Overview of current AI Alignment Approaches

Micah Carroll

11th December 2018

Abstract. This literature survey proposes a categorization for AI alignment approaches within the expected utility maximization framework. We additionally compare such methods with respect to their *training signal*, *assumptions about the human*, *algorithm output*, and *scalability*. Finally, we summarize how all such approaches can be cast as instantiations of the CIRL (*Cooperative Inverse Reinforcement Learning*) framework.

1 Introduction

Surveying the current AI Safety literature, there doesn't seem to be a unique accepted definition of "AI alignment" [1]. One of the most simple definitions that has gained a decent amount of traction is that of Paul Christiano [2]:

> A is **aligned** with H if *A is trying to do what H wants it to do.*

Taken literally, this definition is pretty general. In fact, one could interpret the entire field of AI as trying to get agents do what humans want them to do. All AI algorithms can be thought of as being *designed for alignment* in this sense. In its original context however, the expression "AI alignment" was referred mainly to the alignment of *highly capable AIs*. In that sense, we can draw a (somewhat arbitrary) distinction between AI approaches that can preserve value alignment when approaching tasks with broad-ranging implications, and those that are mainly applicable when solving narrow tasks (e.g. sorting objects in an assembly line).

Currently there is no algorithm or approach that is capable of addressing the AI alignment problem in full generality, neither in theory or in practice. "Solving" Value Alignment for any arbitrary task appears to be an impossibly hard problem – as it would require fully characterizing the values of humans, and in the multi-agent setting, aggregating and trading off between different humans, all the while taking into account value changes over time. However, various directions are currently being explored by the research community to achieve approximate solutions (approximate as far as you believe that people's values cannot be exactly quantified). The purpose of this report is to present an overview of such directions and the preliminary results and properties that various methods in the field have, with a focus on highlighting how such methods can fall short of realistic alignment in real world settings. One should not take the methods considered as a comprehensive view of the field.

Most current approaches to the alignment problem are based on AI systems that are *expected utility maximizers* and that receive training signal from human data. These assumptions are already restrictive, and some have argued that they pose an unsurmountable challenge to achieving full alignment [18]. In addition to these assumptions, phenomena such as bounded rationality - the idea that rationality of individuals is limited by their information state, the cognitive limitations of their minds, and time constraints – are not properly accounted for in the current frameworks. Even though there have been calls to reframe the problem in other lenses [3], in this report we will limit ourselves to work done within these assumptions.

When assessing the various methods, this report will mainly focus on what is referred to as the *specification problem* [18,17], that can be defined as the problem of "*defining* the behavior that

we would want to see from our AI system". Moreover, we will briefly mention the approaches' *scalable supervision*, as defined here [19]. However, we will not discuss AI Safety issues that can occur during the learning process, that are more tangential to the alignment problem and that affect all methods considered at a similar level.

2 Categorizing Alignment Approaches

When assessing current approaches to alignment, various factors emerge as playing a differentiating role between methods. Some of these factors, and how they can affect AI alignment are:

- **Training signal:** In order to train an agent to do anything, we need to have some kind of *training signal* from which the agent can infer what kind of behaviour is appropriate. In the context of classic reinforcement learning, such a training signal is given by the reward, that implicitly defines which behaviours are "good".

  However, for complex contexts, it is hard for humans to explicitly write out a reward function that encompasses everything they care about. As a classic example, maximizing the amount of dirt collected might seem like a reasonable objective to give to a cleaning robot. The problems with such objective only appear obvious in hindsight once the robot learns to continuously dump and re-collect the dirt it found. The difficulty of specifying a proxy reward is equivalent to the problem widely known as Goodhart's Law [20,21].

  There are however other ways for humans to provide training signal that try to avoid making the human explicitly specify a reward, such as *providing near-optimal demonstrations*, *providing preferences* between states or trajectories, or *decomposing problems into subproblems and assembling solutions to subproblems to form the solution of a harder problem* (examples of these are considered in section 3 and 4). In this sense, we can think of the training signal broadly as being the type of interaction between the human and the agent, together with the assumptions about what the meaning of the interaction is towards the larger goal.

  Humans cannot provide training signal of any type for any task. In fact, the type of training signal is an assumption about the approach that greatly determines *feasibility* of obtaining such signal and *tractability* when scaling up the algorithm to more complex domains. For example, humans can't necessarily perform "near-optimal demonstrations" for controlling a robot with many degrees of freedom (for example, 10 arms), or for running an economy (we have trouble with that ourselves). However, humans might still be able to state preferences between trajectories for the movement of such robot, or between trajectories of economies. Another factor to consider is the *cost* of the training signal: if it requires online interaction, the signal can become very time intensive for humans and greatly slow down the training process, potentially impacting the feasibility of the approach at larger scales. In this sense, the type of training signal directly influences the assumptions about the human and the scalability of the approach.

- **Assumptions about the human:** In addition to assuming that the human is capable of providing the training signal, any method is implicitly makes some assumptions about how the human chooses their actions, about their information state, how they manage computation-time tradeoffs, their biases, their action space, and so on. Most methods in fact model the human at some stage as being near-optimal (e.g. Boltzmann rational) in their demonstrations, or try to explicitly model human suboptimalities. However, that of modelling human suboptimalities - either by manually specifying them or learning them -

seems to be a very difficult problem, as there is no way to check whether the model is correct [4].

A nice solution to this problem would be to learn both the planner (i.e. the human model) *and* the reward function that humans optimize concurrently. However, not only it is impossible to learn both planner and reward function from demonstrations without any further assumptions, but it is always possible to find degenerate pairs of planner and reward that perfectly explain the data [32].

Generally, it seems like making some assumptions about human behaviour is going to be necessary to tackle alignment, and due to the issues above all methods are going to incur in some model-misspecification, at least to the degree that the assumptions about the human are too strong. However, an important thing to note is that for different methods such assumptions are more or less fundamental for the correct behaviour of the algorithm, thus having different consequences in terms of alignment.

- **Algorithm Output:** One major difference between approaches is the type of output.

  One basic approach to alignment is to train the AI to act how we want it to without constraining it to explicitly represent "what we value" at any point of the process. We can think of this as a policy-learning approach that preserves alignment. The standard proposal in this direction is to use some form of *Imitation Learning* - incentivizing the agent to *mimic* behaviour that we consider aligned. In this survey, we will refer to such methods as *Imitation Learning based approaches*. IDA and Debate [13,14] - seen later in more detail, can be considered examples of this.

  Another approach is based on trying to obtain an explicit representation of the reward function, and subsequently training an agent with standard Reinforcement Learning techniques using such a reward function. Given the difficulties and risks involved with specifying a reward function by hand when dealing with a highly-capable AI, [22,23] the standard approach in this case is to try and *learn* such a reward, in a process commonly called *Reward* or *Value Learning* [18].

  Both these approaches - Imitation-Based and Value Learning methods - can be thought of as different angles from which to tackle the specification problem [19,17].

  One reason behind choosing to learn a reward function rather than a policy directly is that it is generally assumed that reward functions are able to more succinctly encapsulate everything that is necessary to know about the goal to be able to generalize to new environments. Moreover, one clear advantage of Value Learning over Imitation Learning is that once a reward function is recovered – assuming it is (approximately) correct – it is relatively straightforward to achieve better-than-human performance by just applying standard RL techniques. On the other hand, using Imitation Learning techniques to exceed human performance requires some additional assumptions and ingenuity (e.g. as seen for IDA & Debate later). Other advantages of Value Learning are possibly better generalization and interpretability. If one recovers the "true" reward (or at least something in the spirit of it), one would expect it to better generalize to new situations unseen at training time. Also, having an explicit representation of the reward enables for greater interpretability compared to a black-box policy. However, one disadvantage of Reward Learning is that even small mistakes in the learned reward can be amplified in unpredictable ways by the optimization process, leading to catastrophic failures. This is commonly known as the problem of negative side effects [19].

- **Scalability and Complexity:** In addition to being affected by the type of training signal - and thus by the extent of user interaction necessary - scalability strongly depends on the mathematical framing of the problem and its computational complexity.

We first will consider separately approaches that are based off of value learning and imitation learning respectively, and then place all the methods considered in the CIRL framework (Cooperative Inverse Reinforcement Learning, [8]).

3 Value Learning approaches

3.1 Inverse Reinforcement Learning

Inverse Reinforcement Learning (IRL, [24,11]) is based on trying to infer the human's reward function by observing the human's behaviour, assuming that the human is near-optimal at maximizing their reward.

Assumptions about H. The standard formulations of IRL ([24,11] and others) assume that the human is near-optimal (Boltzmann rational) in giving the demonstrations used as training data. Some attempts have been made to relax some assumption to allow for models of human suboptimality [25]. However, scaling the quality of such models arbitrarily seems like a difficult problem due the reasons mentioned in Section 2. Even with these assumptions about the human, in most scenarios the reward function is not uniquely recoverable. Nevertheless, various ways have been proposed to narrow down to a specific reward function, and seem to perform reasonably well in practice in simple environments [11,37].

Training signal. IRL requires the human to give expert demonstrations for the task at hand. As mentioned in Section 2, this is implicitly assuming that the human *can* give such demonstrations. This might not be possible for instance in tasks with a different action space, or for which humans cannot meet the assumptions made by the human model. It is unclear what the consequences of this are for achieving alignment: one thing to note is that even though humans cannot perform near-optimal demonstrations for some tasks, assuming that human demonstrations in other tasks are sufficient to infer human reward, other non-demonstrable complex behaviours could emerge spontaneously as instrumental goals.

3.2 Inverse Reward Design

Inverse Reward Design (IRD, [6]) tries to tackle the problem of *negative side effects* [19] that arise when the reward function is misspecified. In particular, IRD circumvents the requirement of manually specifying *all-encompassing* reward functions, requiring the human to only specify a *proxy reward function* that is near-optimal in a specific training context. On a high-level, IRD is set up so that the agent is aware that the reward function specified by the human is just a proxy designed with certain test environments in mind, and thus might be misspecified for new environments. By making the AI agent interpret the proxy reward it is given in a pragmatic (i.e. non-literal) way, such agent maintains some uncertainty over what the *true reward function* is, and chooses its actions according to risk-averse planning on its posterior over the true reward.

Assumptions about H. IRD assumes that the human is Boltzmann rational at specifying reward functions for a specific set of training environments. Implicitly, this assumes that the human is capable of specifying a reward function for any of these environments in the first place. It seems like this might not be possible for sufficiently complex environments, for which even coming up with a near-optimal rewards in a specific test environment might be hard, without access to a simulator (e.g. running an economy).

Training signal. In IRD the only interaction with the human is in the form of the proxy reward function it receives at the beginning of training. Extensions on IRD added the option to receive active feedback from the user [7], enabling further disambiguation of the true reward by the human.

Scalability. Computationally, this approach seems challenging, and as such is currently restricted to simple gridworlds. Theoretically, this approach seems only as scalable as the ability for humans to specify any meaningful reward function is.

3.3 Learning from human preferences

Another approach that has been proposed is to learn the desired reward function by preference elicitation.

Training signal. In Deep Reinforcement Learning From Human Preferences [5], humans are shown snippets of trajectories and asked to provide feedback by expressing a preference of one over the other. The snippets shown are chosen to maximize the information gained by the elicitation. Standard RL techniques are applied to this estimate of the reward function to train agents at task to achieve better-than-human performance.

Assumptions about H. This approach assumes that the human is near-optimal (Boltzmann rational) at providing preferences over trajectories. This implicitly requires the trajectories to be informative enough about the true reward (long enough to encompass long-term goals and interpretable enough by humans), and humans to be sufficiently rational in evaluating them. Because of this it is unclear how much this technique can be scaled to achieve alignment for complex tasks in which there are many subtle nuances between trajectories and rewards are on long horizons.

Scalability. The main bottleneck for this method in terms of time complexity is definitely due to the amount of human interaction required. Some attempts have been made to speed up this process [34] but it's unclear how scalable these methods are to complex domains. Moreover, it is unclear if can recover reward functions uniquely, and how the theoretical guarantees compare to those of IRL in learning reward functions.

4 Imitation Learning Based Methods

4.1 Imitation Learning

Some authors argue that approaches [36] that cannot surpass human performance are not viable solutions to the alignment problem, as economic incentives would cause people to turn to unsafe better-than-human performance methods. For this reason, vanilla imitation learning by itself is usually not considered a realistic "solution" to the alignment problem when considered in full generality. However, as it is used by other approaches and is broadly relevant to alignment we will cover it here.

Imitation learning treats the human actions as an example of aligned actions, and directly learns a policy that mimics the policy of the human. There are various imitation learning algorithms [28,29], but we will only consider the method on a high level.

Assumptions about H. As mentioned above, imitation learning assumes that human actions are examples of aligned actions. This assumption could fall short of reality due to human suboptimality, inconsistency, and asymmetry between humans and agents. As an example of the latter, we don't necessarily want to make our trained agent crave coffee in the morning just like the human it was trained from. Another assumption that is made by imitation learning methods is that humans are capable of giving demonstrations whatsoever for the task at hand. As we

have seen before however, this can be difficult. Especially in terms of alignment, gathering large amounts of "aligned" human behaviour to mimic seems to not be a trivial problem.

Performance. As mentioned before, imitation learning approaches are limited by design in not being able to surpass human performance. This is both a blessing and a curse, as it avoids many of the safety issues introduced by the maximization of a reward function, but also does not deliver on the promise of a better-than-human AI.

Scalability. One issue with imitation learning methods is that the learnt policies usually have poor generalization capabilities. Some methods [28] try to address this, but do not solve the distributional shift problems entirely.

4.2 IDA

Iterated Distillation and Amplification (IDA, [13]) constitutes a somewhat different approach to the alignment problem compared to the previous methods considered. The idea is based on the assumption that humans' *considered judgement* is as close as we can get to ground truth *aligned behaviour*, and that the interactions of the human with IDA are representative of such considered judgement [31]. This can be seen as an attempt to address the *ideal specification* problem (according to the categorization advanced by Deepmind [17]), of which IDA can be considered a corresponding *design specification*.

Together with Debate (seen later), IDA can be considered an attempt to overcome the major limitation of not being able to exceed human performance while still maintaining an imitation learning-based approach, thus avoiding some of the risks involved with the unrestricted optimization of a reward function.

Training signal. The idea underlying IDA is to first train an oracle X through imitation learning to be able to solve simple tasks by observing a human. Then have the human solve more difficult tasks by decomposing them into simpler subtasks (that it feeds to X to solve) and assembling the subtasks' solutions provided by X. By training X to imitate this process of decomposition and reassembly, and repeating the entire process arbitrarily many times, the idea is that one could achieve super-human performance while preserving alignment at each step of X's training. For more details about the algorithm, see the original paper [13]. In this sense, the act of decomposition of a problem into subtasks and assembly of subtask solutions is the training signal used to train X.

Assumptions about H. IDA assumes that the human is capable of being near-optimal in performing a complex task (in a way that we would consider *aligned*) if presented with solutions for all possible subtasks H could require. In other words, it is assuming that H is capable of "decomposing" a task and "assembling" solutions to its sub-tasks in an *aligned* way, and that such a process enables human actions to be reflective of their considered judgement. This is possibly a strong assumption, that needs more empirical investigation. However, it seems to assume human optimality in a more restricted way than many other approaches, and in that sense this direction looks promising.

Scalability. The main assumption that underlies the scalability of the method is that the process of training X to imitate the human preserves alignment, and that the human is capable of preserving alignment when scaling up to arbitrarily complex tasks, when given access to an oracle X that is capable of solving problems that are slightly easier than the task at hand in an aligned way. It is unclear how reliably one could get X to preserve alignment at each iteration of IDA. One necessary factor for IDA to scale is for imitation learning to be able to imitate human alignment in tasks robustly. This seems however like it could require an unfeasible amount of training data, especially once the level of tasks reaches a certain complexity.

4.3 Debate

The idea behind the debate alignment approach [14] is to have agents debate topics with one another when they disagree. If there is a disagreement, both agents will try to identify flaws in the other's argument, aiming to win the debate. A human judge will look at the agents arguments, and will pick a winning side. The main underlying assumption is that it is harder to lie than to refute a lie, so that there is no incentive to lie and the agents are driven to try and be as accurate as possible. As both agents in the debate share the same weights (it can be thought of as self-play), and thus have equal capability, and one would expect the agent arguing for the correct position to be able to win the debate.

Assumptions about H. One of the main assumptions is that the human will be capable of identifying which agent is telling the truth by evaluating their arguments. This seems difficult in settings in which the matter being debated is beyond human comprehension, although the paper addresses why one could expect this to be possible. If this assumption were to hold, this approach could lead to better-than-human performance.

Training signal. The human provides training signal to the debate agents by providing it's judgements. The agents are optimized to win the debate. As mentioned in the paper, Debate is related to IDA in that both involve a recursive decomposition of the task in which humans help answer relevant questions. The main difference is that in IDA each question is answered by a different copy of X, while in the Debate each question is answered by one of the two debaters. However, by making the agents commit to just one line of debate, Debate could be preferable in terms of complexity.

Scalability. There are various reasons to be concerned about the scalability of such an approach: in short, it is unclear if debates structured as in this approach would robustly converge to the right answer. This could be due to human biases, to the agent learning to exploiting the human. Moreover, the scalability of the approach relies on having imitation-learning based human models to generate training data in order to reduce the burden on actual humans. However, like in the case of IDA, such models are currently not yet realistic.

5 A unifying perspective: CIRL

As we have seen in the previous sections, alignment approaches generally require taking into account both the human - who determines what the objective is - and the agent, that infers such an objective under some assumptions, and then attempts to maximize it. In this sense, alignment is fundamentally a multi-agent problem. CIRL [8] is an attempt to formally define the alignment problem, formulating it as a two-player game in which a human (H) and a AI agent (R) share a common reward function, but only the human has knowledge of such a reward function. Solving a CIRL game however requires more than multi-agent decision theory, as one has to address the fact that humans don't easily fall in any idealized framework of rational behaviour [10].

Even though there have been attempts to solve the CIRL problem directly, some of the most immediate benefits provided by the CIRL framework are that it enables to collocate approaches to alignment within a unified picture. In particular, we can think of Value Learning methods to alignment as first trying to recover the human reward, and then maximize it, in what can be considered a 2-phase game. Imitation-Learning methods instead set their objective to maximize similarity between the agent and human actions.

All of the methods considered can be thought of in this CIRL framework, with more or less structured assumptions about the game and it's players:

- **IRL** makes the assumption that the cooperative game between H and R has two phases: one in which H displays near-optimal behavior (or with modeled suboptimalities [25]) and R learns from it, and a second phase in which R acts in accordance to such reward function. IRL has actually been proven to be the best response to a fixed demonstration-by-expert policy by H in the CIRL framework [8]. In this sense, R is not taking advantage of the fact that H wants to help R to learn the reward, and more radically, H is just acting as usual in the demonstration phase, without explicitly trying to help R learn the reward. Active IRL [27] tries to remedy the first problem.

- **IRD** can be thought of as a CIRL game in which H specifies a proxy reward function in a first phase of the game, communicating it to R together with a context M of the test environments for which the proxy was designed. This data is interpreted by R pragmatically, which then acts in the second phase of the game by risk-aversely maximizing what the reward based on it's belief about the true reward function is. Active IRD tries to take advantage of the fact that H is still willing to help R narrow down on the true reward function even after the first phase of the game is over [7].

- **Deep RL from Human Preferences** takes advantage of the sequential structure of the game: at each turn R asks H their preference between a pair of trajectories chosen to be maximally informative, and takes H's response into account to update its belief about the true reward function.

- **Imitation Learning** can also be cast as a solution to the CIRL problem. In a first phase H demonstrates the behaviour that it would want R to perform, and in a second phase R acts in the world trying to copy that behaviour, assuming that it will lead to high reward for H.

- **IDA** can be thought of as a game in which there are arbitrarily many instruction phases and then a deployment phase (or also interleaved learning and deployment phases). In each learning phase, H instructs R to imitate H on the current task level decomposition and composition (this consists of basically solving a nested imitation learning CIRL problem). At deployment phase, R answers questions that H poses to it (in the basic question-formulation of IDA).

- **Debate** can be thought of a CIRL game in which H interacts with R by expressing it's judgements, and thus giving R training signal to update it's belief about human values, thus becoming a better debater, and contributing to the H R pair in achieving high reward. The game is structured in phases composed of a round in which H asks a question, rounds in which R expresses it's arguments and counter-arguments, and a final round in which H expresses their judgement.

5.1 Solving CIRL

Some attempts have been made to try and solve the CIRL game directly - finding pairs of policies for H and R that maximize the expected reward obtained by the H and R pair. However, it currently seems necessary to introduce additional assumptions in order to make computation more tractable. In Pragmatic-Pedagogic Value Alignment [9], the authors assume that the human will behave pedagogically in a Boltzmann-rational way. This assumption was partially relaxed in a later paper [10], in which the human model was made to be a arbitrary function of human Q-values. Here we will consider the latter method. In both cases, CIRL captures that H has an incentive for R to infer the correct reward, inasmuch it helps to the maximization of the

reward for the H-R pair. This has nice properties as it creates an incentive to mitigate and avoid unintended consequences from misspecified rewards.

Training signal. the interaction of H with R, under the assumption of the method, is what constitutes the training signal for the agent. In the case of CIRL, the type of this interaction is not specified out of the box, and it will emerge naturally as part of the solution. In particular, under the assumptions in the paper [10], pedagogic behaviours tend to emerge spontaneously, to the extent that it's useful for the H-R pair to achieve high reward.

Assumptions about H. CIRL makes strong assumptions about the human, namely that it has full information, and acts in accordance to a policy parametrized by Q-values of the state. One implicit assumption here is that H is estimating R's belief state, and is using this to plan what actions to perform - taking into account the effect on R's belief about the true reward. These assumptions are particularly strong as it doesn't take into account human's bounded rationality, that seems to be a strong limiting factor in this setting.

Scalability. Even though CIRL seems to solve the problem in most generality out of all methods presented above, it's unclear how scalable of an approach solving the CIRL game itself in full generality is. Even with simplifying assumptions the CIRL problem has bad complexity properties, as it can be formulated as a POMDP that is exponential in the size of the action space. [10] Moreover, it's assumptions about the human are particularly strong.

6 Conclusion

Current progress. As mentioned in the introduction, none of the methods surveyed here can clearly solve the alignment problem at scale. However, this does not take away from their merit, or from the usefulness of further investigating their properties and pushing their capabilities. In fact, it seems at the very least plausible that such methods could be useful as subroutines in more advanced alignment approaches - in a similar way to how IDA and Debate use Imitation Learning and could also use IRL [36].

Not considering computational efficiency concerns, it seems like most of problems of the surveyed alignment approaches stem from their assumptions about humans. As mentioned previously, it seems impossible to totally forgo all such assumptions, but it is possible to try and be as robust as possible to such assumptions not holding.

For example, IDA and Debate make assumptions about human optimality that seem less fundamental than CIRL: CIRL assumes that the entire interaction between H and R - no matter what form it ends up taking - will be shaped according to the way the human acts (encoded in the human model), while IDA and Debate start from the assumption that careful human deliberation is as close as we can get to aligned behaviour and try to isolate such behaviour by a specific choice of setting (debates and task decomposition/composition). However, the practicality of Imitation Learning Based methods depends on having imitation learning capabilities that are far more advanced than the ones we have today.

Further research in existing methods. In general, it seems like further research is necessary in establishing where the assumptions underlying these approaches to alignment are reasonable, where they are not, and working towards trying to relax the strength of such assumptions while still maintaining computational tractability.

For value learning in particular, interesting directions seem to be that of satisficing and mild optimization, such as using quantilizers, which select an action randomly from a top quantile by expected utility [35]. This could help being more robust to model misspecification, to which value learning methods seem to be more susceptible to compared to Imitation Learning Based methods, due to the unbounded optimization of a learned (and probably somewhat misspecified) reward function. Active learning approaches also seem to be promising, allowing to extend

previous methods to make R be capable to take advantage of the fact that H is willing to disambiguate further between what is desired and undesired behavior. For Imitation Learning Based methods instead, it seems important to further test the feasibility and strength of the assumptions made by IDA and Debate, that are still relatively unclear due to the recency of these approaches.

Foundational research. It also seems important to investigate whether some of the fundamental assumptions underlying all these models - such as utility maximization and classical decision theory - are already restricting our capability of solving the alignment problem. In particular, decision theory doesn't have ways to account for bounded rationality, that seems to be very relevant in attempting to model humans, and thus for solving the alignment problem.

As the field is relatively new, there are most likely still low-hanging fruits to be found in solving the alignment problem. These could possibly consist new methods that fall outside of these categories entirely, in a similar way to how IDA and Debate differ pretty strongly from previous alignment approaches. Because of this, there seems to be good reason to hope for additional steps towards in solving the AI alignment problem in the coming years.

References

1. G Gordon Worley. Formally Stating the AI Alignment Problem
2. Paul Christiano. Clarifying "AI alignment"
3. Nate Soares. The Value Learning Problem, 2016; Machine Intelligence Research Institute.
4. Paul Christiano. The easy goal inference problem is still hard
5. Paul Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg and Dario Amodei. Deep reinforcement learning from human preferences, 2017; arXiv:1706.03741.
6. Dylan Hadfield-Menell, Smitha Milli, Pieter Abbeel, Stuart Russell and Anca Dragan. Inverse Reward Design, 2017; arXiv:1711.02827.
7. Sören Mindermann, Rohin Shah, Adam Gleave and Dylan Hadfield-Menell. Active Inverse Reward Design, 2018; arXiv:1809.03060.
8. Dylan Hadfield-Menell, Anca Dragan, Pieter Abbeel and Stuart Russell. Cooperative Inverse Reinforcement Learning, 2016; arXiv:1606.03137.
9. Jaime F. Fisac, Monica A. Gates, Jessica B. Hamrick, Chang Liu, Dylan Hadfield-Menell, Malayandi Palaniappan, Dhruv Malik, S. Shankar Sastry, Thomas L. Griffiths and Anca D. Dragan. Pragmatic-Pedagogic Value Alignment, 2017, International Symposium on Robotics Research, 2017; arXiv:1707.06354.
10. Dhruv Malik, Malayandi Palaniappan, Jaime F. Fisac, Dylan Hadfield-Menell, Stuart Russell and Anca D. Dragan. An Efficient, Generalized Bellman Update For Cooperative Inverse Reinforcement Learning, 2018; arXiv:1806.03820.
11. Ziebart BD. Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy, 2010; PhD thesis, Machine Learning Department, Carnegie Mellon University.
12. Dorsa Sadigh, Anca D. Dragan, Shankar Sastry, and Sanjit A. Seshia. Active Preference-Based Learning of Reward Functions, RSS 2017;
13. Paul Christiano, Buck Shlegeris and Dario Amodei. Supervising strong learners by amplifying weak experts, 2018; arXiv:1810.08575.
14. Geoffrey Irving, Paul Christiano and Dario Amodei. AI safety via debate, 2018; arXiv:1805.00899.
15. Sanjay Krishnan, Animesh Garg, Richard Liaw, Lauren Miller, Florian T. Pokorny and Ken Goldberg. HIRL: Hierarchical Inverse Reinforcement Learning for Long-Horizon Tasks with Delayed Rewards, 2016; arXiv:1604.06508.
16. Adam Gleave and Oliver Habryka. Multi-task Maximum Entropy Inverse Reinforcement Learning, 2018; arXiv:1805.08882.
17. Pedro A. Ortega, Vishal Maini, and the DeepMind safety team. Building safe artificial intelligence: specification, robustness, and assurance., 2018.
18. Rohin Shah. What is ambitious value learning?, 2018.
19. Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman and Dan Mané. Concrete Problems in AI Safety, 2016; arXiv:1606.06565.
20. Steven Kerr. On the folly of rewarding A, while hoping for B, 1975. Academy of Management journal 18.4 (1975): 769-783.
21. David Manheim and Scott Garrabrant. Categorizing Variants of Goodhart's Law, 2018; arXiv:1803.04585.
22. Stephen M. Omohundro. The Basic AI Drives, 2008;
23. Nick Bostrom. Superintelligence: paths, dangers, strategies, 2014;
24. Abbeel, Pieter, and Andrew Y. Ng. Apprenticeship learning via inverse reinforcement learning, 2004. Proceedings of the twenty-first international conference on Machine learning. ACM, 2004.
25. Owain Evans, Andreas Stuhlmueller and Noah D. Goodman. Learning the Preferences of Ignorant, Inconsistent Agents, 2015; arXiv:1512.05832.
26. Aaron Tucker, Adam Gleave and Stuart Russell. Inverse reinforcement learning for video games, 2018; arXiv:1810.10593.
27. Kareem Amin and Satinder Singh. Towards Resolving Unidentifiability in Inverse Reinforcement Learning, 2016; arXiv:1601.06569.
28. Stephane Ross, Geoffrey J. Gordon and J. Andrew Bagnell. A Reduction of Imitation Learning and Structured Prediction to No-Regret Online Learning, 2010; arXiv:1011.0686.

29. Jonathan Ho and Stefano Ermon. Generative Adversarial Imitation Learning, 2016; arXiv:1606.03476.
30. Giusti, Alessandro, et al. A Machine Learning Approach to Visual Perception of Forest Trails for Mobile Robots, 2016. IEEE Robotics and Automation Letters 1.2 (2016): 661-667.
31. Paul Christiano. Humans Consulting HCH
32. Stuart Armstrong. Humans can be assigned any values whatsoever...
33. Owain Evans, Jacob Steinhardt. Model Mis-specification and Inverse Reinforcement Learning
34. Borja Ibarz, Jan Leike, Tobias Pohlen, Geoffrey Irving, Shane Legg and Dario Amodei. Reward learning from human preferences and demonstrations in Atari, 2018; arXiv:1811.06521.
35. Jessica Taylor. Quantilizers: A Safer Alternative to Maximizers for Limited Optimization. AAAI Workshop: AI, Ethics, and Society. 2016.
36. Paul Christiano. AmbVsNarrow
37. Ng, Andrew Y., and Stuart J. Russell. Algorithms for inverse reinforcement learning. Icml. 2000.