Artificial Intelligence Review (2024) 57:322
https://doi.org/10.1007/s10462-024-10974-1

Trustworthy human computation: a survey

Hisashi Kashima[1] · **Satoshi Oyama**[2] · **Hiromi Arai**[3] · **Junichiro Mori**[4]

Accepted: 26 September 2024 / Published online: 12 October 2024

Abstract
Human computation is an approach to solving problems that prove difficult using AI only, and involves the cooperation of many humans. Because human computation requires close engagement with both "human populations as users" and "human populations as driving forces," establishing mutual trust between AI and humans is an important issue to further the development of human computation. This survey lays the groundwork for the realization of trustworthy human computation. First, the trustworthiness of human computation as computing systems, that is, trust offered by humans to AI, is examined using the RAS (reliability, availability, and serviceability) analogy, which define measures of trustworthiness in conventional computer systems. Next, the social trustworthiness provided by human computation systems to users or participants is discussed from the perspective of AI ethics, including fairness, privacy, and transparency. Then, we consider human–AI collaboration based on two-way trust, in which humans and AI build mutual trust and accomplish difficult tasks through reciprocal collaboration. Finally, future challenges and research directions for realizing trustworthy human computation are discussed.

Keywords Human-in-the-loop AI/ML · Reliability/availability/serviceability of human computation · AI Ethics · Collaborative intelligence

1 Trustworthiness in human computation

1.1 Rise of human computation

AI technologies centered on machine learning, as represented by the rise of deep learning, have made remarkable progress in recent years. Its application areas are not limited to information and communication related fields, including recommendation and web marketing, but have expanded across many disciplines to include manufacturing, transportation, medicine, and various science related fields where a variety of innovations are expected. In general, machine learning can only be applied to areas where system inputs and outputs can be formally described and where a large amount of data can be collected. Furthermore, simi-

Extended author information available on the last page of the article

lar to humans, AI decisions always contain errors. This indicates that human teaching, judgment, and feedback at important points are necessary in critical application domains, such as automated driving and medical diagnostics, where research and development is ongoing and human lives are at stake. This trend is more pronounced for complex and critical problems. Real-world problems often do not require AI solution only, but involve some form of human participation. In future, we will face more difficult real-world challenges that require a closer integration of AI and human-driven systems. Over the past decade, AI research has focused on enhancing machine intelligence, especially through deep learning. However, as the impact of AI on society grows, the interface between humans and AI is being discussed in terms of both technology and regulation. For example, Ethically Aligned Design proposed by The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019), Principles on Artificial Intelligence by OECD (2019), and Ethics Guidelines for Trustworthy AI by High-Level Expert Group on Artificial Intelligence of the European Commission (2019) have been proposed to promote the use of AI with respect to humans.

The approach of enlisting the help of (often unspecified) large numbers of humans to solve problems that are difficult for AI to solve by itself is called "human computation" (Law and von Ahn 2011). Early examples include ReCAPTCHA (von Ahn et al. 2008), which solves the limitations of AI recognition performance with the help of humans through tasks hidden in an access control interface, and VizWiz (Bigham et al. 2010), a human-in-the-loop visual question-answering application to aid the visually impaired. The development of human computation was strongly encouraged by the concept of crowdsourcing, which emerged around the same time as human computation, and general-purpose crowdsourcing platforms such as Amazon's Mechanical Turk, which facilitate on-demand employment of an unspecified number of crowd workers. In 2013, HCOMP (AAAI Conference on Human Computation and Crowdsourcing), an academic community dedicated to human computation, was established and steadily progressed.

Many of the early studies in the field of human computation involved testing ability of general public participation with standard capabilities to solve challenges that proved difficult for AI to solve, given the new crowdsourcing platforms. (Various early research studies have been summarized by Law and von Ahn (2011).) Subsequently, attempts were made to establish a systematic design theory for human–AI systems that went beyond a mere collection of successful and unsuccessful cases. The development of a general framework for building various human-computation systems (Little et al. 2010a, b; Kittur et al. 2011) is one example of such an attempt. (More details are provided in Sect. 2.)

1.2 Toward trustworthy human computation

Similar to the discussions on the reliability of AI, it is essential to discuss the trustworthiness of human computation to play a more active role in human society. Before discussing this issue, it is necessary to define trustworthiness in human computation, in which there are at least two kinds. One relates to computing systems, while the other is based on trust between human computation systems and humans who interact with them. The former makes human computation systems as versatile and trustworthy as conventional computer systems, while the latter is described as the social and ethical responsibilities of systems that interact with humans, such as fairness and accountability.

When we implement human computation, we presumably try to achieve a particular goal (perhaps one that is difficult to achieve by using conventional computer systems alone). To achieve this goal, it is desirable to handle human computation with the same trustworthiness as that of conventional computers. Such "human computers" will utilize the collective power of humans as its driving force. Because humans have a great deal of uncertainty compared with conventional computational units, controlling these uncertainties is directly related to the assurance of trustworthiness. In previous studies, such trustworthiness was discussed from various perspectives, such as the motivation to participate in human computation (Mason and Watts 2009; Feyisetan and Simperl 2019) and quality assurance of crowdsourced results (Snow et al. 2008; Whitehill et al. 2009).

Human computation involves not only "human populations as driving forces" but also "human populations as users." The use of AI technology in various aspects of society is currently under consideration, which increases the demand for AI system trust in a social context. Human computation cannot be further expanded without the trust of the latter; conversely, this concept is not sustainable without the trust of the former.

Moreover, beyond the above mentioned trusts between human computation and humans from one side to the other, the ultimate goal of human computation is to solve more difficult problems by having mutual trust and collaborating with each other. Not to mention the anecdote of the victory of a mixed team in the freestyle chess games, in which humans and AI are freely paired up (Brynjolfsson and McAfee 2011), mutually beneficial reciprocal cooperation should enhance each other's capabilities through positive feedback cycles, resulting in the achievement of higher goals. This can be the motivation for mutual trust and cooperation, which we place at the top of our list as the third category of trustworthy human computation.

Based on the above considerations, this study will focus on the interactions with humans inside and outside of human computation systems, and organize various existing studies according to trustworthiness, thereby laying the groundwork for discussions aimed at trustworthy human computation.

1.3 Existing surveys and reviews

Individual surveys have been conducted on human computation, crowdsourcing, and the relationship between humans and AI. This survey differs from previous studies in that it examines the trustworthiness of human computation and organizes the existing studies from this perspective.

An overview of early research on human computation can be found in the book by Law and von Ahn (2011) and the survey by Quinn and Bederson (2011). The handbook (Michelucci 2013) has collected a wide range of related topics. Daniel et al. (2018) surveyed various methods for quality control of human computation.

Several survey papers summarized recent discussions on trustworthy AI. Kaur et al. (2022) conducted a comprehensive review of trustworthy AI, where they discussed the trustworthiness of AI from various perspectives, including ethical discussions such as fairness, human acceptability (explainability and human participation), and safety (privacy and security). Thiebes et al. (2021) summarized the concepts and perspectives of a trustworthy AI. In addition, a survey on AI fairness (particularly in the context of machine learning) was conducted by Mehrabi et al. (2021). Techniques to help humans understand the decisions of

(often black-box) AI are discussed as XAI (explainable AI), and a survey on this topic has bee conducted by Guidotti et al. (2018).

Collaboration with humans is built on trust in reciprocal relationships between humans and AI. Surveys (Vaughan 2017; Mosqueira-Rey et al. 2022) and a book (Monarch 2021) on human-in-the-loop machine learning discuss human participation in the machine learning process. Taking this a step further, some surveys (Seeber et al. 2020; Lai et al. 2021; Vereschak et al. 2021) were conducted on collaborative work and decision-making between AI and humans.

As mentioned above, there are various perspectives on the trustworthiness of AI; however, yet none of the above are based on the unique perspective of human computation.

1.4 Contributions of this survey

The contributions of this survey, which focuses on trust in human computation, are fourfold:

1. We classified trustworthiness in human computation into three perspectives based on the direction of trust.
2. We organized the trust of human computation systems as computing systems in terms of reliability, availability, and serviceability (RAS).
3. We elaborated on the existing discussions of social trust of AI within the context of human computation and clarified its specific characteristics.
4. We identified and examined various approaches to human–AI collaboration in human computation, focusing on their characteristics and trust aspects.

1.5 Structure of this survey

In the following three chapters, this survey summarizes the arguments for achieving trustworthy human computation from three perspectives: (a) trust in humans from the human computation system, (b) trust in human computation systems from humans, and (c) collaborative human computation based on trust in reciprocal relations between AIs and humans (Fig. 1).

In Sect. 2, we discuss the perspective of the trustworthiness of human computation as computing systems, that is, the trust possessed by humans in AI. In particular, we discuss human computation from the perspective of RAS (reliability, availability, and serviceability) (Siewiorek and Swarz 1998), which are standard reliability and security measures in conventional computer systems.

Conversely, in Sect. 3, we discuss the trustworthiness that a human computation system offers to society and participants in human computation. With the social advancement of AI, there has been much discussion about the ethics AI and social trustworthiness of AI, such as fairness and privacy (Kaur et al. 2022). This survey summarizes previous discussions, focusing on perspectives related to crowdsourcing workers who play an important role in the context of human computation.

In Sect. 4, we discuss collaborative human computation, exploring several key forms of human–AI collaboration that is characterized by varying degrees of reciprocity and initiative between humans and AI: first, human-in-the-loop AI systems that explicitly involve

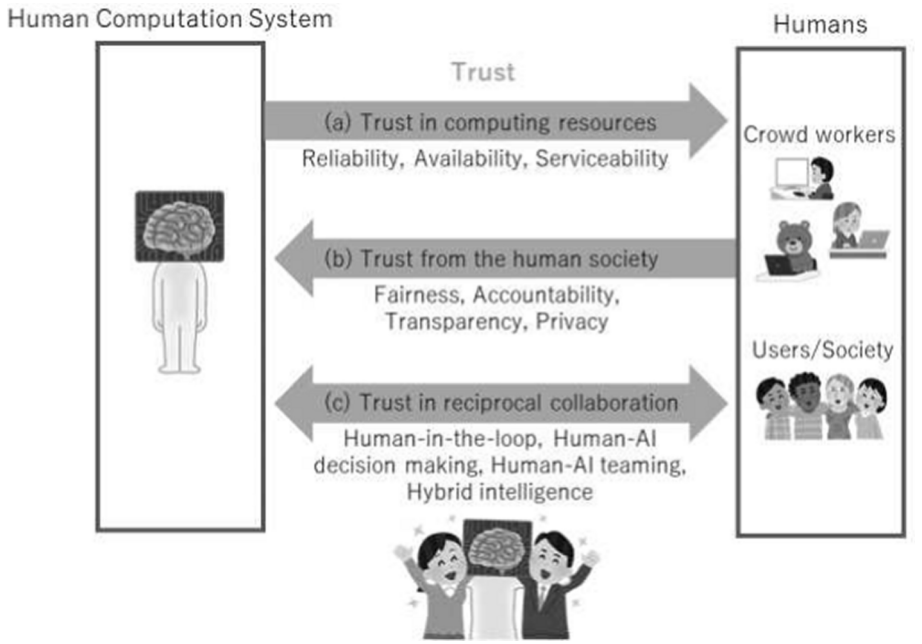

Fig. 1 There are two types of trust in human computation: **a** trust from the human computation system to humans (crowd workers) as computational resources, and **b** trust from human groups (crowd workers, users, and society) to the human computation system. Based on these trusts, the goal of human computation is to solve difficult problems through **c** cooperation between both parties based on trust in reciprocal relationships

human participation in AI systems; second, algorithm-in-the-loop (or decision supporting) systems that focus on AI assisting humans; third, human–AI teaming systems that work together to tackle complex problems; and finally, hybrid intelligence systems that build mutually reciprocal relationships between AI and humans through collaboration. Throughout the chapter, we explore the critical role of trust in collaborative human computation.

In Sect. 5, based on previous discussions, we present future challenges and research directions for achieving trustworthy human computation.

2 Trustworthy human computation systems

One aspect of human computation is the use of humans as a computing resource. In this chapter, we discuss the trustworthiness of human computation in computing systems driven by human labor, that is, in terms of the trustworthiness offered by humans to AI.

Although there have been many discussions on quality control in human computation (Daniel et al. 2018), metrics that consider multiple aspects of quality, as is the case in traditional computer systems, are not evident. One aspect of trustworthiness in human computation is the quality of a correct task execution. Human computation often uses crowdsourcing as a driving force. Crowdsourcing ranges from commercial, such as Amazon Mechanical Turk, to volunteer participation, such as citizen science projects (Willett et

Table 1 Studies on reliability in human computation

Category	References	Description
Worker ability	Whiting et al. (2017)	Mutual rating among workers
	Kazai et al. (2011)	Evaluating worker reliability by tasks with true answers
	Ipeirotis et al. (2010)	Identifying workers with malicious intent or misinterpretation of tasks
	Raykar and Yu (2011)	Introducing the Spammer Score
	Sheng et al. (2008)	Conditions for successful majority voting
	Snow et al. (2008)	Weighted majority voting by worker accuracy
	Sakurai et al. (2013)	Discriminating accurate workers with different reward plans
	Dawid and Skene (1979)	Estimating worker ability without true answers
	Venanzi et al. (2014)	Considering worker group structure
	Li et al. (2019b)	Modeling correlations between workers
Task-worker relation	Whitehill et al. (2009)	Considering task difficulty
	Welinder et al. (2010)	Considering multidimensional characteristics of tasks and workers
	Bachrach et al. (2012)	Bayesian model considering both worker ability and task difficulty
	Oyama et al. (2013)	Considering workers' confidence about their answers
	Ambati et al. (2011)	Predicting preferred tasks based on worker attributes
	Yuen et al. (2012)	Recommendation using task browsing history and past tasks
	Kulkarni et al. (2012b)	Recommendation of tasks by other workers
	Li et al. (2014)	Identifying appropriate workers for a task based on their attributes
	Bender and Friedman (2018)	Different worker attributes between data collection and use
	Sheng et al. (2008)	Calculating the uncertainty of answers
	Donmez et al. (2009)	Trade-off between exploration and exploitation of workers
Various tasks	Chen et al. (2013)	Ranking aggregation from pairwise comparisons
	Matsui et al. (2014)	Aggregation of multiple ranking lists
	Wu et al. (2012)	Aggregation of multiple sequences
	Wang and Dang (2022)	Aggregation of multiple text sentences
	Baba and Kashima (2013)	Quality control using a generation-evaluation process
Cognitive bias	Draws et al. (2021)	Effects of cognitive bias
	Newell and Ruths (2016)	Influence of the previous task
	Eickhoff (2018)	Existence of various cognitive biases
	Barbera et al. (2020)	Effect of confirmation bias
	Coscia and Rossi (2020)	Effect of confirmation bias
	Demartini (2019)	Influence of worker attributes on fact-checking tasks
	Biel and Gatica-Perez (2014)	Presence of bias in impression evaluation
	Kulkarni et al. (2014)	Presence of bias in mutual evaluation
	Eickhoff and de Vries (2013)	Reducing cheat submissions
	Hube et al. (2019)	Making people aware of the presence of biases
	Duan et al. (2020)	Working together with different perspectives
	Faltings et al. (2014)	Influence of monetary incentives on bias
	Gadiraju et al. (2017)	Finding competent workers by the Dunning–Kruger effect
	Gemalmaz and Yin (2021)	Eliminating the influence of confirmation bias
	Echterhoff et al. (2022)	Eliminating the influence of anchoring bias

al. 2013; Xue et al. 2013). Many crowdsourcing tasks are designed such that they can be performed by workers who do not necessarily have expertise. However, the quality of tasks vary greatly depending on the knowledge and dedication of the workers. On commercial crowdsourcing platforms particularly, there are so-called "spam workers" who seek to maximize reward for as little effort and time as possible. To ensure quality in these situations, many studies have been conducted on how to select workers who produce high-quality deliverables and how to evaluate the quality of work (Ipeirotis et al. 2010; Kazai et al. 2011). Furthermore, when considering the trustworthiness of human computation, it is necessary to consider not only the correctness of the computation results, but also a broader perspective. This includes the availability of results, or if there is a flaw in the computation, whether it can be corrected.

Human computation, which is a fundamental framework for humans and AI to cooperate in solving problems, is not able to compare and evaluate different systems from multiple perspectives because there is no standard quality evaluation metric like that for conventional computer systems. For human computation to be trusted and used in the real world as a more practical approach, it must be designed considering quality from various perspectives. In the field of computer engineering, RAS is used to evaluate systems from multiple perspectives to quantitatively and qualitatively assess the possibility of problems arising during system operations (Siewiorek and Swarz 1998). However, previous human computation research has not comprehensively addressed the system trustworthiness from an RAS perspective. It is also unclear whether RAS, which targets systems consisting only of computers, can be directly applied to systems that include humans as components. From this point forward, we will discuss the differences between what should be considered in computer engineering and human computation for each aspect of RAS, and describe existing human computation studies that are relevant to each aspect.

2.1 Reliability

In computer system design, reliability is defined by mean time to failure (MTTF), which is calculated by dividing the system uptime by the number of failures. However, hardware failures are rarely a problem in human computation, whereas human errors are far more frequent and of main concern. Therefore, the error rate of the task execution results would be a more appropriate criterion for reliability.

The research most related to the reliability in RAS is referred to as *quality control* in human computation, and is commonly found in the literature. Statistical quality control, which improves the overall reliability through redundancy introduced by asking multiple workers to perform the same task, is the mainstay of quality control research (Daniel et al. 2018). Furthermore, an inherent problem with human computation is that the people participating are highly heterogeneous in terms of motivation. Therefore, research has also been conducted to design mechanisms that provide incentives for participants to work diligently from a game-theory perspective (Muldoon et al. 2018).

In human computation, even if the results of some tasks are incorrect, they can be resolved and the final result of the entire system's computation can be correct. In the context of system design, this corresponds to a *fault-tolerant* design in which individual component failures or operational defects do not significantly affect the overall process. Fault tolerance is a concept that is addressed in the field of reliability design (Shooman 2002). Qual-

ity control of human computation includes concepts pertaining to both (narrowly defined) quality control to enhance the quality of individual components and reliability design when multiple components are combined to build a system (Lease 2011).

When controlling product quality at a manufacturing site, product samples are inspected for defects. Once defective products are found, the causes of them at the manufacturing site are investigated. For example, if defective products are concentrated in a particular production line, the line is stopped to investigate the cause. In crowdsourcing, which is a platform for human computation, people are the main cause task errors, and the work lines at the manufacturing site can be considered to correspond to crowdsourced workers. (Of course, there can also be causes other than workers, such as unclear task descriptions or faulty work interfaces.) Therefore, quality modeling which focuses on workers has been widely used for crowdsourcing quality control (Ipeirotis et al. 2010). In the following subsections, we discuss existing research in terms of worker ability, cognitive bias, relationship between tasks and workers, and types of tasks. Table 1 summarizes related studies based on the reliability of human computation.

2.1.1 Considering worker abilities

The most important factors in obtaining high-quality results is assigning tasks to workers with high ability. (Hereafter, we use the term "ability" not only in the narrow sense of worker knowledge and skills, but also in the broader sense that includes other factors that affect the accuracy of work results, such as worker morale.) In many crowdsourcing markets, workers are not paid until their work is reviewed and approved by the requester. If the requester is not satisfied with the results of the work, he/she may refuse to pay. The percentage of approved results submitted by a worker is an important indicator of the worker's ability to perform tasks. In many crowdsourcing markets, filtering functions are available, such as requesting workers with an approval rate of 95% (Amazon Web Services 2017).

The task approval rate can be thought of as the evaluation of the worker by requesters. The idea of computational trust and reputation (Sabater and Sierra 2005; Braga et al. 2018) is also useful in considering long-term trust between human computation systems and humans. Many crowdsourcing platforms have implemented mutual rating items between task requesters and workers, and these have made a certain contribution to quality assurance of tasks and outcomes. There is also research on quality assurance by mutual rating among workers (Whiting et al. 2017).

However, these approaches cannot be applied to new workers with a limited work history. In addition, if there is little relationship between previous and current tasks (e.g., a task to answer a question in English and a task to answer a question in Chinese), the past evaluation may not be directly applicable to the estimation of ability to perform in the current task.

If the correct answers to some of the tasks are known, they can be used to filter workers by evaluating their responses. Alternatively, worker abilities can often be measured by performing spot checks in which tasks with known answers are mixed with those with unknown answers (Kazai et al. 2011). If a worker's ability is determined to be low, measures can be taken such as stopping further work or discarding results obtained from the worker.

Even if a worker performs a task by guesswork, the accuracy is not zero because there is a chance that he/she will obtain a correct answer by fluke. Therefore, workers whose answer rate is close to zero (i.e., workers who almost always answer the opposite of the correct

answer) can be considered malicious workers or workers who misunderstand the question and provide incorrect answers. For example, in the task of classifying Web sites into "porn" and "not porn," some workers mistakenly checked "porn" when they were instructed to check "not porn." However, if such workers can be identified using data with known correct answers, other correct answers can be estimated with high accuracy by replacing their answers in the opposite manner (Ipeirotis et al. 2010).

On the other hand, workers who answer randomly may be so-called spam workers who answer without looking at the question because they only want money from the beginning, workers who answer by a guess without sufficient ability to answer, or bots or other programs other than humans. These workers answer randomly, independent of the question; thus, the probability of a worker answering the question does not depend on the correct answer. A spammer score that identifies such workers has been proposed (Raykar and Yu 2011).

The simplest way to provide robustness against errors in the results of worker tasks is to introduce redundancy by requesting the same task to multiple workers, rather than requesting it to only one worker, based on the idea that "two heads are better than one." For example, in the case of a binary classification problem, if three people are requested to perform a task and a majority vote is taken, the correct answer can be obtained even if one person makes a mistake. It is important to note that the assumption for majority voting is that the percentage of correct answers by workers is high to some degree (Sheng et al. 2008).

As described above, simple majority voting, in which a majority is determined by one person with one vote, can be regarded as assuming that all workers have equal ability (i.e., equal chance of obtaining the correct answers). However, worker abilities often vary with crowdsourcing. In such cases, it is a bad idea to treat the opinions of workers with a high and low probability of obtaining the correct answers equally when taking a majority vote. If the worker's accuracy is known from a pre-test using a task for which the correct answer is known, the possibility that the correct answer can be derived increases by assigning more weight to the opinion of the worker with higher accuracy. In fact, it has been shown that using such weighted majority voting in multiple natural language processing tasks can estimate the correct answer with higher accuracy than simple majority voting (Snow et al. 2008).

For tasks that predict the future, the correctness/incorrectness of the worker's answers becomes known in the future. In such tasks, a method to select high-quality workers has been proposed by preparing two types of reward plans: a high-risk, high-return reward plan in which the difference in reward between correct and incorrect answers is large, and a low-risk, low-return reward plan in which there is little difference between correct and incorrect answers (Sakurai et al. 2013).

In the weighted majority voting method based on worker ability, the parameters representing worker ability are estimated by having workers carry out work on tasks for which the correct answers are known in advance. In practice, there are many cases in which it is not possible to prepare enough tasks with correct answers, or the worker has just started working. Therefore, it is desirable to have a method that simultaneously estimates the worker's ability and the correct answer from the results of their work on tasks for which the correct answer is not known. One method uses a latent class model that alternately estimates the correct answer and the worker's ability parameters (Dawid and Skene 1979). In fact, the latent class model was proposed to derive a more reliable diagnosis from the results of

medical diagnoses by multiple doctors long before the advent of crowdsourcing, and only recently has been used in crowdsourcing research to derive correct answers from multiple worker responses. The EM (Expectation Maximization) method, which is commonly used for parameter estimation in models with hidden variables (correct answers), was used in this study. Additionally, several Bayesian extensions of latent-class models have been proposed to address a small number of worker labels and complex relations among workers. For example, Community BCC (Venanzi et al. 2014) considers group structures within workers, and Enhanced BCC (Li et al. 2019b) models the correlation between workers.

2.1.2 Considering the relationship between tasks and workers

In the latent-class model introduced in Sect. 2.1.1, worker ability was estimated by assuming that the difficulty of the problem was constant. However, to assess worker ability accurately, it is necessary to consider the difficulty of a task. The item response theory used in the test design (Linden and Hambleton 1997) can be used to estimate worker ability using a task in which the correct answer is known. In crowdsourcing, Whitehill et al. (2009) proposed a method that can simultaneously determine the worker's ability and the correct answer while considering the difficulty of the task, even when the correct answer is not given. While in the latent class model (Dawid and Skene 1979), the only parameters to be estimated were those related to the worker's ability, this model also explicitly introduces parameters related to the difficulty of tasks to be estimated. This makes it possible to fairly evaluate the abilities of workers who worked on tasks of different difficulty levels. Bachrach et al. (2012) proposed an Bayesian extension called joint difficulty-ability-response estimation (DARE) model.

The models described thus far assumed that workers could be represented by a one-dimensional ability parameter and tasks by a one-dimensional difficulty parameter. However, human ability is multifaceted and includes various factors, such as language ability, computational ability, and memory. On the other hand, tasks also differ in terms of the abilities required to perform them, with some tasks requiring a high level of language skills but not computational skills, and other tasks requiring a high level of computational skills but not language skills. This leads to differences in worker-task "compatibility," such that a task that is easy for one worker (high language ability but low computational ability) may be difficult for another worker (low language ability but high computational ability). To incorporate such situations into the model, it is necessary to represent the worker and task parameters as multi-dimensional feature vectors rather than one-dimensional parameter. Welinder et al. (2010) proposed a method for estimating worker and task feature vectors from worker responses, using a model in which worker responses are determined probabilistically given the above worker and task feature vectors.

Another approach is to directly ask the worker about the difficulty of the task or the worker's confidence in his/her answer, rather than using computationally estimated task difficulty. This approach may be more in line with the idea of human computation in that it makes active use of human meta-cognitive abilities. Although there is a correlation between confidence and the rate of correct answers, there are many overconfident workers whose confidence is higher than the rate of correct answers, and many underconfident workers whose confidence is lower than the rate of correct answers. The accuracy of confidence judgments differs from person to person, and it is necessary to consider differences in the accuracy of workers' self-assessments, rather than treating the confidence level given by

each worker uniformly. Therefore, by extending the latent-class model and introducing a parameter indicating the accuracy of a worker's self-reported confidence, a model that can derive the correct answer with good accuracy has also been proposed (Oyama et al. 2013).

To obtain quality work results, it is important to ensure that the task is performed by a worker who has the necessary competence for the task. Task recommendation formulates task assignments as a recommendation problem: Ambati et al. (2011) predicts preferred tasks based on worker attributes and other information through supervised learning; Yuen et al. (2012) recommends tasks using task browsing history and past tasks; Kulkarni et al. (2012b) also introduced a mechanism for other workers to recommend tasks. Li et al. (2014) used a model that predicted task accuracy based on worker attributes such as nationality, education level, gender, major, and personality test scores to identify appropriate workers. In addition, there have been reports of application accuracy drops in tasks dealing with language, even for the same English task, when country and speaker attributes differ between the data-collection task and the resulting application domain. Bender and Friedman (2018) suggested utilizing data statements that include speaker and annotator demographics for data dealing with language.

The methods described so far assume that batch processing is performed after all workers have completed their work on all tasks, and that the degree of redundancy, i.e. the number of workers requested to perform the same task, for each task is predetermined. However, because task execution involves both human and financial costs, it is desirable to minimize redundancy if the quality can be guaranteed. For this purpose, instead of allocating the same redundancy to all tasks, it is necessary to distinguish between tasks with different levels of answer uncertainty. For tasks in which the correct answer is mostly clear from the worker's response, no further requests are made to the worker. Contrary, for tasks in which the correct answer is uncertain, the worker's response is continuously collected until the correct answer is assured. To this end, Sheng et al. (2008) proposed a method of Bayesian estimation of the probability that the answer decided by the majority vote is incorrect, which is then used to calculate the uncertainty of the answer. Sheng et al. (2008) assumed that the abilities of all workers are equal. However, workers have different abilities, and it is possible to improve the efficiency of quality control by considering these differences when assigning tasks. As previously mentioned, worker ability can be evaluated using tasks in which the correct answers are known. However, if the number of tasks performed by a worker is small, the worker's ability remains uncertain. For example, we cannot assume that a worker's ability is low only because of his/her first few failures.

In the problem of selecting a task to collect responses, workers were preferentially assigned to tasks with high uncertainty levels. On the other hand, in the problem of selecting which worker to request, assigning a task to a worker with high uncertainty in his/her ability is not a good method. Rather, if a worker is known to have high ability, it is more efficient to continuously assign tasks to that worker. However, if we assign tasks to the same workers only, we may omit a high-capable worker who has not yet been assigned many tasks. (Conversely, there is no need to assign tasks only to those who are certain to have low ability.) This is the so-called trade-off between exploration and exploitation and appears in various decision-making problems when dealing with uncertainty. To address such problems, a method called *interval estimation* (Kaelbling 1990) was introduced for crowdsourcing and used for worker selection (Donmez et al. 2009).

2.1.3 Quality control for other types of tasks

There are many forms of human computation output, not just the simple labeling of data. Ranking is the ordering of given items and is used in a variety of tasks, such as evaluating search engine results. Chen et al. (2013) proposed a method to determine the overall ranking while considering the work quality when pairwise comparisons between two items were provided during crowdsourcing. Furthermore, they generalized traditional active learning and proposed a method for selecting pairs to be assigned to workers while considering worker quality, uncertainty in the order of pairs, and uncertainty in the model. This study determined a single ranking for all items. In contrast, Matsui et al. (2014) proposed a method for simultaneously determining multiple rankings from worker ordering data for three or more items in a task, such as sorting words in a sentence.

The quality control studies described thus far have dealt with cases in which the worker produces structured outputs, such as classification and ranking. In human computation, many tasks such as writing articles also deal with unstructured outputs. For example, Wu et al. (2012) proposed an aggregation method of sequence labels that assumed a sequential noise model. Wang and Dang (2022) proposed a sentence integration model based on a language generation model using a transformer. On the other hand, Baba and Kashima (2013) proposed a general quality control model consisting of two stages: producing unstructured outputs and evaluating them. They introduced a probabilistic generative model for quality control that considered the author's ability and evaluator bias.

2.1.4 Considering the cognitive bias of workers

Efforts have been made to improve the quality of human computation by focusing on the characteristics of human information processing. Cognitive bias is a decision-making bias that occurs systematically because of the tendencies and limitations of human cognitive functions. Whereas the abilities discussed in the previous sections refer to variation among individuals within the human species, cognitive biases are tendencies that are common to the human species, which is to say variation among different species. Naturally, the effects of cognitive bias are also expected to appear in human-driven computations. Indeed, in crowdsourcing, which is often a platform for human computation, human cognitive bias is known to have a significant impact on the results (Draws et al. 2021). From a technical perspective, ability is often treated in a way that includes variability of crowd workers in a model, while cognitive biases are modeled as features common to all or a specific group of workers.

Eickhoff (2018) demonstrated the existence of various cognitive biases (ambiguity effect, anchoring, bandwagon effect, and decoy effect) in crowdsourcing tasks through experiments. For example, Eickhoff (2018) confirms the existence of an ambiguity effect in the document relevance evaluation task; when information presented is reduced, even if it is essentially irrelevant to the evaluation, it is perceived as more ambiguous and therefore rated lower. Newell and Ruths (2016) also showed that in a repetitive image-labeling task, the crowd worker's previous task influences the results for the current task. This can be considered as a type of anchoring bias, which is the tendency for judgments about information presented later to be influenced by previously presented information. Confirmation bias is a type of cognitive bias in which workers attach importance to information that supports their

hypotheses and beliefs. In crowdsourcing, workers' personal beliefs and backgrounds have also been observed to influence the outcome of their judgments (Barbera et al. 2020; Coscia and Rossi 2020), which is also confirmed in fact-checking tasks by Demartini (2019). Leniency and Halo errors (Balzer and Sulsky 1992) arise from the difficulty for raters to independently verify multiple rating axes. The presence of these biases in impression (Biel and Gatica-Perez 2014) and mutual ratings (Kulkarni et al. 2014) has also been noted in crowdsourced annotations.

The above studies primarily suggest the existence of negative effects of cognitive biases on task results and the need to remove these biases. Typical ways to mitigate these biases involve ingenious task design and interventions in the annotation process. Eickhoff and de Vries (2013) explored various factors for reducing cheat submissions in crowdsourcing tasks. They reported that rewriting the task to avoid repetition or reducing the batch size alone reduced the rate of cheating. Faltings et al. (2014) showed that monetary incentives affect worker bias and proposed a bonus-adding scheme based on game theory control the biases. Hube et al. (2019) instructs participants to consider social opinions in subjective-judgement tasks and to instruct people not to bring their own potential biases into the task. By making people aware of possible biases, they are effectively mitigated. On the other hand, Duan et al. (2020) also reported that workers with different perspectives did not necessarily reduce bias when they worked together.

There have been several attempts to explicitly incorporate cognitive biases into quality control methods for individual worker responses. Gadiraju et al. (2017) used a pre-screening task that required workers to submit self-assessments as well as responses to a task, aiming to identify competent workers based on the Dunning–Kruger effect, in which people with low capabilities overestimate themselves. Gemalmaz and Yin (2021) proposed a method that introduced the degree to which each worker is affected by a confirmation bias in a label-integration model, which allowed for integration that eliminated the influence of bias. Echterhoff et al. (2022) proposed a method to remove anchoring bias, whereby a decision on the same item will differ depending on one's past decisions, by detecting it in the data or by controlling the order in which the items are presented.

2.2 Availability

In a traditional RAS, availability is calculated using the system uptime relative to the total time. Because human labor is the bottleneck in increasing availability in human computation, various studies on worker availability have been conducted. In the following subsections, we describe existing research on monetary and non-monetary incentives, worker behavior prediction, and crowdsourcing contests to achieve availability. Table 2 summarizes the related studies on the availability of human computation.

2.2.1 Monetary incentives

One way to increase the availability of the workforce is to increase the reward, which is known to increase the quantity of work, although this does not necessarily improve the quality of the work (Mason and Watts 2009). Setting appropriate rewards is critical to ensuring worker availability. If the reward is too low, it will be difficult to attract enough workers, whereas if the reward is too high, it will be less profitable for the client. Miao et al. (2022)

Table 2 Studies on availability in human computation

Category	References	Description
Monetary incentives	Mason and Watts (2009)	Increasing monetary incentives
	Miao et al. (2022)	Dynamic pricing of tasks
	Bigham et al. (2010)	Recruiting workers for real-time service
	Bernstein et al. (2011)	Recruiting workers for real-time service
	Bernstein et al. (2012)	Estimating waiting time by queueing theory
	Oka et al. (2014)	Giving incentive to work during declared period
	Bacon et al. (2012)	Having workers declare their effort and completion time
	d'Eon et al. (2019)	Rewarding collaborative work
Non-monetary incentives	von Ahn and Dabbish (2008)	Introducing games with a purpose
	Ipeirotis and Gabrilovich (2014)	Recruiting volunteer workers through advertisements
	Dai et al. (2015)	Providing occasional entertainments for workers
Worker behavior prediction	Cheng et al. (2019)	Worker availability prediction using machine learning
	Mao et al. (2013)	Predicting volunteer worker participation
Crowdsourcing contests	DiPalantino and Vojnovic (2009)	Relationship between rewards and participation in contests
	Truong et al. (2022)	Selecting the optimal incentive design
	Feyisetan and Simperl (2019)	Introducing contests to microtask crowdsourcing

used a deep time series model to predict the relationship between bonus and task quality and proposed a dynamic pricing mechanism for tasks by introducing bonuses.

Labor availability is particularly important for crowdsourcing real-time services. VizWiz (Bigham et al. 2010) hires workers before they are needed and keeps them on standby to solve past tasks, enabling the service to answer questions about images from blind people in real time. Similarly, the Retainer Model (Bernstein et al. 2011) pays workers a small fee to wait for a task, and when it becomes available, the worker can start working immediately. Furthermore, Bernstein et al. (2012) used queueing theory to analyze requester waiting times and costs in the Retainer Model.

Participatory (Burke et al. 2006) and crowd sensing (Ra et al. 2012) asks users with mobile terminals or other devices to collect data such as images, sound, and location information from real world environments, which are often used in citizen science. The difference between conventional distributed sensing and crowdsensing is that the sensors are owned by ordinary users and require their involvement in data collection. The key challenge in crowdsourcing is obtaining participants to contribute to data collection. Therefore, research is being conducted to provide incentives for participants to work according to their declared schedules (Oka et al. 2014).

For long-running tasks such as software development, it is important to know the task's completion time. The problem here is that the workers themselves, who predict the completion time, can influence this time by adjusting the amount of effort. Bacon et al. (2012) proposed an incentive mechanism to elicit maximum effort, while allowing workers to report their expected time of completion with honesty.

In conventional crowdsourcing, workers receive payment for work completed on an individual basis. However, some tasks may require the cooperation of multiple workers. When paying for group work, paying all members the same remuneration may cause higher performing workers to withhold their efforts. d'Eon et al. (2019) compares rewards proportional to outcomes based on equity theory and Shapley values in cooperation game theory with uniform rewards. The results suggest that pay-for-performance rewards increase the effort extended by workers to a small extent.

2.2.2 Non-monetary incentives

Gamification or games with a purpose (von Ahn and Dabbish 2008) are representative methods for encouraging participation with non-monetary incentives. Games with a purpose have known templates called output-agreement, inversion-problem, and input-agreement, which can be used to transform a task into a game with a desired outcome. Furthermore, the authors stated that timed responses, score keeping, player skill level, high-score lists, and randomness are important factors in making a game interesting.

When collecting specialized knowledge in areas such as the medical field, it is necessary to recruit workers who are knowledgeable in these areas. Ipeirotis and Gabrilovich (2014) recruits the participation of workers who are not paid for their labor by using an advertising platform. By not paying rewards, they are able to target only those workers who are knowledgeable about the task, and attract only those who have the internal motivation to perform the task. Additionally, advertising campaigns can be optimized by providing feedback on worker conversions to the advertising platform. Furthermore, they found that providing feedback, such as the correct answers to tasks and performance of other participants led to continued participation.

Unlike machines, humans do not like performing monotonous tasks for extended periods of time. Dai et al. (2015) proposed a method to keep workers engaged by occasionally providing them with entertainment, such as games and cartoons, as micro-diversions. Experimental results on several types of tasks have shown that introducing entertainment improves worker retention and response speed while maintaining quality. Furthermore, it is suggested that micro-diversion is more effective in complex cognitive tasks than in reflective tasks.

2.2.3 Worker behavior prediction

Research has also been conducted to predict the availability of workers using machine learning. For example, the crowdsourcing framework FROG (Cheng et al. 2019) uses Kernel Density Estimation to predict future availability based on a worker's past work history. It also introduces smoothing based on information from social networks among workers to solve the cold-start problem for new participants with no work history.

In volunteer-based crowdsourcing, there is no financial incentive therefore, it is important to predict worker disengagement and provide appropriate intervention to workers who

are likely to disengage. Mao et al. (2013) developed a model for predicting worker disengagement through supervised learning, based on task and worker characteristics, as well as worker activity data during task sessions.

2.2.4 Crowdsourcing contests

In crowdsourcing contests, the number of participants has a significant impact on the final output. There are usually multiple contests running simultaneously, and they compete for participation. DiPalantino and Vojnovic (2009) model contests as all-pay auctions, where all bidders must pay their bids regardless of whether they win, and the relationship between prize money and the number of participants is analyzed. The results show that the number of participants is related to the number of prize winners. They theoretically show that the number of participants is logarithmically proportional to the amount of prize money, which is consistent with actual data when the target is limited to users who repeatedly use the crowdsourcing site.

Prize distribution in contests has been analyzed using game theory (Moldovanu and Sela 2001; Archak and Sundararajan 2009) to determine whether prize money should be paid to the highest ranking participant only or to multiple participants. These theoretical studies show that the optimal prize distribution depends on the risk preferences of participants and the form of the cost function. Therefore, Truong et al. (2022) formulated the incentive-selection problem of choosing the optimal design among multiple incentive designs as a multi-armed bandit problem and proposed an efficient algorithm to solve it.

In conventional crowdsourcing of microtasks, workers are paid a fixed amount upon completion of the task. Contest elements have also been introduced into microtask crowdsourcing to accelerate the time required to complete tasks. Feyisetan and Simperl (2019) introduced a leaderboard based on the number and quality of tasks completed in microtask crowdsourcing to allow workers to compete. They showed that paying only the highest-ranked workers increased the speed of task completion, but decreased the quality of some types of tasks. In contrast, they also showed that increasing the number of workers rewarded can increase the number of completed tasks.

2.3 Serviceability

Serviceability indicates the ease of recovery when a system fails, and is evaluated by an index such as MTTR (Mean Time To Repair). To the best of our knowledge, no research has directly and quantitatively evaluated serviceability in human computation. In HC, where many workers are involved, human-caused computation failures can occur at any time, and to improve the serviceability of HC, it is necessary to facilitate the management of complex workflows that represent the entire computation and allow for immediate recovery from failures. Of particular importance is the introduction of knowledge developed in software engineering, which abstracts programming details and improves the ability to maintain complex systems. These include techniques such as hierarchical decomposition of complex programs into simpler modules, reuse of modules, and automatic synthesis of programs, as well as programming languages to support their implementation. Table 3 summarizes related studies on serviceability in human computation.

Table 3 Studies on serviceability in human computation

Category	References	Description
Workflow management	Kittur et al. (2011)	Decomposing complex tasks into simple tasks
	Dai et al. (2010)	Automated control of crowdsourcing workflows
	Zhang et al. (2013)	Synthesizing optimal workflows
	Kulkarni et al. (2012a)	Asking workers to decompose and integrate tasks
	Kittur et al. (2012)	Visual management of complex tasks
Programming language	Little et al. (2010a)	Imperative programming of fault-tolerant human computation
	Minder and Bernstein (2012)	Programming using abstract patterns of human computation
	Park et al. (2012)	Writing data retrieval queries using an SQL-like language
	Morishima et al. (2012)	Declarative programming using a Prolog-like language
	Tranquillini et al. (2015)	Visual programming using a business process modeling language

2.3.1 Workflow management

In human computation, when complex tasks are performed, their workflows become more complex, therefore it is necessary to abstract them and make them easier to manage. CrowdForge (Kittur et al. 2011) automates the flow of decomposing complex tasks into simple tasks by abstracting them with a distributed computing framework similar to MapReduce. Conventional crowdsourcing workflows are complex and repetitive, making it difficult to maintain optimal workflow. TurKontrol (Dai et al. 2010) used a mathematical model for decision making to automate the control of crowdsourcing workflows. In the field of software engineering, there is a process called program synthesis, which generates a program from a problem specification; Zhang et al. (2013) introduced this idea to crowdsourcing workflows. They proposed a method for synthesizing optimal workflows based on inference arising in the workflow structure and learning about worker performance.

Turkomatic (Kulkarni et al. 2012a) allows workers to participate in the design and execution of workflows such as task division and integration, and visualizes the process so that requesters can supervise it. CrowdWeaver (Kittur et al. 2012) is a system for visually managing complex workflows. It allows requesters to reuse existing tasks to create workflows, manage data flow between tasks and task progress, and modify workflows in real time.

2.3.2 Programming languages

There have been studies on programming language-level support for describing complex human-computation workflows. Turkit (Little et al. 2010a) introduced a crash-and-return programming model that allows imperative programming to describe fault-tolerant human computation. CrowdLang (Minder and Bernstein 2012) is a programming language for

designing and implementing human computation, which can incorporate patterns that appear in human computation through abstraction, such as group decision making. One type of abstract programming language is a declarative language termed SQL, which is a standard declarative data-retrieval language. That is, the user only describes what he/she wants to retrieve from the database, and the steps for retrieval are left to the optimization process and are hidden from the user. Deco (Park et al. 2012) is a system that allows data retrieval using crowdsourcing to be written in SQL. CyLog (Morishima et al. 2012), a logic programming language similar to Prolog, is a declarative language for describing crowdsourcing. BPMN4Crowd (Tranquillini et al. 2015) is a business process model and notation (BPMN)-based modeling language. It allows crowdsourcing workflows to be intuitively described and integrated into business processes by means of diagrams using a visual editor.

3 Social trust of human computation

In this chapter, we discuss the opposing perspective that expressed in the previous chapter, that is, the trustworthiness that a human computation system instills in its users, society, and participants (crowd workers in many cases) in human computation. Similar to the previous discussion on the trustworthiness of AI (Kaur et al. 2022), human computation systems must be operated ethically to earn the trust of society. The processes and deliverables of human computation systems should also be ethical and explicable, such that general public can understand and accept them.

Human computation systems must handle data generated by workers in a scientifically and socially appropriate manner. When a system learns from the crowd-generated data and makes decisions based on them, its behavior is expected to follow the AI ethics guidelines. In terms of security, an AI system should ensure that socially undesirable, malicious, or privacy-invasive tasks are not performed. Moreover, it is necessary to consider ethical requirements such as fairness, accountability, and transparency of the outcomes of the system. For human computation systems, data biases, such as sampling and annotation biases of annotators, have a significant impact on the fairness of their outcomes. The transparency of an AI system refers to the need to explain, interpret,and reproduce its decisions (Dignum 2017). Since reproducibility of the system is ensured by by the reliability and serviceability described in Sect. 2, we focus on explainability for transparency. The explanation of a human computation system should be reasonable and understandable for non-experts. Along with the development of explanation methods, it is necessary to evaluate the understandability of explanations. Accountability is the need for a system to explain and justify its actions and decisions to various stakeholders.

3.1 Accountability of human computation systems

Accountability is the need for a system to explain and justify its actions and decisions to its various stakeholders. Providing accountability to human computation systems requires transparency of the system and the implementation of appropriate governance. This subsection organizes accountability in human computation systems from two perspectives: accountability to the users and society, who benefit from the human computation system,

and accountability to the crowd workers, who participate in and serve as the driving force behind the system. These two parties correspond to the elements on the right side of Fig. 1.

A discussion of accountability in AI in general (Kaur et al. 2022) provides more details on how to make human computation systems decision-making accountable. The methods include methods that can be incorporated into the algorithm design process, methods to ensure transparency, and strict laws and policies for better governance of algorithms.

For accountable system development, various stakeholders may be involved in the design process. Furthermore, algorithms that are secure and fair are required. Here, safety includes cyber security as well as privacy protection of data providers. Security and fairness are discussed in more detail later in this chapter.

Model impact assessment also improves accountability. For example, available evaluation methods such as privacy impact assessment (PIA) and ethical impact assessment (EIA). When conducting such an impact assessment, the responsible parties and data handling flows will be clarified and responsibilities will be clearly defined. The need for multi-layered explanations has also been pointed out, and accountability is expected to be improved by conducting multiple such assessments.

Monitoring after the model has been deployed also contributes to better governance. The development of laws and other mechanisms for monitoring (Kroll et al. 2017) is necessary. In addition, it is needed to select indicators for evaluation of the system and evaluate them using algorithms. Proper selection of indicators is important here. For example, there are various indicators for fairness (Narayanan 2018), and it is necessary to use appropriate ones according to the context.

Although the above mentioned various approaches for accountability of the decision-making process in human computation systems have been raised, it is not easy to clarify where the responsibility lies in practice. Continuous review is important for accountable human computation systems.

The fair treatment of human computation system participants is also important for system accountability. Crowd workers must be treated fairly and respectfully. Crowd workers often conduct subjective tasks such as data annotation and questionnaire responses. Therefore, it is necessary to protect workers from possible risk during task execution. Therefore, it is important that the human computation processes follow appropriate standards and guidelines for participants treatment.

In human computation, tasks are often based on subjective ratings or sensory evaluations conducted by humans. These should meet the same standards as the ethics of research and experiments in the behavioral sciences and psychology, which involve questionnaires and other surveys of human subjects. The Belmont report (National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research 1978) summarizes ethical principles and guidelines for research involving human subjects. The report identified three core principles: respect for persons, beneficence, and justice. In applying these principles, we need to consider informed consent, the assessment of risks and benefits, and selection of subjects. All relevant information should be provided in an understandable and accessible manner. In particular, possible risks in task execution, such as psychological invasion, must be noted. It is also required that workers should not be disadvantaged in the task assignments. All relevant studies should be reviewed and approved by the Institutional Review Board (IRB).

As human computation is invariably conducted in an online environment, we should also refer to the guidelines for participant protection during online research. Several guidelines has been proposed for participant protection in online research (Frankel and Siang 1999; Bruckman 2002). To protect privacy in online interactions, the requester must protect workers' private information such as email addresses. When obtaining consent, they must be careful to provide information in a manner that the participants can comprehend. Although the online work environment has changed significantly since these guidelines were proposed, and crowd workers have become more proficient at online work, the aforementioned points should still be considered.

If a task contains deception, intentional misleading of subjects, workers must be informed. User studies related to privacy and security such as phishing unavoidably contain deception (Jagatic et al. 2007). While research involving deception can be of great benefit to academia and product development, additional care must be taken with crowd workers. The requesters must design research protocols that minimize risk. They must also conduct debriefing sessions for workers, and the deception in the task and its necessity must be explained in an understandable manner (Finn and Jakobsson 2007).

It is also necessary to carry out proper labor management of participants and disclose information about it. The crowdsourcing work environment must be considered when implementing human computation on a crowdsourcing platform. It has been noted that, while the crowdsourcing marketplace allows workers to work without geographic or time constraints, labor exploitation can occur due to differences in position and economic disparity (Silberman et al. 2010). One of the concerns raised on low compensation is that crowdsourcing of translation tasks may upset the balance of the market by competing with professional translators (Dolmaya 2011). Research communities currently suggest that rewards should at least be the minimum wage (Silberman et al. 2018), and research papers should clearly state how workers are compensated. For example, Prolific, a research-purpose crowdsourcing platform, has a minimum hourly wage that encourages appropriate payments. On the other hand, a survey of workers revealed that, they are often not trained in labor management and sometimes work on several jobs at simultaneously (Kaplan et al. 2018), while labor management on crowdsourcing still has many problems. Human computation systems should ensure fair evaluation of workers and requesters. Mutual evaluation between clients and workers is a prominent feature of crowdsourcing. For workers to receive appropriate compensation and work opportunities, appropriate evaluation of deliverables must be considered. In addition, proper evaluation of the requesters is also important, as they may be unfairly undervalued, resulting in loss of task assignment opportunities. It is also proposed to provide a mechanism for crowdsourcing workers to disclose and evaluate their interactions with human computation systems, thereby making the trust of workers visible (Irani and Silberman 2013).

3.2 Privacy and security in human computation

The privacy of workers and tasks must be properly protected. Crowd workers' private information such as names and e-mail addresses must be secured. In addition, there is a risk of sensitive information being inferred from task execution. For example, in a mobile crowdsourcing which uses mobile devices and sensing to collect data, the workers' location can be inferred from task outcomes and the information exchanged during task execution. Feng

et al. (2018) identified privacy and security threats in mobile crowdsourcing and reviewed existing countermeasures.

Worker privacy protection by using privacy-preserving data mining techniques has also been studied. Kajino et al. (2014a) proposed a secure computation protocol for crowd label aggregation, in which labels provided by crowd workers and their ability parameters are kept secret, but the client can obtain aggregated labels. The protocol uses homomorphic encryption schemes to perform the secure computation to estimate the true aggregated label from encrypted labels. Similarly, secret computation techniques were applied for task-worker matching (Shu et al. 2018b) and for task recommendation (Shu et al. 2018a), while protecting worker privacy.

Task instances may also contain private information. For example, in medical tasks, personal information can be inferred from images and texts of medical records. Task division is an approach for coping with such risks. For text privacy, Little and Sun (2011) proposed a method that divides medical records and presents them to different workers during transcription tasks. To protect image privacy, Kajino et al. (2014b) investigated the properties of the instance clipping protocol that clips the images by a window of a fixed size.

The human computation systems must avoid inappropriate tasks to ensure security of human computation. Examples of inappropriate tasks include the unauthorized acquisition of personal information and cooperation in stealth marketing and slander. System administrators must constantly monitor and remove inappropriate tasks. From the perspective of ensuring the security of human computation, the systems have to avoid inappropriate tasks. An automated monitoring system using machine learning was proposed to improve the efficiency of such monitoring (Baba et al. 2013).

3.3 Fairness and biases in human computation

Accounting for the fairness of AI systems has recently become increasingly important. AI systems sometimes exhibit discriminatory behavior toward certain groups or populations in high-stakes decisions, such as hiring and personnel evaluations. Examples of cases that drew criticisms include a hiring AI systems discriminating against women (Dastin 2022) and a facial recognition system that reported a difference in accuracy between groups of different gender and race (Buolamwini and Gebru 2018). Furthermore, a chat-bot reflected hate speeches contained in its training data (Neff 2016).

As most AI systems and algorithms are data-driven, they reflect biases in the training data (Barocas et al. 2019). For example, if the data has a sampling bias, the characteristics of a minority in the data are less reflected in the result. The labels or annotations may have biases that discriminate against certain groups or populations. Models learned from biased labels prone to reflect biases in their predictions. For natural language processing model, Shah et al. (2020) suggested general mathematical definitions of predictive biases. They differentiated four potential origins of biases: label bias, selection bias, model over–amplification, and semantic bias. Here we focus on workers' annotation or label bias in the human computation system. Table 4 summarizes the studies related to the annotation biases.

Crowd workers, who are often participants in decision-making as well as providers of data and inputs in human computation, tend to be biased in subjective tasks, depending their personal preferences or inclinations. Several studies have shown systematic annotation biases. Sen et al. (2015) showed that different communities create different gold standards

Table 4 Studies on annotation bias in human computation

Category	References	Description
Systematic annotation biases	Sen et al. (2015)	Difference in semantic relatedness judgments between communities
	Dong et al. (2012)	Difference in image tagging tasks between countries
	Otterbacher et al. (2018)	Gender bias of sexists in search result evaluation
Modeling worker biases	Liu et al. (2022)	Modeling annotator group bias in label aggregation
	Davani et al. (2022)	Modeling each annotator using multi-task learning
Countermeasures for biases	Barbosa and Chen (2019)	Task allocation considering the human factors of workers
	Ueda et al. (2022)	Enhancement of the minority labels in label aggregation
Multi-modality in annotations	Gordon et al. (2021)	Classification using disagreement-adjusted metric
	Gordon et al. (2022)	Label aggregation for each subgroup

in semantic relatedness judgments. Dong et al. (2012) found that European-Americans and Chinese crowd workers may provide different tags in image tagging tasks. Otterbacher et al. (2018) also reported that sexist workers were less likely to report gender bias in evaluating image search results.

Methods for modeling and estimating worker bias have also been proposed. Liu et al. (2022) developed a framework to capture annotator group bias. The framework extends the probabilistic graphical model for label aggregation. It uses the probabilistic graphical model with additional variables representing annotator group bias. Davani et al. (2022) investigated a multi-task learning approach that treats the prediction of each annotator' judgments as separate subtasks, while sharing a common learned representation of the subtasks.

The presence of biases has a significant impact on the outputs of the AI systems trained on them, which can be a major social risk, for example, discrimination against certain groups or populations. In order to deal with such biases, there is an attempt to assign workers with an appropriate demographic composition for given tasks. Barbosa and Chen (2019) proposed a framework that allocates tasks considering the human factors of workers. The frameworks translate the task assignment process into a multi-objective optimization problem. By routing tasks to workers based on demographics, it mitigates biases in the worker sampling. On the other hand, Ueda et al. (2022) viewed this as a correction problem of sampling biases in worker recruitment, and enhanced the minority in the data in the results of label aggregation.

Certain subgroups of crowd workers may be biased in subjective tasks such as detecting emotion, aggression, and hate speech, which often reflects their preferences or inclinations. Such biases may result in annotation disagreements in label aggregation. However, such disagreements do not necessarily indicate low-quality annotation. Alm (2011) identified potential challenges for subjective tasks including the absence of the grand truth. Alm (2011) also discussed the implications. Evaluation techniques deserve careful thought if the ground truth needs to be reassessed. Inter-annotator agreement schemes may not be appropriate and divergence and variation in annotation may provide a useful understanding of the task.

The presence of subgroups with multiple systematic differences can lead to multi-modality in their opinions. Gordon et al. (2021) proposed an approach that separates bias from noise and uncertainty, rather than aggregating annotation labels into a single ground truth label in situations where there is multi-modality in the opinions. Although there is currently no standardized method for determining which subgroups' opinions are more important than the others, Gordon et al. (2022) also proposes "jury learning" that presents aggregated labels for each subgroup defined by their attributes.

Note that fairness is sometimes a trade-off with privacy. It is often necessary to use the worker's attributes to deal with biases in human computation systems. However, workers' demographic or profiled information is sometimes privacy-sensitive. For example, the GDPR states that when profiling based on personal data, it is necessary to inform the data provider of the purpose and use of the data (Voigt and Von dem Bussche 2017). Appropriate notices must be provided to crowd workers. Also, privacy-sensitive information must be protected in the human computation systems.

3.4 Explainability of human computation systems

To understand the system's performance and limitations, both ex-ante and ex-post explanations are essential for the transparency of the system (Kaur et al. 2022). For human computation systems, the ex-ante explanation is to explain the use, working, and the feature of the system, and the ex-post explanation is to explain the reasons that led to a particular decision.

Along with the recent spread of AI, there has been a growing demand for explaining AI. To gain the trust of the users, it is necessary to explain the behavior of AI systems (Ribeiro et al. 2016). However, what should be explained depends on each individual case (Adadi and Berrada 2018).

To ensure the transparency of human computation systems, in addition to the explanation of AI, profiles and characteristics of data produced by workers is important. Besides, explanation of AI should be understandable and justified for general users. This section covers studies on explanation of the data and evaluation of interoperability by humans. The related studies are summarized in Table 5.

There have been several proposals for explaining data. Bender and Friedman (2018) proposed data statements for natural language processing. Data statements were proposed to address scientific and ethical issues that result from the use of data from certain populations for other populations. A data statement is a characterization of a dataset that provides a context of the dataset for developers and users. Gebru et al. (2021) proposed datasheets for datasets. A datasheet describes characteristics of datasets for machine learning. It documents its motivation, composition, collection process, recommended uses, and other information. Gebru et al. (2021) provided a set of questions designed to elicit the information and a workflow for dataset creators to use when answering these questions. Note that privacy risks in these explanation should be considered. There is a risk of privacy leakage in the explanation of data creation, for example, by disclosing detailed profiles of a small number of crowd workers (Bender and Friedman 2018).

To explain machine learning models and their outcomes, various interpretation methods have been proposed (Guidotti et al. 2018). Major approaches of interpretation methods include local explanation, global explanation, and example-based explanation. LIME (Ribeiro et al. 2016) is a representative local explanation method. LIME estimates

Table 5 Studies on transparency in human computation

Category	References	Description
Data explanation	Bender and Friedman (2018)	Dataset description for natural language processing
	Gebru et al. (2021)	Dataset description for machine learning
Evaluation of interpretability	Doshi-Velez and Kim (2017)	Classification of crowdsourced evaluation schemes
	Hutton et al. (2012)	Worker preference of explanations for text classifiers
	Selvaraju et al. (2017)	Worker rating of relative reliability of the visual explanations
	Jeyakumar et al. (2020)	Comparison of user preferences for the explanation methods
	Can et al. (2018)	Evaluation of visual ambiance cues
	Lu et al. (2021)	Evaluation based on a human computation game

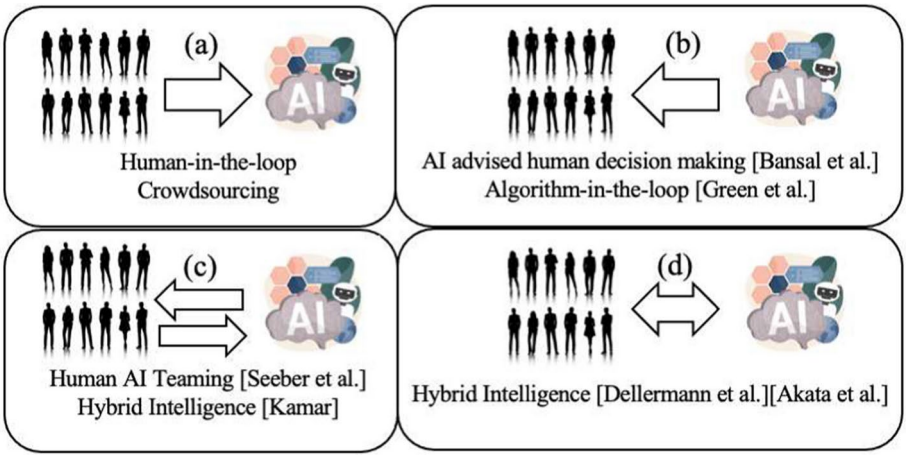

Fig. 2 Collaborative human computation can take several forms: **a** humans intervene and participate in decision-making of AI. This includes human-in-the-loop. **b** AI supports human decision-making. This includes AI advised human decision-making and algorithm-in-the-loop. **c** humans and AI support each other and sometimes work together as a team to solve problems. This includes human–AI teaming. **d** humans and AI co-evolve and build mutually reciprocal relationships. This includes Hybrid Intelligence

local linear models for explanation of a certain outcome. In the case of an image classifier, one approach of local explanation is to identify pixels that strongly influence the outcome (Baehrens et al. 2010).

Although the aforementioned explanation methods provide explanations and interpretations of algorithmic decision processes, it is not clear whether a certain explanation is appropriate, suitable, and reasonable for a certain task. Several methods have been proposed to automatically evaluate the performance of explanation methods. For example, Samek et al. (2016) proposed a methodology based on region perturbation for evaluating ordered col-

lections of pixels such as heatmaps. However, such computational evaluation of explanation is not necessarily the same for humans' evaluation (Narayanan et al. 2018). Rationalization provided by machine learning algorithms can be difficult to interpret. For example, algorithms may recognize objects based on their relationship to the background rather than the object itself. An airplane may be recognized by the sky, because the background of the image of an airplane is often the sky. This is a reasonable strategy for algorithms to make decisions based on statistical information. However, this is not justified for humans. Verifying the rationality of the explanation can be effective to build interpretable model for humans.

It is reasonable to use crowdsourcing as a basis for scaling up human-based evaluation of interpretability. Several human-based evaluation schemes using crowdsourcing have been proposed to measure the capability of explanation methods. In the classification by Doshi-Velez and Kim (2017), crowdsourced evaluation schemes can be divided into three types: binary forced choice, forward simulation/prediction, and counterfactual simulation. Hutton et al. (2012) assessed the explanations for text classification by using crowd worker evaluation. The workers compared human- and computer-generated explanations and indicated which they preferred and why. They demonstrated a slight preference for computer-generated explanations. Selvaraju et al. (2017) conducted user studies to measure the reliability of the visual explanation for image classification. Workers were instructed to rate the reliability of the models relative to each other. Jeyakumar et al. (2020) performed a cross-analysis user study to compare the explanation methods to assess user preferences. The study was conducted across applications spanning image, text, audio, and sensory domains. Can et al. (2018) conducted crowd-based assessment of machine recognition of ambiance. Lu et al. (2021) proposed a human-based evaluation method based on Peek-a-boom (Von Ahn et al. 2006), a human computation game used for image annotation. In Peek-a-boom, the Boom player selects important parts of an image and presents them to the Peek player. An XAI method plays the Boom instead of another human.

4 Collaborative human computation

Sections 2 and 3 discussed the one-way trust between humans and AI. In this chapter, we consider two-way trust, or human–AI collaboration, in which humans and AI build reciprocal relationships and work together to accomplish difficult tasks (which cannot be done by one or the other in isolation). Bedwell et al. (2012) organized the cross-disciplinary concept of collaboration, with "evolving process whereby two or more social entities actively and reciprocally engage in joint activities aimed at achieving at least one shared goal" as the definition of collaboration. They define collaboration as a superordinate construct, which is distinguished from related constructs, such as teamwork, coordination, and cooperation, under their evaluation criteria and definition components of "evolving process," "two or more social entities," "actively and reciprocally participate," and "achieving at least one shared goal."

Drawing from the definition of collaboration, this chapter explores how humans and AI, as social entities, can forge reciprocal relationships and achieve collaborative human computation to tackle complex problems. Through a comprehensive literature survey, we identify and categorize various approaches to realizing human–AI collaboration within the

context of human computation into four forms (Fig. 2). These forms can be conceptualized as progressively exhibiting a higher degree of reciprocity and initiative between humans and AI. In the following sections, we provide an overview of the characteristics and discuss representative studies for each form of human–AI collaboration, with a particular emphasis on its associated trust aspects.

4.1 Human-in-the-loop AI/machine learning

In human-in-the-loop AI (Wu et al. 2022; Mosqueira-Rey et al. 2022), humans are involved in part of the workflow to accomplish tasks, with AI taking the initiative. This human involvement includes providing information for AI to make decisions, as well as intervening and participating in decision-making by confirming or modifying the AI's decisions. We regard human-in-the-loop AI as a form of human–AI collaboration.

An early attempt to involve humans in the AI/machine learning process was VizWiz (Bigham et al. 2010), a human-in-the-loop visual question answering application to assist visually impaired persons. VizWiz uses the recognition results of the AI as-is when the AI is sufficiently confident, but automatically issues crowdsourcing tasks that leave the decision to the human when it is not. In addition, Wilder et al. (2021) proposed a method to quantify the cost of querying a human and the cost of making an incorrect decision using AI alone, and to make an overall decision that balances these costs. A more complex case is a workflow such as content generation, which consists of various steps such as generation, modification, and evaluation. AI controls the workflow, which consists of multiple humans performing various tasks, so that it is executed in an overall optimal manner. Dai et al. (2010, 2011) considered workflow control as a Markov decision process and proposed an approach for optimally controlling this process.

In human-in-the-loop machine learning, humans participate in certain parts of the machine learning process, such as providing data and data modeling. Providing annotations for target labels in supervised learning is a typical example of this. For example, Raykar et al. (2010) proposed the learning from crowds setting that simultaneously estimates ground truth target labels, worker confidence, and prediction models based on annotations provided by crowd workers with different (unknown) reliability levels. from crowds setting. We will not cover all of them here, but there are developments such as active learning (Yan et al. 2011), which actively selects unlabeled data and workers to request annotations to reduce the total annotation costs, and extensions to deep learning (Rodrigues and Pereira 2018).

What can be expected from human contributions is not limited to providing target labels, but also includes these input features for machine-learning prediction models. For example, Branson et al. (2010) proposed a human-in-the-loop image recognition system in which humans were responsible for extracting (abstract) features for the inputs of prediction models. This approach is effective in difficult tasks where sufficient data are not available for AI to be able to identify images, while even typical humans cannot directly recognize target labels owing to the lack of expertise. Cheng and Bernstein (2015) proposed the idea of asking humans to define features through tasks that ask them to distinguish between positive and negative examples. Furthermore, Takahama et al. (2018) proposed a boosting-based method that adaptively and efficiently performs the training.

Another approach is to obtain a representation of the data from the similarity evaluations of the data provided by humans. Gomes et al. (2011) proposed a Bayesian method for

obtaining data embeddings that reflect the results of similarity comparisons of pairs of data by crowd workers. For humans, relative evaluation is often easier than absolute evaluation. That is, it is easier to answer the relative similarity question "Is object A more similar to object B or object C?" than the absolute similarity question "Are objects A and B similar?" Tamuz et al. (2011) and Wilber et al. (2014) obtained object similarity and embedding from these triplet-wise comparisons. Furthermore, Amid and Ukkonen (2015) proposed a multi-view embedding that considered comparisons from several different viewpoints.

The execution of data analysis processes using machine learning algorithms by a large number of humans can also be considered a human-in-the-loop process in a broad sense. A typical example is data analysis competitions such as Kaggle. In actual data modeling, no single method always achieves the best performance (Wolpert 2002), and it is very effective to search extensively for a model that fits data using many people. Typical competitions employ winner-take-all competitive mechanisms, that does not motivate cooperation among participants. To overcome this problem, a competition mechanism has been proposed that allocates rewards linked to the performance gains brought by participants in order to motivate them to contribute and share their models early (Abernethy and Frongillo 2011).

4.2 Algorithm-in-the-loop

In contrast to human-in-the-loop, which involves human participation in AI and machine learning processes, the idea of algorithm-in-the-loop, where AI and machine learning models support human decision-making, has also been proposed. Green and Chen (2019) propose the three core principles of algorithm-in-the-loop. They include accuracy ("people using the algorithm should make more accurate predictions than they could without the algorithm"), reliability ("people should accurately evaluate their own and the algorithm's performance and should calibrate their use of the algorithm to account for its accuracy and errors", fairness ("people should interact with the algorithm in ways that are unbiased with regard to race, gender, and other sensitive attributes").

In connection with algorithm-in-the-loop to support human decision-making, AI-advised human decision-making, where humans and AI make decisions as a team and solve complex problems, has emerged as another form of human–AI collaboration (Bansal et al. 2019b). In this paradigm, users receive action recommendations (or predicted results in simpler cases) from AI systems to solve problems. Users may choose to follow the AI's proposed action or decide on an alternative course. Such decision-making processes can be applied to decision-critical domains, such as medical diagnosis, recruitment selection, and loan approval.

4.3 Human–AI teaming

As seen in human-in-the-loop, human intervention can help prevent errors that AI systems alone may cause. Moreover, human feedback contributes to improving the AI system's learning cycle. In order to enable AI systems to learn continuously in human-in-the-loop, AI systems must determine when and how to leverage human knowledge. Kamar (2016) defined hybrid intelligence as a human–AI collaboration that incorporates human knowledge into AI systems to complement AI capabilities, where AI systems can delegate tasks to humans when necessary. They claim that humans can assume various roles within hybrid intelligence, ranging from crowdsource workers and citizen scientists to teachers actively

intervening in AI systems. Such active human intervention in AI systems leads to crucial human–AI teamwork dynamics. In this section, we discuss such human–AI teaming as a form of human–AI collaboration where humans and AI work together as a team to solve problems. While human-in-the-loop AI involves human intervention in a specific part of the AI system, human–AI teaming is characterized by more mutual and continuous collaboration between humans and AI.

Seeber et al. (2020) outline a research agenda on AI in team collaboration, based on research questions derived from a survey of collaboration studies researchers. They considered a hypothetical scenario in which a devastating hurricane impacts a small country. In such complex situations, AI systems, as teammates, are expected to collaborate with humans to address a range of intricate and diverse problems. The scenario exemplifies a complex situation in which human and AI teammates must swiftly analyse circumstances, communicate and cooperate, coordinate emergency responses, and identify feasible solutions. Collaborative problem-solving between humans and AI involves identifying problem causes, proposing and evaluating solutions, selecting options, planning, taking act, learning from past interactions, and participating in after-action reviews. In this context, they define machines as teammates as "those technologies that draw inferences from information, derive new insights from information, find and provide relevant information to test assumptions, debate the validity of propositions offering evidence and arguments, propose solutions to unstructured problems, and participate in cognitive decision-making processes with human actors." They design machines as teammates, considering seven perspectives: appearance, sensing and awareness, learning and knowledge processing, conversation, architecture, visibility, and reliability.

Zhang et al. (2021) investigate the relationship between humans and AI as teammates in a multiplayer online game. Specifically, they addressed three research questions: (1) how people perceive AI teammates; (2) what factors influence people to team up with AI; and (3) what people expect from their AI teammates. Their results reveal that perceptions of AI teammates encompass both positive and negative perceptions. AI is often perceived as a tool, and past experiences and attitudes in collaboration with AI influence people's willingness to team up with AI. Furthermore, AI teammates are expected to possess proficient skills, share common understandings with humans, and engage in effective communication.

4.4 Hybrid intelligence

In the context of extensive discussions on human–AI collaboration, where humans and AI work together as a team to solve problems, several studies propose a new form of collaboration in which both human and AI co-evolve and establish mutually reciprocal relationships. Dellermann et al. (2019a) characterize such co-evolving and reciprocal relationships between humans and AI as "Systems that have the ability to accomplish complex goals by combining human and artificial intelligence to collectively achieve superior results than each of the entities could have done in separation and continuously improve by learning from each other." They refer to such systems as hybrid intelligence, which is distinct from Kamar (2016)'s definition. To systematically organize the knowledge for designing hybrid intelligence systems, They conduct an extensive literature review based on search queries

related to hybrid intelligence system.[1] As a result, they develop a comprehensive taxonomy consisting of four meta-dimensions, 16 sub-dimensions, and 50 categories (Table 6).

Similarly, Akata et al. (2020) define Hybrid Intelligence, with a particular focus on AI amplifying human intellectual capabilities, as "the combination of human and machine intelligence, augmenting human intellect and capabilities instead of replacing them, to make meaningful decisions, perform appropriate actions, and achieve goals that were unreachable by either humans or machines." They organized research questions for hybrid intelligence systems in terms of collaborative, adaptive, responsible, and explainable aspects.

Based on the key concepts common to these definitions, Hybrid Intelligence is characterized by its aim for co-evolution and mutual augmentation of humans and AI. In contrast, human–AI teaming rather focuses on mutually complementary collaboration between humans and AI. Hybrid Intelligence can be defined as a form of human–AI collaboration that enables the achievement of complex goals through a continuous learning process in which humans and AI continuously learn from and utilize each other's knowledge and capabilities, eventually co-evolving together.

An prime example of the potential of hybrid intelligence systems is AlphaGO. In AlphoGO, the AI learns from a vast amount of game data and amplifies human knowledge by introducing novel Go strategies to human players. In this manner, humans and AI in hybrid intelligence systems learn from each other through various mechanisms, such as labelling, demonstrating, teaching adversarial moves, criticizing, and rewarding. This enables them to co-evolve and establish reciprocal relationships. The complex problems addressed by hybrid intelligence are typically time-variant, dynamic, require considerable domain knowledge, and lacking specific ground truth (Dellermann et al. 2019b). Application areas of hybrid intelligence include strategic decision-making, science, healthcare, education, innovation and creativity (Dellermann et al. 2019b; Akata et al. 2020).

4.5 Trust aspects in collaborative human computation

Our exploration of collaborative human–AI systems, ranging from human-in-the-loop approaches to hybrid intelligence, has revealed a recurring theme: the fundamental role of trust. This element underpins successful collaboration between humans and AI, shap-

[1] The search queries included "hybrid intelligence OR human-in-the-loop OR interactive machine learning OR machine teaching OR machine learning AND crowdsourcing OR human supervision OR human understandable machine learning OR human concept learning."

Table 6 Taxonomy of hybrid intelligence design (Dellermann et al. 2019a)

Meta-dimensions	Dimensions
Task characteristics	Type, goals, data representation, timing
Learning paradigm	Augmentation, machine learning, human learning
Human–AI interaction	Machine teaching, teaching interaction, expertise requirements, amount of human input, aggregation, incentives
AI–human interaction	Query strategy, machine feedback, interpretability

ing human willingness to engage with AI systems, the reliability of AI-assisted decision-making, and the overall performance of human–AI teaming.

The following subsections investigate crucial aspects of trust in collaborative human computation: how trust is established and sustained, the role of explainable AI in fostering trust, the interplay between trust and performance, and evaluating trust in the context of human–AI collaboration. These trust aspects are crucial to all forms of collaborative human computation discussed earlier and play a pivotal role in realizing the full potential of trustworthy human computation.

4.5.1 Trust in collaborative human computation

In human–AI collaboration, such as the hybrid intelligence described earlier, establishing appropriate trust between humans and AI is crucial. As discussed in Sects. 2 and 3, the mutual evaluation between human computation systems and human participants contributes to visualizing and fostering trust on both sides. Similarly, trust between humans and AI based on mutual evaluation in human–AI collaboration is also discussed (Jorge et al. 2022).

Regarding trust in human–AI collaboration, Dellermann et al. (2019b) emphasize the interpretability and transparency of AI as the foundations of trust. They argue that interpretability is important for removing bias, achieving reliability and robustness, causality of learning, debugging learning, and establishing trust. Specifically, they argue that interpretability can be achieved through (1) transparency, which allows the "black box" of algorithms to be opened; (2) global interpretability, which provides general interpretability of machine learning models; and (3) local prediction interpretability, which makes more complex models interpretable for single predictions. In a related discussion, Vössing et al. (2022) propose that explanations (descriptions of reasoning processes) (Gilpin et al. 2018) representing the internal states of the system in an understandable manner could help build trust. In human–AI collaboration, complete understanding of AI decisions is often unattainable. Therefore, they also emphasize the importance of establishing trust to manage the complexity and uncertainty of AI systems.

In situations where humans and AI work together as a team to solve problems, building trust significantly influences rational decisions regarding when to follow AI recommendations. Concerning when to trust AI and its recommendations, Bansal et al. (2019b) introduce the concept of mental models of AI. In connection with this discussion on mental models and trust (particularly learned trust (Marsh and Dibben 2003) based on contexts and past experiences), previous research in cognitive psychology demonstrates that people construct mental models when interacting with complex systems, which facilitate their use of the system (Norman 1988). Another study showed that people tend to build mental models even for autonomous systems, such as AI agents (Kulesza et al. 2012). Furthermore, Hoff and Bashir (2015) explore the relationship between mental models and trust in systems.

Building upon these discussions, Bansal et al. (2019b) argue that constructing sound mental models of AI helps humans decide when to trust AI, which in turn leads to improving human–AI team performance. Bansal et al. (2019a) also argue that maximizing AI accuracy does not necessarily lead to maximising human–AI team performance in human–AI collaboration. They emphasize the importance of assessing when and how humans and AI should complement each other to enhance human–AI team performance. In relation to the earlier discussion of mental models, they also propose developing mental models of error bound-

aries regarding when AI makes errors, which enables people to decide when to follow AI recommendations appropriately. The importance of mental models has also been recently achnowledged in human-in-the-loop systems (Chakraborti and Kambhampati 2018) and human-computer interactions (Kaur et al. 2019).

In summary, establishing trust and building accurate mental models of AI systems are crucial factors in achieving effective human–AI collaboration and its team performance. These factors help humans make informed decisions about when to rely on AI recommendations and how to best complement AI capabilities.

4.5.2 Explainable AI in collaborative human computation

Previous research has shown that building more accurate mental models of a system generally enhances trust in the system when explanations are provided (Staab et al. 2002). Regarding explanations and trust in human–AI collaboration, Lai and Tan (2019) investigated whether AI explanations improved human decision-making performance in deception-detection tasks. Their experimental results demonstrated that human performance improved slightly when AI presented only explanations without predicted labels. They argue that there is a trade-off between human performance and independence in AI-assisted decision-making, and that explanations from AI may mitigate this trade-off. In line with this discussion, Lai et al. (2021) further conducted a comprehensive survey on human–AI decision-making. In the survey, they categorize the elements of explanations from AI (called "AI assistance elements") as "Prediction", "Information about predictions", "Information about models (and training data)", and "Other" (Table 7), providing an overview of which explanations are used in existing studies on human–AI decision-making.

Previous studies have demonstrated the effects of explanations from AI in human–AI collaboration. Nourani et al. (2019) investigate how explanations affect people's mental models of AI. Specifically, they examined how people perceive the accuracy of AI in an image classification task, depending on meaningful and non-meaningful explanations by AI (highlighting the areas of an image on which predictions are based). Their experimental results show that people estimate the accuracy of AI with non-meaningful explanations lower. They argue that if explanations are not based on rationality that people can understand, they cannot accurately estimate AI's accuracy, leading to less trust in AI. Smith-Renner et al. (2020) discuss that explanations from AI and human feedback are complementary. They show that a lack of opportunities for users to provide feedback on AI explanations (which highlight words on which predictions are based in their text categorization task) negatively impacts

Table 7 AI assistance elements in human–AI decision-making (Lai et al. 2021)

AI assistance elements	Examples
Information about predictions	Model uncertainty, local feature importance, rule-based explanations, example-based methods, counterfactual explanations, natural language explanations, partial decision boundary
Information about models (and training data)	Model performance, global feature importance, presentation of simple models, global example-based explanations, model documentation, information about training data
Others	Level of user agency, interventions or workflows affecting cognitive process

user experiences. They note that human feedback should not be requested without explanations from AI. In another study conducted on an image face detection task, Honeycutt et al. (2020) report that requesting feedback makes people perceive the system as more inaccurate, resulting in reduced trust. Wang and Yin (2021) identify three desirable properties of AI explanations in the context of human–AI decision-making: (1) they improve people's understanding; (2) they help people to recognize uncertainty in the AI model; and (3) they help increase trust in the AI model. They also pointed out that people's domain knowledge of tasks is required when receiving AI' explanations. They compare several explanatory methods, including feature importance-based, feature contribution-based, nearest neighbors and counter-factual-based, in two types of human–AI decision-making tasks, recidivism prediction and forest cover prediction. Their experimental results reveal that none of the explanatory methods satisfy the aforementioned properties when people have limited domain knowledge of the tasks.

4.5.3 Performance in collaborative human computation

While explanations from AI can enhance trust and improve human–AI collaboration, various side effects of explanations from AI on the performance of human–AI collaboration have been also identified. The first problem is that explanations from AI may lead to information overload. Alufaisan et al. (2021) show that while AI's predictions improve the accuracy of human decision-making, AI's explanations (based on LIME) about the predictions have no effect on improving the accuracy in two types of decision-making tasks, recidivism prediction and income status prediction, This may be because AI's explanations lead to an information overload and therefore affect human cognitive abilities to detect errors in AI's predictions. The second problem is that explanations from AI may cause humans to unconditionally trust AI and accept its suggestions. Bansal et al. (2021) conducted an experiment investigating the impact of explanations on human–AI team performance in two types of decision-making tasks: text classification (sentiment analysis) and question answering. Therein, their system highlights words as local explanations on which AI's predictions are based. Their experimental results show that while their subjects accept AI suggestions because of their explanations, the explanations do not necessarily contribute to complementary performance improvements in human–AI collaboration. They argue that as explanations increase trust in AI, humans are accepting AI suggestions regardless of whether they are correct or incorrect.

Concerning the impact of AI's explanations on human–AI team performance, some issues exist when evaluating the performance. Buçinca et al. (2020) argue that the inferiority of human–AI team performance is attributed to the fact that its evaluation is based on the proxy task of how well humans can accurately predict AI's decisions (or its decision boundaries). The proxy task situation differs from directly evaluating how well humans and AI jointly perform as a team. They also point out that explanations from AI can be inductive (e.g., example-based explanations) or deductive (e.g., rule-based explanations). While inductive explanations are less cognitively demanding, deductive explanations are more cognitively demanding. In real decision-making situations, humans tend to avoid analytical thinking, which is cognitively demanding. In the proxy task, humans explicitly pay attention to and deliberate the AI's deductive explanations, which creates a different situation from an actual decision-making scenario.

Regarding cognitive loads on explanations from AI, Buçinca et al. (2021) point out that people tend to create heuristics to decide whether to follow AI suggestions to avoid the cognitive loads required to understand explanations from AI. Consequently, as discussed earlier, humans tend to unconditionally trust AI, which leads to the inferiority of human–AI team performance. To address this issue, they propose introducing cognitive forcing functions in human–AI collaboration, which encourages humans to engage in analytical thinking about AI's explanations. Cognitive forcing functions are defined as "interventions that are applied at the decision-making time to disrupt heuristic reasoning and thus cause the person to engage in analytical thinking". They exemplify several strategies for such cognitive forcing functions, including "asking the person to make a decision before seeing the AI's recommendations", "slowing down the process", and "letting the person choose whether and when to see the AI recommendation". In a different study, Green and Chen (2019) point out that even simple cognitive forcing, which allows people to make their own decisions before presenting AI's decisions, improves decision-making by a human–AI team. The implications of these studies indicate that introducing cognitive forcing functions in human–AI collaboration would reduce unconditional or excessive trust in AI caused by people avoiding their cognitive loads, and improve human–AI team performance.

4.5.4 Evaluating trust in collaborative human computation

As discussed earlier, explanations from AI can increase the interpretability and transparency of a model, leading to establishing trust in the model. Explanations from AI also help people build more accurate mental models of a system, resulting in enhancing trust. In this section, we discuss how to evaluate trust in the context of human–AI collaboration.

Vereschak et al. (2021) surveyed how to evaluate trust in human–AI decision-making. According to the definition of trust, "an attitude that an agent will achieve an individual's goal in a situation characterized by uncertainty and vulnerability" (Lee and See 2004), they propose three theoretical elements of trust, "vulnerability," "positive expectations," and "attitude." Vulnerability indicates that a person is in a situation where the outcome of a decision involves uncertainty and potentially negative or undesired consequences. From a social-cognitive viewpoint, they claim that trust is "attitude" rather than "behavior" as it cannot be systematically replaced by behavior and cannot be fully observed by a third party. They also pointed out that it is appropriate to discuss "confidence" instead of trust in situations where there are no elements of vulnerability and discuss "distrust" in the situations with no positive expectations. According to their organization of the concepts, "reliance" refers to whether a person follows the system's suggestions, and "compliance" refers to whether a person asks the system for suggestions. Those are not "attitude" but actual "behavior." Then, they propose objective, quantitative behavioral indicators for evaluating trust as shown in Table 8. In addition, Lai et al. (2021) propose both subjective and objective indicators for evaluating trust and reliance in human–AI decision-making as shown in Table 9.

We discussed earlier that explanations would lead to humans unconditionally trusting AI and degrading the human–AI team performance. To address this issue, the importance of making humans think analytically about AI's explanations by introducing cognitive forcing functions has been proposed. By paying more attention to AI' predictions and explanations, humans can carefully examine the correctness or incorrectness of AI' predictions. This also

Table 8 Trust-related behavioral measures (Vereschak et al. 2021)

Measures	Definition
Decision time	How fast a recommendation is accepted
Compliance	The number of times participants follow the systems' recommendations
Reliance	The number of times participants asked for a recommendation
Agreement/disagreement	How quickly a recommendation is accepted
Switch ratio	The number of times a participant who initially disagreed with the system decided to follow its recommendation in the end

Table 9 Evaluation metrics of trust and reliance (Lai et al. 2021)

Metrics	Examples
Subjective	Self-reported trust, model confidence/acceptance, self-reported agreement/reliance, perceived accuracy, perceived capability/benevolence/integrity, usage intention/willingness
Objective	Agreement/acceptance of model suggestions, switch, weight of advice, model influence (difference between conditions), disagreement/deviation, choice to use the model, over-reliance, under-reliance, appropriate reliance

leads to the development of accurate mental models of AI, which fosters trust in AI. The importance of such trust calibration (knowing when AI is wrong and when to trust/distrust AI) is commonly discussed in the context of machine automation (Pop et al. 2015). Zhang et al. (2020) show that presenting confidence scores (the chance that AI is correct, e.g., the probability of every single prediction) on AI's predictions helps calibrate trust (they evaluate using behavioral indicators of switch percentage and agreement percentage) in AI. Their experimental results show that AI's local explanations (feature weights-based) do not contribute to trust calibration because of the information overload that AI brings. They also point out that calibrating trust alone is insufficient to improve human–AI decision-making performance and that humans should have relevant domain knowledge to appropriately complement AI errors.

Finally, regarding the relationship between trust and ethics, Flathmann et al. (2021) point out that ethical AI gains trust from humans and therefore improves human–AI team performance to solve problems. Ethical AI is also important in fairness of human–AI decision-making. They identify the ethical requirements for human–AI team: (1) AI has at least partial autonomy in decision-making; (2) AI has a clear role in the team; and (3) AI is interdependent with humans on its activities and outcomes. They then propose a model of ethical human–AI teamwork, in which humans share their ethical ideology and AIs share their joint team-specific ethical ideology with each other. Such sharing of ethical ideology allows for building more robust trust between humans and AI as teammates. Ethical AI should gain trust over time by fulfilling its role in the team.

5 Challenges and open questions

Thus far, we have summarized previous research related to trustworthy human computation from the perspective of achieving this goal. Finally, we discuss four research directions for further development of trustworthy human computation: (1) tools and libraries; (2) secure and distributed human computation; (3) bias control; (4) reciprocal human computation; and (5) grand challenges.

5.1 Tools and libraries for trustworthy human computation

In Sect. 2, we surveyed the quality evaluation criteria for human computation from various perspectives such as reliability, availability, and serviceability. For example, methods for reliability include estimating worker ability, ground truth answers, as well model parameters. Methods for availability motivate active participation of workers. There are also techniques for serviceability such as human computation programming and workflow control. The current remarkable expansion of deep learning is supported by frameworks such as PyTorch and TensorFlow, which enable the use of deep learning without considering the details of learning algorithms and other complications. The development and standardization of these techniques as tools and libraries is an important factor for human computation to spread as a trustworthy problem-solving infrastructure. Specifically, although several research prototypes exist for workflow modeling and programming languages to improve the serviceability of human computation, none of them are widely used on commercial platforms. Similar to general software development, standardization of modeling methods and languages are important for proliferation of human computation.

5.2 Secure and distributed human computation

In addition to RAS, security is another factor that improves the trustworthiness of human computation as computational systems. Current human computation platforms operate on centralized servers. Because requesters and workers information is concentrated on these servers, there are concerns about the leakage of such information if the operator is untrustworthy. To address such security problems, a decentralized platform using blockchain was developed (Li et al. 2019a). On this platform, human computation processes such as user registration, task registration, and task assignment are realized without a centralized server. It will be important to securely connect multiple human computation platforms to work together to achieve complex workflows.

5.3 Control of various biases and variances

In Sect. 2, we saw that humans have an overwhelmingly larger range of biases and variances than conventional computers pertaining to various factors such as attributes, knowledge, ability, and motivation, and these factors greatly affect the results of human computation. There is still room for further development of techniques to improve the quality of human computation by suppressing of unfavorable biases and variances and utilizing beneficial ones. For example, variation in worker ability is undesirable in terms of instability in the quality of each individual response, but is desirable for obtaining the highest quality results

in multiple responses. Worker self-selection bias, if well controlled, can also encourage the participation of workers who are closest to the correct response. Above all, cognitive biases are unique to humans. Although the reproducibility and universality of various types of cognitive biases are still under debate in the field of psychology, one important research direction would be to examine cognitive biases in the context of human computation and to use them for controlling human biases and variances.

In Sect. 3, we examined attempts to make human computation trustworthy in society. Because social biases are reflected in data, which directly affect the trustworthiness of human computation, the control of these biases is clearly important. On the other hand, regardless of how successful social biases can be technically removed, technology itself will not lead to social trust if it does not gain a certain level of understanding from society. Although this survey also introduced the XAI (explainability of AI), it is not yet clear how this can be connected to the accountability of fairness of human computation.

5.4 Reciprocal human computation

In Sect. 4, we examined the idea of an AI and groups of humans working together to solve difficult problems. This entails going beyond the limited/one-way trust described in Sects. 2 and 3, and building a mutually trusting relationship between AIs and humans. The construction of mutually beneficial and sustainable relationships, in which both parties can amplify their abilities and grow together through mutual trust is ideal for human computation. One of our major challenges for the future is to find common patterns by accumulating and analyzing successes and failures in various real-world human computation applications, and to develop this into a systematic theory of system construction that transcends individual cases.

5.5 Grand challenges in human computation

Human computation is a complex and interdisciplinary theme that develops through problems and solutions in various fields and through the fusion of knowledge. Just as the "grand challenges" are symbolic goals of AI, such as winning against top-tier players in board games or determining the three-dimensional structure of proteins in bioinformatics, and they have served as driving forces to propelling the development of the field, the grand challenges of human computation are also expected to play an important role in the further development of human computation. For example, research is among the most intelligent of human activities. Accelerating scientific discovery through collaboration between large numbers of AI and humans, contributing to the achievement of common human goals such as the SDGs (Sustainable Development Goals), solving pressing global issues such as COVID, and developing human resources through these efforts are major milestones which demonstrate the potential of human computation, which has previously been focused on performing relatively simple tasks.

6 Conclusion

In this survey, we organized various efforts related to the trustworthiness of human computation, which is closely related to both "human populations as users" and "human populations as driving forces," to establish mutual trust between these two human populations. First, we organized the existing research related to the trustworthiness of human computation as computing engines, that is, the trust experienced by humans of AI, by mapping them to the RAS trustworthiness measures in conventional computer systems. Thereafter, we summarized past discussions on the trustworthiness between human computation systems and users or participants, focusing on ethics such as fairness and privacy. We also summarized the discussion on human-in-the-loop human computation from the viewpoint of problem solving through collaboration between groups of humans and AI, as well as the viewpoint of hybrid intelligence, in which both parties cooperate more closely and grow together.

We believe that human computation, which involves the cooperation of many humans to solve problems that are difficult to solve by AI alone, is a concept that will serve as a foundation for exploring the use of AI with respect to humans, and the issues and research directions for trustworthy human computation described at the end of this survey will become a key factor in realizing human-centered AI.

Author contributions The authors contributed equally to this work.

Funding The authors were supported by Japan Science and Technology Agency (JST), Core Research for Evolutionary Science and Technology CREST Program, Grant Number JPMJCR21D1.

Declarations

Conflict of interest The authors declare that they have no Conflict of interest.

References

Abernethy JD, Frongillo R (2011) A collaborative mechanism for crowdsourcing prediction problems. Adv Neural Inf Process Syst 24:2600–2608

Adadi A, Berrada M (2018) Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6:52138–52160

Akata Z, Balliet D, de Rijke M et al (2020) A research agenda for hybrid intelligence: augmenting human intellect with collaborative, adaptive, responsible, and explainable artificial intelligence. Computer 53(8):18–28

Alm CO (2011) Subjective natural language problems: motivations, applications, characterizations, and implications. In: Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies (HLT), pp 107–112

Alufaisan Y, Marusich LR, Bakdash JZ et al (2021) Does explainable artificial intelligence improve human decision-making? In: Proceedings of the 35th AAAI conference on artificial intelligence (AAAI), pp 6618–6626
Amazon Web Services (2017) Amazon mechanical turk developer guide. https://docs.aws.amazon.com/pdfs/AWSMechTurk/latest/AWSMechanicalTurkRequester/amt-dg.pdf
Ambati V, Vogel S, Carbonell J (2011) Towards task recommendation in micro-task markets. In: Proceedings of the third human computation workshop (HCOMP), pp 80–83
Amid E, Ukkonen A (2015) Multiview triplet embedding: Learning attributes in multiple maps. In: Proceedings of the 32nd international conference on machine learning (ICML), pp 1472–1480
Archak N, Sundararajan A (2009) Optimal design of crowdsourcing contests. In: Proceedings of the 30th international conference on information systems
Baba Y, Kashima H (2013) Statistical quality estimation for general crowdsourcing tasks. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 554–562
Baba Y, Kashima H, Kinoshita K, et al (2013) Leveraging crowdsourcing to detect improper tasks in crowdsourcing marketplaces. In: Proceedings of the 25th conference on innovative applications of artificial intelligence (IAAI), pp 1487–1492
Bachrach Y, Minka T, Guiver J, et al (2012) How to grade a test without knowing the answers: a Bayesian graphical model for adaptive crowdsourcing and aptitude testing. In: Proceedings of the 29th international conference on international conference on machine learning (ICML), pp 819–826
Bacon DF, Parkes DC, Chen Y, et al (2012) Predicting your own effort. In: Proceedings of the 11th international conference on autonomous agents and multiagent systems (AAMAS), pp 695–702
Baehrens D, Schroeter T, Harmeling S et al (2010) How to explain individual classification decisions. J Mach Learn Res (JMLR) 11:1803–1831
Balzer WK, Sulsky LM (1992) Halo and performance appraisal research: a critical examination. Appl Psychol 6:975–985
Bansal G, Nushi B, Kamar E et al (2019a) Beyond accuracy: the role of mental models in human–AI team performance. In: Proceedings of the AAAI conference on human computation and crowdsourcing (HCOMP), pp 2–11
Bansal G, Nushi B, Kamar E et al (2019b) Updates in human–AI teams: understanding and addressing the performance/compatibility tradeoff. In: Proceedings of the 33rd AAAI conference on artificial intelligence (AAAI), pp 2429–2437
Bansal G, Wu T, Zhou J et al (2021) Does the whole exceed its parts? the effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI conference on human factors in computing systems (CHI)
Barbera DL, Roitero K, Demartini G, et al (2020) Crowdsourcing truthfulness: the impact of judgment scale and assessor bias. In: Proceedings of the 42nd European conference on IR research (ECIR), pp 207–214
Barbosa NM, Chen M (2019) Rehumanized crowdsourcing: a labeling framework addressing bias and ethics in machine learning. In: Proceedings of the 2019 CHI conference on human factors in computing systems (CHI), pp 1–12
Barocas S, Hardt M, Narayanan A (2019) Fairness and machine learning. fairmlbook.org. http://www.fairmlbook.org
Bedwell WL, Wildman JL, DiazGranados D et al (2012) Collaboration at work: an integrative multilevel conceptualization. Hum Resour Manag Rev 22(2):128–145
Bender EM, Friedman B (2018) Data statements for natural language processing: toward mitigating system bias and enabling better science. Trans Assoc Comput Linguist (TACL) 6:587–604
Bernstein MS, Brandt J, Miller RC et al (2011) Crowds in two seconds: enabling realtime crowd-powered interfaces. In: Proceedings of the 24th annual ACM symposium on user interface software and technology (UIST), pp 33–42
Bernstein MS, Karger DR, Miller RC et al (2012) Analytic methods for optimizing realtime crowdsourcing. In: Proceedings of the collective intelligence conference (CI)
Biel JI, Gatica-Perez D (2014) Mining crowdsourced first impressions in online social video. IEEE Trans Multimed 16:2062–2074
Bigham JP, Jayant C, Ji H, et al (2010) Vizwiz: nearly real-time answers to visual questions. In: Proceedings of the 23nd annual ACM symposium on user interface software and technology (UIST), pp 333–342
Braga DDS, Niemann M, Hellingrath B et al (2018) Survey on computational trust and reputation models. ACM Comput Surv 51(5)
Branson S, Wah C, Schroff F et al (2010) Visual recognition with humans in the loop. In: Proceedings of the 11th European conference on computer vision (ECCV), pp 438–451
Bruckman A (2002) Ethical guidelines for research online

Brynjolfsson E, McAfee A (2011) Race against the machine: how the digital revolution is accelerating innovation, driving productivity, and irreversibly transforming employment and the economy. Brynjolfsson and McAfee

Buçinca Z, Lin P, Gajos KZ et al (2020) Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. In: Proceedings of the 25th international conference on intelligent user interfaces (IUI), pp 454–464

Buçinca Z, Malaya MB, Gajos KZ (2021) To trust or to think: cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. In: Proceedings of the ACM on human–computer interaction 5(CSCW1)

Buolamwini J, Gebru T (2018) Gender shades: intersectional accuracy disparities in commercial gender classification. In: Proceedings of the 2018 conference on fairness, accountability and transparency (FAccT), pp 77–91

Burke JA, Estrin D, Hansen M, et al (2006) Participatory sensing. In: Proceedings of the world sensor web workshop (WSW)

Can G, Benkhedda Y, Gatica-Perez D (2018) Ambiance in social media venues: visual cue interpretation by machines and crowds. In: Proceedings of 2018 IEEE/CVF conference on computer vision and pattern recognition workshops (CVPRW), pp 2363–2372

Chakraborti T, Kambhampati S (2018) Algorithms for the greater good! on mental modeling and acceptable symbiosis in human–AI collaboration. arXiv:1801.09854

Chen X, Bennett PN, Collins-Thompson K et al (2013) Pairwise ranking aggregation in a crowdsourced setting. In: Proceedings of the 6th ACM international conference on web search and data mining (WSDM), pp 193–202

Cheng J, Bernstein MS (2015) Flock: Hybrid crowd-machine learning classifiers. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing (CSCW), pp 600–611

Cheng P, Lian X, Jian X et al (2019) Frog: A fast and reliable crowdsourcing framework. IEEE Trans Knowl Data Eng (TKDE) 31(5):894–908

Coscia M, Rossi L (2020) Distortions of political bias in crowdsourced misinformation flagging. J R Soc Interface 17:20200020

Dai P, Mausam, Weld DS (2010) Decision-theoretic control of crowd-sourced workflows. In: Proceedings of the 24th AAAI conference on artificial intelligence (AAAI), pp 1168–1174

Dai P, Mausam, Weld DS (2011) Artificial intelligence for artificial artificial intelligence. In: Proceedings of the 25th AAAI conference on artificial intelligence (AAAI), pp 1153–1159

Dai P, Rzeszotarski JM, Paritosh P et al (2015) And now for something completely different: improving crowdsourcing workflows with micro-diversions. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing (CSCW), pp 628–638

Daniel F, Kucherbaev P, Cappiello C et al (2018) Quality control in crowdsourcing: a survey of quality attributes, assessment techniques, and assurance actions. ACM Comput Surv (CSUR) 51(1):1–40

Dastin J (2022) Amazon scraps secret AI recruiting tool that showed bias against women. In: Ethics of data and analytics: concepts and cases, p 296

Davani AM, Díaz M, Prabhakaran V (2022) Dealing with disagreements: looking beyond the majority vote in subjective annotations. Trans Assoc Comput Linguist (TACL) 10:92–110

Dawid AP, Skene AM (1979) Maximum likelihood estimation of observer error-rates using the EM algorithm. J R Stat Soc Ser C (Appl Stat) 28(1):20–28

Dellermann D, Calma A, Lipusch N et al (2019a) The future of human–AI collaboration: a taxonomy of design knowledge for hybrid intelligence systems. In: Proceedings of the 52nd Hawaii international conference on system sciences (HICSS), pp 274–283

Dellermann D, Ebel P, Söllner M et al (2019) Hybrid intelligence. Bus Inf Syst Eng 61(5):637–643

Demartini G (2019) Implicit bias in crowdsourced knowledge graphs. In: Companion proceedings of the 2019 world wide web conference (WWW), pp 624–630

d'Eon G, Goh J, Larson K et al (2019) Paying crowd workers for collaborative work. In: Proceedings of the ACM human–computer interaction 3(CSCW)

Dignum V (2017) Responsible artificial intelligence: designing AI for human values. ICT Discoveries 1:1–8

DiPalantino D, Vojnovic M (2009) Crowdsourcing and all-pay auctions. In: Proceedings of the 10th ACM conference on electronic commerce (EC), pp 119–128

Dolmaya JM (2011) The ethics of crowdsourcing. Linguistica Antverpiensia New Seri Themes Transl Stud 10:97–110

Dong Z, Shi C, Sen S et al (2012) War versus inspirational in forrest gump: Cultural effects in tagging communities. In: Proceedings of the international AAAI conference on web and social media (ICWSM), pp 82–89

Donmez P, Carbonell JG, Schneider J (2009) Efficiently learning the accuracy of labeling sources for selective sampling. In: Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 259–268
Doshi-Velez F, Kim B (2017) Towards a rigorous science of interpretable machine learning. arXiv:1702.08608
Draws T, Rieger A, Inel O et al (2021) A checklist to combat cognitive biases in crowdsourcing. In: Proceedings of the AAAI conference on human computation and crowdsourcing (HCOMP), pp 48–59
Duan X, Ho CJ, Yin M (2020) Does exposure to diverse perspectives mitigate biases in crowdwork? An explorative study. In: Proceedings of the AAAI conference on human computation and crowdsourcing (HCOMP), pp 155–158
Echterhoff JM, Yarmand M, McAuley J (2022) AI-moderated decision-making: capturing and balancing anchoring bias in sequential decision tasks. In: Proceedings of CHI conference on human factors in computing systems (CHI)
Eickhoff C (2018) Cognitive biases in crowdsourcing. In: Proceedings of the eleventh ACM international conference on web search and data mining (WSDM), pp 162–170
Eickhoff C, de Vries AP (2013) Increasing cheat robustness of crowdsourcing tasks. Inf Retr 16:121–137
Faltings B, Pu P, Duy B et al (2014) Incentives to counter bias in human computation. In: Proceedings of the second AAAI conference on human computation and crowdsourcing (HCOMP), pp 59–66
Feng W, Yan Z, Zhang H et al (2018) A survey on security, privacy, and trust in mobile crowdsourcing. IEEE Internet Things J 5(4):2971–2992
Feyisetan O, Simperl E (2019) Beyond monetary incentives: experiments in paid microtask contests. ACM Trans Soc Comput (TSC) 2(2)
Finn P, Jakobsson M (2007) Designing ethical phishing experiments. IEEE Technol Soc Mag 26(1):46–58
Flathmann C, Schelble BG, Zhang R et al (2021) Modeling and guiding the creation of ethical human–AI teams. In: Proceedings of the 2021 AAAI/ACM conference on AI, ethics, and society (AIES), pp 469–479
Frankel MS, Siang S (1999) Ethical and legal aspects of human subjects research on the internet
Gadiraju U, Fetahu B, Kawase R et al (2017) Using worker self-assessments for competence-based preselection in crowdsourcing microtasks. ACM Trans Comput Hum Interact 24(4):30
Gebru T, Morgenstern J, Vecchione B et al (2021) Datasheets for datasets. Commun ACM 64(12):86–92
Gemalmaz MA, Yin M (2021) Accounting for confirmation bias in crowdsourced label aggregation. In: Proceedings of the thirtieth international joint conference on artificial intelligence (IJCAI), pp 1729–1735
Gilpin LH, Bau D, Yuan BZ, et al (2018) Explaining explanations: an overview of interpretability of machine learning. In: Proceedings of the fifth IEEE international conference on data science and advanced analytics (DSAA), pp 80–89
Gomes R, Welinder P, Krause A et al (2011) Crowdclustering. In: Advances in neural information processing, vol 24
Gordon ML, Zhou K, Patel K et al (2021) The disagreement deconvolution: Bringing machine learning performance metrics in line with reality. In: Proceedings of the 2021 CHI conference on human factors in computing systems (CHI)
Gordon ML, Lam MS, Park JS et al (2022) Jury learning: integrating dissenting voices into machine learning models. In: Proceedings of the 2022 CHI conference on human factors in computing systems (CHI), pp 1–19
Green B, Chen Y (2019) The principles and limits of algorithm-in-the-loop decision making. Proceedings of the ACM on human–computer interaction, vol 3, no (CSCW)
Guidotti R, Monreale A, Ruggieri S et al (2018) A survey of methods for explaining black box models. ACM Comput Surv (CSUR) 51(5):1–42
High-Level Expert Group on Artificial Intelligence of the European Commission (2019) Ethics guidelines for trustworthy AI. https://www.aepd.es/sites/default/files/2019-12/ai-ethics-guidelines.pdf
Hoff KA, Bashir M (2015) Trust in automation: Integrating empirical evidence on factors that influence trust. Hum Factors 57(3):407–434
Honeycutt D, Nourani M, Ragan E (2020) Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In: Proceedings of the AAAI conference on human computation and crowdsourcing (HCOMP), pp 63–72
Hube C, Fetahu B, Gadiraju U (2019) Understanding and mitigating worker biases in the crowdsourced collection of subjective judgments. In: Proceedings of the 2019 CHI conference on human factors in computing systems (CHI)
Hutton A, Liu A, Martin C (2012) Crowdsourcing evaluations of classifier interpretability. In: AAAI Spring symposium series
Ipeirotis PG, Gabrilovich E (2014) Quizz: targeted crowdsourcing with a billion (potential) users. In: Proceedings of the 23rd international conference on world wide web (WWW), pp 143–154

Ipeirotis PG, Provost F, Wang J (2010) Quality management on amazon mechanical turk. In: Proceedings of the ACM SIGKDD workshop on human computation (HCOMP), pp 64–67
Irani LC, Silberman MS (2013) Turkopticon: interrupting worker invisibility in amazon mechanical turk. In: Proceedings of the SIGCHI conference on human factors in computing systems (CHI), pp 611–620
Jagatic TN, Johnson NA, Jakobsson M et al (2007) Social phishing. Commun ACM 50(10):94–100
Jeyakumar JV, Noor J, Cheng YH et al (2020) How can I explain this to you? An empirical study of deep neural network explanation methods. Adv Neural Inf Process Syst 33:4211–4222
Jorge CC, Tielman ML, Jonker CM (2022) Artificial trust as a tool in human–AI teams. In: Proceedings of the 17th ACM/IEEE international conference on human–robot interaction (HRI), pp 1155–1157
Kaelbling LP (1990) Learning in Embedded Systems. PhD thesis, Department of Computer Science, Stanford University
Kajino H, Arai H, Kashima H (2014) Preserving worker privacy in crowdsourcing. Data Min Knowl Discov (DMKD) 28(5–6):1314–1335
Kajino H, Baba Y, Kashima H (2014b) Instance-privacy preserving crowdsourcing. In: Proceedings of the second AAAI conference on human computation and crowdsourcing (HCOMP), pp 96–103
Kamar E (2016) Directions in hybrid intelligence: complementing AI systems with human intelligence. In: Proceedings of the 30th international joint conference on artificial intelligence (IJCAI), pp 4070–4073
Kaplan T, Saito S, Hara K et al (2018) Striving to earn more: a survey of work strategies and tool use among crowd workers. In: Proceedings of the sixth AAAI conference on human computation and crowdsourcing (HCOMP), pp 70–78
Kaur H, Williams A, Lasecki WS (2019) Building shared mental models between humans and AI for effective collaboration. In: Proceedings of CHI 2019 workshop on where is the human? Bridging the gap between AI and HCI
Kaur D, Uslu S, Rittichier KJ et al (2022) Trustworthy artificial intelligence: a review. ACM Comput Surv (CSUR) 55(2):1–38
Kazai G, Kamps J, Koolen M et al (2011) Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In: Proceedings of the 34th international ACM SIGIR conference on research and development in information retrieval (SIGIR), pp 205–214
Kittur A, Smus B, Khamkar S et al (2011) Crowdforge: crowdsourcing complex work. In: Proceedings of the 24th annual ACM symposium on user interface software and technology (UIST), pp 43–52
Kittur A, Khamkar S, André P et al (2012) Crowdweaver: visually managing complex crowd work. In: Proceedings of the ACM 2012 conference on computer supported cooperative work (CSCW), pp 1033–1036
Kroll JA, Huey J, Barocas S et al (2017) Accountable algorithms. University of Pennsylvania Law Review, p 165
Kulesza T, Stumpf S, Burnett M et al (2012) Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In: Proceedings of the CHI conference on human factors in computing systems (CHI), pp 1–10
Kulkarni A, Can M, Hartmann B (2012a) Collaboratively crowdsourcing workflows with turkomatic. In: Proceedings of the ACM 2012 conference on computer supported cooperative work (CSCW), pp 1003–1012
Kulkarni A, Gutheim P, Narula P et al (2012) Mobileworks: designing for quality in a managed crowdsourcing architecture. IEEE Internet Comput 16(5):28–35
Kulkarni CE, Socher R, Bernstein MS et al (2014) Scaling short-answer grading by combining peer assessment with algorithmic scoring. In: Proceedings of the first ACM conference on learning @ Scale (L@S), pp 99–108
Lai V, Chen C, Liao QV et al (2021) Towards a science of human–AI decision making: a survey of empirical studies. arXiv:2112.11471
Lai V, Tan C (2019) On human predictions with explanations and predictions of machine learning models: a case study on deception detection. In: Proceedings of the conference on fairness, accountability, and transparency (FAT*), pp 29–38
Law E, von Ahn L (2011) Human computation. Morgan & Claypool Publishers
Lease M (2011) On quality control and machine learning in crowdsourcing. In: Proceedings of the third human computation workshop (HCOMP)
Lee JD, See KA (2004) Trust in automation: designing for appropriate reliance. Hum Factors 46(1):50–80
Li H, Zhao B, Fuxman A (2014) The wisdom of minority: discovering and targeting the right group of workers for crowdsourcing. In: Proceedings of the 23rd international conference on world wide web (WWW), pp 165–176
Li M, Weng J, Yang A et al (2019a) Crowdbc: a blockchain-based decentralized framework for crowdsourcing. IEEE Trans Parallel Distrib Syst (TPDS) 30(6):1251–1266

Li Y, Rubinstein B, Cohn T (2019b) Exploiting worker correlation for label aggregation in crowdsourcing. In: Proceedings of the international conference on machine learning (ICML), pp 3886–3895
Linden WJ, Hambleton RK (eds) (1997) Handbook of modern item response theory. Springer, Berlin
Little G, Sun YA (2011) Human OCR: insights from a complex human computation process. In: Proceedings of CHI 2011 workshop on crowdsourcing and human computation, pp 8–11
Little DG, Chilton LB, Goldman M et al (2010a) Turkit: human computation algorithms on mechanical turk. In: Proceedings of the 23nd annual ACM symposium on user interface software and technology (UIST), pp 57–66
Little G, Chilton LB, Goldman M et al (2010b) Exploring iterative and parallel human computation processes. In: Proceedings of the ACM SIGKDD workshop on human computation (HCOMP), pp 68–76
Liu H, Thekinen J, Mollaoglu S et al (2022) Toward annotator group bias in crowdsourcing. In: Proceedings of the 60th annual meeting of the association for computational linguistics (ACL), pp 1797–1806
Lu X, Tolmachev A, Yamamoto T et al (2021) Crowdsourcing evaluation of saliency-based XAI methods. In: Proceedings of the European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD), pp 431–446
Mao A, Kamar E, Horvitz E (2013) Why stop now? Predicting worker engagement in online crowdsourcing. In: Proceedings of the first AAAI conference on human computation and crowdsourcing (HCOMP), pp 103–111
Marsh S, Dibben MR (2003) The role of trust in information science and technology. Ann Rev Inf Sci Technol (ARIST) 37:465–98
Mason W, Watts DJ (2009) Financial incentives and the "performance of crowds". In: Proceedings of the ACM SIGKDD workshop on human computation (HCOMP), pp 77–85
Matsui T, Baba Y, Kamishima T et al (2014) Crowdordering. In: Proceedings of the 18th Pacific-Asia conference on knowledge discovery and data mining (PAKDD), pp 336–347
Mehrabi N, Morstatter F, Saxena N et al (2021) A survey on bias and fairness in machine learning. ACM Comput Surv (CSUR) 54(6)
Miao X, Peng H, Gao Y et al (2022) On dynamically pricing crowdsourcing tasks. In: ACM Transactions on knowledge discovery from data (TKDD). Just Accepted
Michelucci P (2013) Handbook of human computation. Springer, Berlin
Minder P, Bernstein A (2012) Crowdlang: a programming language for the systematic exploration of human computation systems. In: Proceedings of the fourth international conference on social informatics (SocInfo), pp 124–137
Moldovanu B, Sela A (2001) The optimal allocation of prizes in contests. Am Econ Rev 91(3):542–558
Monarch RM (2021) Human-in-the-loop machine learning: active learning and annotation for human-centered AI. Simon and Schuster
Morishima A, Shinagawa N, Mitsuishi T et al (2012) Cylog/crowd4u: a declarative platform for complex data-centric crowdsourcing. Proc VLDB Endow 5(12):1918–1921
Mosqueira-Rey E, Hernández-Pereira E, Alonso-Ríos D et al (2022) Human-in-the-loop machine learning: a state of the art. Artif Intell Rev 1–50
Muldoon C, O'Grady MJ, O'Hare GM (2018) A survey of incentive engineering for crowdsourcing. Knowl Eng Rev 33
Narayanan A (2018) Translation tutorial: 21 fairness definitions and their politics. In: Proc. Conf. Fairness accountability Transp., New York, USA, p 3
Narayanan M, Chen E, He J et al (2018) How do humans understand explanations from machine learning systems? An evaluation of the human-interpretability of explanation. arXiv:1802.00682
National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research (1978) The Belmont report: ethical principles and guidelines for the protection of human subjects of research, vol 2. Department of Health, Education, and Welfare
Neff G (2016) Talking to bots: symbiotic agency and the case of Tay. Int J Commun 10:4915–4931
Newell E, Ruths D (2016) How one microtask affects another. In: Proceedings of the 2016 CHI conference on human factors in computing systems (CHI), pp 3155–3166
Norman DA (1988) The psychology of everyday things. Basic books
Nourani M, Kabir S, Mohseni S et al (2019) The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems. In: Proceedings of the AAAI conference on human computation and crowdsourcing (HCOMP), pp 97–105
OECD (2019) OECD principles on AI. https://www.oecd.org/going-digital/ai/principles/
Oka M, Todo T, Sakurai Y et al (2014) Predicting own action: self-fulfilling prophecy induced by proper scoring rules. In: Proceedings of the second AAAI conference on human computation and crowdsourcing (HCOMP), pp 184–191

Otterbacher J, Checco A, Demartini G et al (2018) Investigating user perception of gender bias in image search: the role of sexism. In: Proceedings of the 41st international ACM SIGIR conference on research & development in information retrieval (SIGIR), pp 933–936
Oyama S, Baba Y, Sakurai Y et al (2013) Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In: Proceedings of the 23rd international joint conference on artificial intelligence (IJCAI), pp 2554–2560
Park H, Garcia-Molina H, Pang R et al (2012) Deco: a system for declarative crowdsourcing. Proc VLDB Endow 5(12):1990–1993
Pop VL, Shrewsbury A, Durso FT (2015) Individual differences in the calibration of trust in automation. Hum Factors 57(4):545–556
Quinn AJ, Bederson BB (2011) Human computation: a survey and taxonomy of a growing field. In: Proceedings of the CHI conference on human factors in computing systems (CHI), pp 1403–1412
Ra MR, Liu B, La Porta TF et al (2012) Medusa: a programming framework for crowd-sensing applications. In: Proceedings of the tenth international conference on mobile systems, applications, and services (MobiSys), pp 337–350
Raykar VC, Yu S (2011) Ranking annotators for crowdsourced labeling tasks. Adv Neural Inf Process 24:1809–1817
Raykar VC, Yu S, Zhao LH et al (2010) Learning from crowds. J Mach Learn Res 11:1297–1322
Ribeiro MT, Singh S, Guestrin C (2016) "Why should I trust you?" explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 1135–1144
Rodrigues F, Pereira FC (2018) Deep learning from crowds. In: Proceedings of the 32nd AAAI conference on artificial intelligence (AAAI)
Sabater J, Sierra C (2005) Review on computational trust and reputation models. Artif Intell Rev 24:33–60
Sakurai Y, Okimoto T, Oka M et al (2013) Ability grouping of crowd workers via reward discrimination. In: Proceedings of the first AAAI conference on human computation and crowdsourcing (HCOMP), pp 147–155
Samek W, Binder A, Montavon G et al (2016) Evaluating the visualization of what a deep neural network has learned. IEEE Trans Neural Netw Learn Syst (TNNLS) 28(11):2660–2673
Seeber I, Bittner E, Briggs RO et al (2020) Machines as teammates: a research agenda on AI in team collaboration. Inf Manag 57(2):103174
Selvaraju RR, Cogswell M, Das A et al (2017) Grad-CAM: visual explanations from deep networks via gradient-based localization. In: Proceedings of 2017 IEEE international conference on computer vision (ICCV), pp 618–626
Sen S, Giesel ME, Gold R et al (2015) Turkers, scholars, "arafat" and "peace": cultural communities and algorithmic gold standards. In: Proceedings of the 18th ACM conference on computer supported cooperative work & social computing (CSCW), pp 826–838
Shah DS, Schwartz HA, Hovy D (2020) Predictive biases in natural language processing models: a conceptual framework and overview. In: Proceedings of the 58th annual meeting of the association for computational linguistics (ACL), pp 5248–5264
Sheng VS, Provost F, Ipeirotis PG (2008) Get another label? Improving data quality and data mining using multiple, noisy labelers. In: Proceeding of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (KDD), pp 614–622
Shooman ML (2002) Reliability of computer systems and networks: fault tolerance, analysis, and design. Wiley, London
Shu J, Jia X, Yang K et al (2018a) Privacy-preserving task recommendation services for crowdsourcing. IEEE Trans Serv Comput 14(1):235–247
Shu J, Liu X, Jia X et al (2018b) Anonymous privacy-preserving task matching in crowdsourcing. IEEE Internet Things J 5(4):3068–3078
Siewiorek DP, Swarz RS (1998) Reliable computer systems: design and evaluation. AK Peters/CRC Press
Silberman MS, Irani L, Ross J (2010) Ethics and tactics of professional crowdwork. ACM XRDS 17(2):39–43
Silberman MS, Tomlinson B, LaPlante R et al (2018) Responsible research with crowds: pay crowdworkers at least minimum wage. Commun ACM 61(3):39–41
Smith-Renner A, Fan R, Birchfield M et al (2020) No explainability without accountability: an empirical study of explanations and feedback in interactive ml. In: Proceedings of the 2020 CHI conference on human factors in computing systems (CHI), pp 1–13
Snow R, O'Connor B, Jurafsky D et al (2008) Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the conference on empirical methods in natural language processing (EMNLP), pp 254–263
Staab S, Werthner H, Ricci F et al (2002) Intelligent systems for tourism. IEEE Intell Syst 17(06):53–64

Takahama R, Baba Y, Shimizu N et al (2018) Adaflock: adaptive feature discovery for human-in-the-loop predictive modeling. In: Proceedings of the 32nd AAAI conference on artificial intelligence (AAAI), pp 1619–1626
Tamuz O, Liu C, Belongie S et al (2011) Adaptively learning the crowd kernel. In: Proceedings of the 28th international conference on machine learning (ICML), pp 673–680
The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems (2019) Ethically aligned design: a vision for prioritizing human well-being with autonomous and intelligent systems, first edition. IEEE
Thiebes S, Lins S, Sunyaev A (2021) Trustworthy artificial intelligence. Electron Mark 31(2):447–464
Tranquillini S, Daniel F, Kucherbaev P et al (2015) Modeling, enacting, and integrating custom crowdsourcing processes. ACM Trans Web 9(2)
Truong NVQ, Dinh LC, Stein S et al (2022) Efficient and adaptive incentive selection for crowdsourcing contests. Appl Intell
Ueda R, Takeuchi K, Kashima H (2022) Mitigating observation biases in crowdsourced label aggregation. In: Proceedings of the 26th international conference on pattern recognition (ICPR)
Vaughan JW (2017) Making better use of the crowd: how crowdsourcing can advance machine learning research. J Mach Learn Res (JMLR) 18(1):7026–7071
Venanzi M, Guiver J, Kazai G et al (2014) Community-based Bayesian aggregation models for crowdsourcing. In: Proceedings of the international conference on world wide web (WWW), pp 155–164
Vereschak O, Bailly G, Caramiaux B (2021) How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies. In: Proceedings of the ACM on human–computer interaction, vol 5(CSCW2)
Voigt P, Von dem Bussche A (2017) The EU general data protection regulation (GDPR). Springer, Berlin
von Ahn L, Dabbish L (2008) Designing games with a purpose. Commun ACM 51(8):58–67
von Ahn L, Maurer B, McMillen C et al (2008) reCAPTCHA: Human-based character recognition via Web security measures. Science 321(5895):1465–1468
Von Ahn L, Liu R, Blum M (2006) Peekaboom: a game for locating objects in images. In: Proceedings of the CHI conference on human factors in computing systems (CHI), pp 55–64
Vössing M, Kühl N, Lind M et al (2022) Designing transparency for effective human–AI collaboration. Inf Syst Front 24:877–895
Wang S, Dang D (2022) A generative answer aggregation model for sentence-level crowdsourcing task. IEEE Trans Knowl Data Eng (TKDE)
Wang X, Yin M (2021) Are explanations helpful? A comparative study of the effects of explanations in AI-assisted decision-making. In: Proceedings of the 26th international conference on intelligent user interfaces (IUI), pp 318–328
Welinder P, Branson S, Belongie S et al (2010) The multidimensional wisdom of crowds. In: Advances in neural information processing systems, vol 23
Whitehill J, Ruvolo P, Wu T et al (2009) Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In: Advances in neural information processing systems, vol 22
Whiting ME, Gamage D, Gaikwad SNS et al (2017) Crowd guilds: worker-led reputation and feedback on crowdsourcing platforms. In: Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing (CSCW), pp 1902–1913
Wilber M, Kwak I, Belongie S (2014) Cost-effective hits for relative similarity comparisons. In: Proceedings of the AAAI conference on human computation and crowdsourcing (HCOMP), pp 227–233
Wilder B, Horvitz E, Kamar E (2021) Learning to complement humans. In: Proceedings of the 29th international joint conference on artificial intelligence (IJCAI), pp 1526–1533
Willett KW, Lintott CJ, Bamford SP et al (2013) Galaxy zoo 2: detailed morphological classifications for 304 122 galaxies from the Sloan digital sky survey. Mon Not R Astron Soc stt1458
Wolpert DH (2002) The supervised learning no-free-lunch theorems. Soft Comput Ind 25–42
Wu X, Fan W, Yu Y (2012) Sembler: ensembling crowd sequential labeling for improved quality. In: Proceedings of the AAAI conference on artificial intelligence (AAAI), pp 1713–1719
Wu X, Xiao L, Sun Y et al (2022) A survey of human-in-the-loop for machine learning. Futur Gener Comput Syst 135:364–381
Xue Y, Dilkina B, Damoulas T et al (2013) Improving your chances: boosting citizen science discovery. In: Proceedings of the first AAAI conference on human computation and crowdsourcing (HCOMP)
Yan Y, Rosales R, Fung G et al (2011) Active learning from crowds. In: Proceedings of the 28th international conference on machine learning (ICML), pp 1161–1168
Yuen MC, King I, Leung KS (2012) Taskrec: probabilistic matrix factorization in task recommendation in crowdsourcing systems. In: Proceedings of the 19th international conference on neural information processing (ICONIP), pp 516–525
Zhang H, Horvitz E, Parkes D (2013) Automated workflow synthesis. In: Proceedings of the 27th AAAI conference on artificial intelligence (AAAI), pp 1020–1026

Zhang Y, Liao QV, Bellamy RK (2020) Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Proceedings of the 2020 conference on fairness, accountability, and transparency (FAccT), pp 295–305

Zhang R, McNeese NJ, Freeman G et al (2021) "An ideal human" expectations of AI teammates in human–AI teaming. In: Proceedings of the ACM on human–computer interaction, vol 4(CSCW3)

Authors and Affiliations

Hisashi Kashima[1] · Satoshi Oyama[2] · Hiromi Arai[3] · Junichiro Mori[4]

✉ Hisashi Kashima
kashima@i.kyoto-u.ac.jp

Satoshi Oyama
oyama@ds.nagoya-cu.ac.jp

Hiromi Arai
hiromi.arai@riken.jp

Junichiro Mori
mori@mi.u-tokyo.ac.jp

[1] Graduate School of Informatics, Kyoto University, Yoshida Honmachi, Kyoto 6068501, Japan

[2] School of Data Science, Nagoya City University, 1 Yamanobata, Mizuho-cho, Mizuho-ku, Nagoya, Aichi 467-8501, Japan

[3] Center for Advanced Intelligence Project, RIKEN, Nihonbashi 1-chome Mitsui Building, 15th floor, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

[4] Graduate School of Information Science and Technology, The University of Tokyo, 7-3-1 Hongo, Bunkyo, Tokyo 1138656, Japan