Directions in Hybrid Intelligence: Complementing AI Systems with Human Intelligence

Ece Kamar
Microsoft Research
eckamar@microsoft.com

Abstract

Hybrid intelligence systems combine machine and human intelligence to overcome the shortcomings of existing AI systems. This paper reviews recent research efforts towards developing hybrid systems focusing on reasoning methods for optimizing access to human intelligence and on gaining comprehensive understanding of humans as helpers of AI systems. It concludes by discussing short and long term research directions.

1 Introduction

Historically, an overarching goal for AI has been exhibiting abilities that come naturally to people such as making sense of the world and acting in it to achieve goals. Despite recent advances, AI systems are far from being perfect. When these systems are left to function without human assistance, they may occasionally make mistakes or completely fail. With AI systems becoming a bigger part of daily lives by carrying out critical tasks such as driving, mistakes and failures of these systems negatively affect user trust and can even lead to drastic consequences such as loss of lives.

The central idea of this paper is that instead of designing AI systems that function alone, we should focus on hybrid systems that can benefit from partnership with humans. In a hybrid system, human intelligence can be integrated into an AI system to complement machine capabilities (i.e., to form hybrid intelligence) throughout its life cycle. Hybrid systems can offload computational tasks to humans on demand to overcome the deficits of AI systems. Human involvement can prevent the mistakes and failures that would be caused by an AI system working alone and the feedback from humans can lead to a virtuous improvement cycle for the system to continuously learn from.

The need for human involvement to overcome the mistakes and limitations of AI systems is already acknowledged in critical domains such as medicine and driving. For example, a driver of a semi-autonomous car is expected to continuously watch over the decisions of the machine and correct it when needed to prevent accidents. However, successfully integrating human and machine intelligence together has its challenges. Human intelligence is a valuable resource associated with costs and constraints. The quality and availability of human input may vary depending on many factors including the state of the human. AI systems need to be equipped with reasoning capabilities that can make effective decisions about accessing human intelligence.

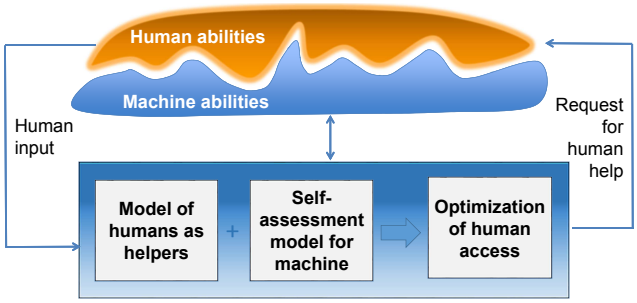

Figure 1: Reasoning capabilities for hybrid intelligence.

Recent advances in human computation closely relate to the efforts on hybrid intelligence. Crowdsourcing platforms provide easy access to human intelligence on demand in a scalable and versatile way. For AI systems in which a user is not in the loop to provide help, the human help needed by the system may be provided by the crowd. For many research efforts, including the ones presented here, crowdsourcing platforms function as testbeds for data collection and experimentation to inform us about the challenges about accessing and working with human intelligence.

This paper summarizes research efforts influenced from the vision of hybrid intelligence. These efforts aim at developing reasoning capabilities needed by an AI system to determine when and how it can benefit from the complementary strengths of human intelligence (see Figure 1 for an outline). These reasoning capabilities build on self-assessment models for assessing machine capabilities as well as models specifying human helpers with respect to their abilities, availabilities and costs. Overview of these research efforts in this paper are grouped under two sections; Section 2 on reasoning capabilities for optimizing access to human intelligence and Section 3 on specifying humans as helpers to AI systems.

We review the efforts on optimizing access to human intelligence in two settings. In the first setting, a system can acquire additional evidence from humans to accomplish crowdsourcing tasks. We show how machine learning and decision-

theoretic optimization can be used together to guide the allocation of human effort. Second, we describe how an agent can learn more effectively with help from a "teacher" (i.e., more experienced agent or human) than learning alone when it can query the teacher about how to act. The algorithms designed to decide when the agent receives teacher input not only reasons about the performance of the agent but also the cognitive and communication overheads on the teacher for providing assistance.

Next, we present an overview of efforts towards a comprehensive understanding of humans as helpers to AI systems. These efforts focus on accessing human input through crowdsourcing and investigate how high-quality human input can be acquired with advances in task design, quality control models, incentives, interaction techniques and worker training.

The paper concludes by outlining future directions for hybrid intelligence. In short term, we discuss different ways human intelligence can be integrated into AI systems through training, execution and evaluation. For the long term, we discuss how stronger models of hybrid intelligence can be achieved by treating humans as partners rather than helpers by making teamwork an integral component of AI systems.

2 Optimizing Access to Human Intelligence

This section reviews models and algorithms designed for reasoning about the value of two different forms of human assistance for AI systems; when a system seeks additional evidence from humans to accomplish tasks, and when a learning agent has access to teacher advice on how to act.

2.1 Solving Tasks with Human Help

The CrowdSynth effort combines machine learning and decision-theoretic optimization techniques to leverage the complementary strengths of humans and machines for solving crowdsourcing tasks [Kamar *et al.*, 2012]. This effort focuses on solving consensus tasks, a common task type in crowdsourcing, where the goal is uncovering the true answer of each task by collecting multiple assessments from human workers in addition to machine analysis that may be performed on tasks. It uses Galaxy Zoo, a citizen science project that seeks volunteers' input to classify images of millions of celestial objects, as a testbed for studies.

Each Galaxy Zoo task is associated with 453 image features generated with automated computer vision. CrowdSynth uses supervised learning to infer accuracy of automated analysis for labeling images as well as the accuracies of individual Galaxy Zoo workers. Optimizing access to human input for a given Galaxy Zoo task hinges on trading off the long-term expected value of acquiring an additional assessment from a crowd worker with the immediate cost of hiring. CrowdSynth formalizes this decision-making problem as a Partially Observable Markov Decision Process, which makes calls to the machine learned models to make inferences about the current state and future transitions.

This decision-making problem introduces a number of planning challenges which generalize beyond of solving consensus tasks. These challenges are formalized under the class of long evidential sequence (LES) tasks in which each observation provides a weak evidence about the state of the world but sets of observations may provide significant value [Kamar and Horvitz, 2013]. Accurately reasoning about the value of acquiring an observation (e.g., input from a single human) requires searching a space that grows exponentially in the horizon. We formulated MC-VOI, a Monte-Carlo planning algorithm that exploit the structure of LES tasks to successfully cut through the intractability of the combinatorial space.

Evaluations of the CrowdSynth effort on Galaxy Zoo demonstrate that significant gains can be achieved from the optimization of access to human intelligence. The experiments show that CrowdSynth can achieve the maximum accuracy of the original system by hiring only 47% of the workers who participated in the open world run of the system. Under a fixed budget, the gains from CrowdSynth is 2.4 times of allocating workers to tasks with a random policy.

In follow-up work, we extend the modeling and algorithmic approaches of CrowdSynth to solving Hierarchical Consensus Tasks (HCTs) [Kamar and Horvitz, 2015]. Hierarchical consensus tasks seek correct answers to a hierarchy of subtasks, where branching depends on answers at preceding levels of the hierarchy. We show that solving HCTs are exponentially harder than solving consensus tasks due to the branching of task hierarchy. We address this complexity by customizing MC-VOI to HCTs and experimental evaluations show the approach achieving additional gains by reasoning not only about hiring workers but also about which part of the task hierarchy to cover with each worker.

Another extension is CrowdExplorer, designed for *lifelong learning* settings, in which historical data for learning models of human help is not available [Kamar *et al.*, 2013]. CrowdExplorer combines Monte-Carlo planning with Bayesian learning to manage the exploration-exploitation trade-off in adaptive control of crowdsourcing tasks. The proposed technique can simultaneously optimize access to human intelligence while learning models of workers from their contributions.

2.2 Learning How to Act

Agents learning how to act in new environments can benefit from the input of humans or more experienced teachers on which action to take next. This framework of student-teacher training has been proposed by Torrey and Taylor and strategies on when to ask/provide advise have been studied from the student's and the teacher's perspectives respectively [Torrey and Taylor, 2013; Clouse, 1996]. Previous work demonstrated that the teacher continuously watching over the decisions of the student and guiding decisions on when to provide advice causing significant learning gains for the student. However these teacher-initiated approaches introduce unrealistic attention and communication demands when a human acts as the teacher.

We address this shortcoming by proposing interactive teaching strategies in which the student and the teacher jointly identify advising opportunities [Amir *et al.*, 2016]. These strategies do not require the teacher to continuously monitor the student but instead involve the teacher to verify the decisions of the student when the student asks for advice. Evaluations demonstrated that these approaches reduce the amount

of attention required of the teacher compared to teacher-initiated strategies, while maintaining similar learning gains.

The work on interactive teaching strategies builds on CrowdSynth effort by reasoning about humans not only as providers of help but also as active participators of the decision-making process that manage when help is needed. It is also motivated by the special considerations around having humans as helpers of AI systems.

3 Specifying Humans as Helpers

An AI system grappling with the decision of accessing human help needs to have an understanding of the capabilities of its helper and the costs and constraints associated with asking for help. As opposed to the computational resources used in the development of an AI system, human helpers do not come with a specification; even through a constrained interaction such as crowdsourcing, many factors including task design, incentives and training may affect human behavior. This section presents an overview of research efforts on gaining an understanding of human work in crowdsourcing in order to develop ideal methods for accessing human input.

3.1 Task Design

A core requirement for integrating human input into an AI system is having a translation of system state and needs to humans in an understandable way. In crowdsourcing, this corresponds to the challenge of task design to elicit high quality work from the crowd. We performed two separate studies on how to elicit high-quality crowd input to be integrated into an existing spoken dialog system. The first study explored task designs to collect language diversity corresponding to semantic forms used by the system so that the resulting corpora can be used in the training of the system for language understanding [Wang *et al.*, 2012]. The second study investigated how crowd can participate in the language generation process of a spoken dialog system [Mitchell *et al.*, 2014].

The goal of these studies were not creating task templates that generalize to any AI system. Instead, they provided guidelines on eliciting high-quality crowd work by drawing attention to the errors and biases created by task design and described workflows to mitigate these errors.

3.2 Modeling Worker Bias

A well-known problem with crowd work is the noise in the contributions of individual workers. Consequently researchers have developed machine learning models that can learn about workers' quality, bias and expertise and the relationship of their contributions to ground truth answers of tasks (i.e., [Ipeirotis *et al.*, 2010]). When individual workers' noise is independent, these models can accurately correct individuals' mistakes. However, they may fail when task characteristics induce a population-wide or subgroup specific bias in worker contributions. We developed a family of probabilistic graphical models that can successfully learn about the task-dependent worker bias and correct it to accurately infer ground truth answers [Kamar *et al.*, 2015].

3.3 Monetary Incentives and Performance

People may have different motivations to contribute to crowd work; participation to volunteer crowd work like citizen science may emerge from altruism and interest in science, whereas workers in paid crowdsourcing may be influenced from the amount and structure of monetary incentives offered for a task. How to design the best incentive structure for a given task depends on understanding the way different incentives affect the quality of work obtained from the crowd.

Through a set of experiments using tasks from a well-known citizen science project called Planet Hunters, we studied how different incentive structures lead to tradeoffs in quality and effort in paid crowd work [Mao *et al.*, 2013a]. The analysis provided generalizable insights on the way crowd workers adjust their work to maximize monetary payments. The comparison of the paid crowd work with the work of Planet Hunters' volunteers demonstrated that the quality of paid crowd work can be comparable to the quality of self-motivated crowd with the right incentive structure, showing the viability of paid crowd work for acquiring human intelligence for complex tasks.

3.4 Engagement in Volunteer Work

Success of volunteer crowdsourcing such as citizen science depend on the continuous contributions of volunteers. Like paid crowdsourcing, participants of citizen science projects follow a power-law distribution, where the majority produces few contributions. Unlike paid crowdsourcing, citizen science platforms cannot increase worker contributions by offering higher monetary incentives. How to improve engagement is a vital problem for the well-being of these platforms.

We investigated this engagement problem in two steps. First, we studied the historical data collected from Galaxy Zoo project to develop machine learned models predicting whether a volunteer worker is likely to disengage from the current session in a limited time window [Mao *et al.*, 2013b]. In a follow-up study, we used the predictions of the machine learned models to time intervention messages that are hand-crafted to increase engagement of workers. Controlled studies run live on the Galaxy Zoo platform showed that a message emphasizing the helpfulness of individuals' contributions significantly improved the amount of work produced by workers when the messages are delivered according to the predicted times of disengagement [Segal *et al.*, 2016].

This work presents a general methodology that combines machine learning with intervention design to study and improve engagement in volunteer crowd work. In addition, it demonstrates an example of models of human helpers being used to generate significant improvements in the way human input is acquired from crowdsourcing.

3.5 Training Strategies

Another characteristic of human help separating it from computational resources used in AI systems is that human capabilities are not static, they extend to many different types of tasks. A promising direction for crowd work is whether crowd workers can be trained to acquire new capabilities. We performed controlled studies on a paid crowdsourcing platform to measure the effectiveness of different training strate-

gies in improving the performance of workers for accomplishing complex tasks [Doroudi *et al.*, 2016]. The studies compared the effectiveness of five different training strategies including expert examples, which require additional work from a domain expert, and peer validation, which asks workers to validate the work of their peers as a form of training. The comparisons showed both expert examples and peer validation methods leading to significant improvements in worker performance for these complex tasks and that the validation approach can be as good as the expert examples approach if the task solutions to be validated are pre-filtered to produce better learning outcomes. The positive results associated with the validation condition suggest a self-sufficient automated training pipeline within crowdsourcing platforms in which the community of crowd workers train each other to accomplish difficult tasks accurately without the involvement of an expert. The results also provide additional support to the observation that paid crowd work is not limited to microtasks that come naturally to humans but it is a viable approach to accomplishing complex tasks.

4 Future Directions

With AI systems becoming an integral part of our daily lives, the exciting and timely research directions emerging from the partnership of humans and AI systems are not limited to the research efforts reviewed in this paper. We seek innovative applications of hybrid intelligence to understand its potential as well as limitations. There is need for generalizable models, algorithms and workflows to move away from hand-crafted hybrid systems to optimized access to human input. More work is needed to develop a comprehensive specification of humans as helpers that can be used in deciding whether, when and how to access human input.

Integration of human input into AI systems offers great promise for the development of practical applications. Human computation is already an important resource in the training of AI systems. Advances in crowd platforms allowing for real-time access to crowd can enable crowd input to be used in the execution of AI systems to prevent mistakes and shortcomings. AI systems can be paired with humans through crowdsourcing to enable testing these systems with a large and diverse set of subjects. These crowd powered testbeds can be shared among researchers to evaluate AI readiness and for tracking progress in the field.

In the longer term, involvement of humans as helpers to AI systems may be too limiting for critical applications that seek a deeper integration of human and machine intelligence to function. For example the driver of a semi-autonomous car does not only provide assistance when being asked, but proactively engages in the activity of driving. Developing AI systems that can function as effective team members to humans requires a paradigm shift from hybrid systems to hybrid teamwork. It requires deeper reasoning capabilities for machines to make decisions not only about how they are accomplishing their tasks, but also about how they can support their teammates towards the success of the collaborative activity. This promising direction can build on the rich literature on formal models of teamwork to develop new representations and decision-making approaches that can reason about the hybrid nature of human-computer teamwork.

References

[Amir *et al.*, 2016] Ofra Amir, Ece Kamar, Andrey Kolobov, and Barbara Grosz. Interactive teaching strategies for agent training. In *IJCAI*, 2016.

[Clouse, 1996] Jeffery Allen Clouse. On integrating apprentice learning and reinforcement learning. 1996.

[Doroudi *et al.*, 2016] Shayan Doroudi, Ece Kamar, Emma Brunskill, and Eric Horvitz. Toward a learning science for complex crowdsourcing tasks. In *CHI*, 2016.

[Ipeirotis *et al.*, 2010] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, 2010.

[Kamar and Horvitz, 2013] Ece Kamar and Eric Horvitz. Light at the end of the tunnel: A Monte Carlo approach to computing value of information. In *AAMAS*, 2013.

[Kamar and Horvitz, 2015] Ece Kamar and Eric Horvitz. Planning for crowdsourcing hierarchical tasks. In *AAMAS*, 2015.

[Kamar *et al.*, 2012] Ece Kamar, Severin Hacker, and Eric Horvitz. Combining human and machine intelligence in large-scale crowdsourcing. In *AAMAS*, 2012.

[Kamar *et al.*, 2013] Ece Kamar, Ashish Kapoor, and Eric Horvitz. Lifelong learning for acquiring the wisdom of the crowd. In *IJCAI*, 2013.

[Kamar *et al.*, 2015] Ece Kamar, Ashish Kapoor, and Eric Horvitz. Identifying and accounting for task-dependent bias in crowdsourcing. In *HCOMP*, 2015.

[Mao *et al.*, 2013a] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *HCOMP*, 2013.

[Mao *et al.*, 2013b] Andrew Mao, Ece Kamar, and Eric Horvitz. Why stop now? Predicting worker engagement in online crowdsourcing. In *HCOMP*, 2013.

[Mitchell *et al.*, 2014] Margaret Mitchell, WA Redmond, Dan Bohus, and Ece Kamar. Crowdsourcing language generation templates for dialogue systems. In *INLG and SIGDIAL 2014*, 2014.

[Segal *et al.*, 2016] Avi Segal, Ya'akov Gal, Ece Kamar, Eric Horvitz, Alex Bowyer, and Grant Miller. Intervention strategies for increasing engagement in crowdsourcing: Platform, predictions, and experiments. In *IJCAI*, 2016.

[Torrey and Taylor, 2013] Lisa Torrey and Matthew Taylor. Teaching on a budget: Agents advising agents in reinforcement learning. In *AAMAS*, 2013.

[Wang *et al.*, 2012] Wei Yu Wang, Dan Bohus, Ece Kamar, and Eric Horvitz. Crowdsourcing the acquisition of natural language corpora: Methods and observations. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, 2012.