Inferring human values for safe AGI design

Can Eren Sezener

Department of Computer Science
Ozyegin University
Istanbul, Turkey
eren.sezener@ozu.edu.tr

Abstract. Aligning goals of superintelligent machines with human values is one of the ways to pursue safety in AGI systems. To achieve this, it is first necessary to learn what human values are. However, human values are incredibly complex and cannot easily be formalized by hand. In this work, we propose a general framework to estimate the values of a human given its behavior.

Keywords: Value Learning, Inverse Reinforcement Learning, Friendly AI, Safe AGI

1 Introduction

Intelligence cannot be defined in the absence of goals[1]. Superintelligent machines will pursue some goals and if their goals are very different than those of humans', the results will likely be catastrophic. Therefore, it is of great importance to align AGI goals with human values, at least to some extent. However, this is not an easy task. Humans have complex value systems [1] and it is shown that humans are unable to determine what they value [2]. Therefore, crafting utility functions for AGI systems that encapsulate human values by hand is not viable.

Hibbard [3] suggests that learning models of humans is a viable solution for avoiding unintended AI behaviors. The agent architecture Hibbard suggests asks modeled humans to assign utility values to outcomes. However, a shortcoming of this approach is that what human models say they value and what they value can still be different.

Another possible approach is to directly estimate what humans find rewarding. Ng [4] suggests that rewards are more compact and robust descriptions of intended behaviors than full policies or models of agents. In fact, for imitation learning, it is argued that just learning the policy of the teacher is more limited and hence less powerful than extracting the teacher's reward function and then calculating a policy. Furthermore, once we obtain a reward function, we can modify it to alter the agent's behavior, which is easier than modifying the full policy of the agent directly. Soares [5] suggests using methods similar to *inverse reinforcement learning* (IRL) for learning human values. However, the

[1] We use goals, rewards, utilities, and values interchangeably in this work.

current IRL methods are limited and cannot be used for inferring human values because of their long list of assumptions. For instance, in most IRL methods the environment is usually assumed to be stationary, fully observable, and sometimes known; the policy of the agent is assumed to be stationary and optimal or near-optimal; the reward function is assumed to be stationary as well; and the Markov property is assumed. Such assumptions are reasonable for limited motor control tasks such as grasping and manipulation; however, if our goal is to learn high-level human values, they become unrealistic. For instance, assuming that humans have optimal policies discards the possibility of superintelligent machines and ignores the entire cognitive biases literature. In this work, we propose a general framework for inferring the reward mechanisms of arbitrary agents that relaxes all the aforementioned assumptions. Through this work, we do not only intend to offer a potential solution to the problem of inferring human values (i.e., the so-called Value Learning Problem [5]), but also stimulate AI researchers to investigate the theoretical limits of IRL.

2 Inferring human values

As in Hutter's work [6], we model an agent by a program p_A that determines the policy of the agent when run on a universal Turing machine (UTM), and the environment by an arbitrary function. In Hutter's AIXI model [6], the rewards are computed by the environment. We assume that rewards are computed by a distinct process called the *reward mechanism*, which we model by the program p_R. This is a reasonable assumption from a neuroscientific point of view because all reward signals are generated by brain areas such as the striatum. We model the agent, the reward mechanism, and the environment as processes that work in synchronization and in a sequential manner as illustrated in Figure 1. p_A reads $r_t \in [r_{min}, r_{max}]$ and $o_t \in O$ and writes $a_t \in A$, where O and A are sufficiently large and finite observation and action spaces. Then, the environment reads a_t and writes o_{t+1}. Subsequently, p_R reads o_{t+1} and writes r_{t+1} and so on. Now our problem reduces to finding the most probable p_R given the entire action-observation history $a_1 o_1 a_2 o_2 \dots a_n o_n$.

Solomonoff [7] proposed the *universal prior* $M(x)$ as the probability of a UTM outputting a string with the prefix x. Formally, $M(x) := \sum_{p:U(p)=x*} 2^{-l(p)}$ is the universal prior where $l(p)$ is the length of the program p, $U(p)$ is the output of a UTM that simulates p, and $x*$ is a string with the prefix x. Hutter extended the definition of universal prior to programs, and defined a universal prior over programs as $m(p) := 2^{-l(p)}$ [6]. Similarly, by assuming the independence of prior probabilities of p_R and p_A, we can get their joint prior as $m(p_A, p_R) = 2^{-(l(p_A)+l(p_R))}$. Then, we can obtain the probability of p_R being the true reward generating program given an action-observation history as:

$$m(p_R||a_{1:n}, o_{1:n}) = \sum_{p_A : p_A(p_R(o_{1:n}), o_{1:n}) = a_{1:n}} 2^{-(l(p_R)+l(p_A))} \tag{1}$$

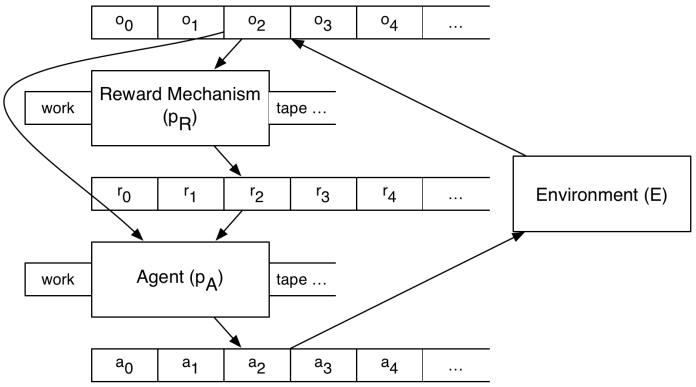

Fig. 1. The interaction between the agent, the environment, and the reward mechanism.

where $a_{1:n} := a_1 a_2 \dots a_n$, $o_{1:n} := o_1 o_2 \dots o_n$, and $p_R(o_{1:n}) = r_1 r_2 \dots r_n$. It should be noted that $\sum_{p_R} m(p_R||a_{1:n}, o_{1:n}) \neq 1$ and the true probability measure can be obtained via normalization. We also assume that the agent cannot access the reward mechanism directly, but can only sample it. If the agent has access to the reward mechanism, $p_A(p_R(o_{1:n}), o_{1:n})$ in (1) should be replaced with $p_A(p_R(o_{1:n}), p_R, o_{1:n})$.

Equation 1 provides a simple way to to estimate reward mechanisms of arbitrary agents with a very few assumptions. We do not assume Markov property, fully-observable and stationary environments, optimal and stationary policies, or stationary rewards. However, this degree of generality comes with high computational costs. Due to the infinite loop over the programs and the existence of non-halting programs, this solution is incomputable. Nevertheless, one can obtain approximations of (1) or use different complexity measures (such as Schmidhuber's Speed Prior [8]) in order to obtain computable solutions.

It should also be noted that even though we assumed deterministic agents and reward mechanisms and fully-observable action-observation histories, these assumption can be relaxed and a framework that assumes probabilistic agent and reward functions and noisy action-observation histories can be developed.

3 Discussion

In principle if we can capture the actions and observations of a human with high accuracy, we might be able to estimate its values. This is a potential solution for the Value Learning Problem [5]. For example, we can infer the values of some individuals who are 'good' members of the society and possess 'desirable' values. Then we can preprocess the inferred values and give a mixture of them to an AGI system as its reward mechanism. The preprocessing stage would involve

weeding out states/activities that are valuable for biological agents but not for robots such as eating[2]. How to achieve this is an open problem.

Dewey [9] suggests an AGI architecture that replaces the rewards in AIXI with a utility function as well. The proposed agent can either be provided with a hand-crafted utility function or a set of candidate, weighted utility functions. If the latter is the case, the agent can improve its utility function by adjusting the weights. However, it is not specified *how* the agent should or can do the adjustments. Furthermore, the proposed agent improves its utility function through interacting with the environment, whereas we suggest that human values should be estimated and processed first and then be provided to an AGI system.

Acknowledgements

I would like to thank Erhan Oztop for helpful discussions and comments and the anonymous reviewers for their suggestions.

References

1. Eliezer Yudkowsky. Complex value systems in friendly ai. In Jürgen Schmidhuber, Kristinn R. Thrisson, and Moshe Looks, editors, *Artificial General Intelligence*, volume 6830 of *Lecture Notes in Computer Science*, pages 388–393. Springer Berlin Heidelberg, 2011.
2. Luke Muehlhauser and Louie Helm. The singularity and machine ethics. In Amnon H. Eden, James H. Moor, Johnny H. Sraker, and Eric Steinhart, editors, *Singularity Hypotheses*, The Frontiers Collection, pages 101–126. Springer Berlin Heidelberg, 2012.
3. Bill Hibbard. Avoiding unintended ai behaviors. In Joscha Bach, Ben Goertzel, and Matthew Ikl, editors, *Artificial General Intelligence*, volume 7716 of *Lecture Notes in Computer Science*, pages 107–116. Springer Berlin Heidelberg, 2012.
4. Andrew Y. Ng and Stuart J. Russell. Algorithms for inverse reinforcement learning. In *Proceedings of the Seventeenth International Conference on Machine Learning*, ICML '00, pages 663–670, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.
5. Nate Soares. The value learning problem. Technical report, Machine Intelligence Research Institute, Berkeley, CA, 2015.
6. Marcus Hutter. *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability*. Springer, Berlin, 2005.
7. R.J. Solomonoff. A formal theory of inductive inference. part i. *Information and Control*, 7(1):1 – 22, 1964.
8. Jürgen Schmidhuber. The speed prior: A new simplicity measure yielding near-optimal computable predictions. In *Computational Learning Theory, 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8-10, 2002, Proceedings*, pages 216–228, 2002.
9. Daniel Dewey. Learning what to value. In Jürgen Schmidhuber, Kristinn R. Thorisson, and Moshe Looks, editors, *AGI*, volume 6830 of *Lecture Notes in Computer Science*, pages 309–314. Springer, 2011.

[2] This should be done such that the robot will not value consuming food but will value providing humans with food.