga, Britta Wrede, Katharina Rohlfing, Elio Tuci, Kerstin Dautenhahn, Joe Saunders, and Arne Zeschel. Integration of action and language knowledge: A roadmap for developmental robotics. *IEEE Transactions on Autonomous Mental Development*, 2010.

[109] G. Rizzolatti and M. A Arbib. Language within our grasp. *Trends in Neurosciences*, 21(5):188–194, 1998.