On the Perception of Difficulty: Differences between Humans and AI

Philipp Spitzer[1,*,†], Joshua Holstein[1,†], Michael Vössing[1] and Niklas Kühl[2]

[1]Karlsruhe Institute of Technology, Kaiserstraße 89-93, Karlsruhe, 76133, Germany
[2]University of Bayreuth, Wittelsbacherring 10, Bayreuth, 95444, Germany

Abstract

With the increased adoption of artificial intelligence (AI) in industry and society, effective human-AI interaction systems are becoming increasingly important. A central challenge in the interaction of humans with AI is the estimation of difficulty for human and AI agents for single task instances. These estimations are crucial to evaluate each agent's capabilities and, thus, required to facilitate effective collaboration. So far, research in the field of human-AI interaction estimates the perceived difficulty of humans and AI independently from each other. However, the effective interaction of human and AI agents depends on metrics that accurately reflect each agent's perceived difficulty in achieving valuable outcomes. Research to date has not yet adequately examined the differences in the perceived difficulty of humans and AI. Thus, this work reviews recent research on the perceived difficulty in human-AI interaction and contributing factors to consistently compare each agent's perceived difficulty, e.g., creating the same prerequisites. Furthermore, we present an experimental design to thoroughly examine the perceived difficulty of both agents and contribute to a better understanding of the design of such systems.

Keywords

Artificial Intelligence, Human-AI Interaction, Confidence Estimation, Instance Difficulty

1. Introduction

In recent decades, technological advances have led to artificial intelligence (AI) applications becoming part of our everyday lives, e.g., when learning a new language [1] or driving autonomous cars [2]. Like many other examples of human-AI interaction, it comes down to appropriately assessing the difficulty of different situations for each agent (human and AI). The consequences for incorrect estimates can range from rejecting such systems, e.g., when the human learner is given too difficult words or grammar without being ready, to potentially severe consequences, e.g., autonomously driving cars on a foggy night. Consequently, it is necessary to estimate each agent's difficulty for an instance adequately.

Further examples of human-AI interaction that draw from an estimation of instance difficulty are *human-AI complementarity* [3–11], *curriculum learning* [12–14], and *machine teaching* [12, 13, 15–18]. Accurately assessing the difficulty of single instances for both human and AI agents is central to developing these forms of human-AI interaction to fully exploit their complementary capabilities while creating pleasant automation experiences.

By reviewing related literature, we observe different methods and terms, most prominently *uncertainty*, *confidence*, *performance* (e.g., in [19]), for measuring the difficulty of human and AI agents, which is why we aim to delimit our research in the following and create a shared understanding of the relevant terms. Before diving into the frequently used methods, we elaborate on the commonly used terms to describe the difficulty. *Performance* represents the aggregated accuracy over multiple instances for a task or over multiple agents for an instance [10, 19].

Further, *confidence* and *uncertainty* are often used interchangeably [20] and serve as a proxy for the difficulty of an instance. However, Pouget et al. [20] argues that these notions are not synonyms. Instead, uncertainty describes the distribution of probabilities over all possible outcomes, while confidence represents the probability that a particular decision is correct. When it comes to *difficulty*, one must differentiate between *objective difficulty* and *perceived difficulty*. The former, for instance, can be measured by comparing the number of features of a given task [21]. For the *perceived difficulty*, one must distinguish between task and instance difficulty. On the task level, a common method for human and AI agents depicts the usage of the average performance over multiple instances (for example, in Hemmer et al. [8], Geirhos et al. [22]) to determine the perceived difficulty.

However, on an instance level, the perceived difficulty of human and AI agents is assessed differently. First, a potential issue arises from an existing gap in access to relevant information. Usually, the AI agent is trained and tested on data drawn from the same distribution, thus having information on the label distribution. However,

AutomationXP23: Intervening, Teaming, Delegating - Creating Engaging Automation Experiences, CHI '23, April 23rd, Hamburg, Germany
*Corresponding author.
†These authors contributed equally.
Philipp.Spitzer@kit.edu (P. Spitzer); Joshua.Holstein@kit.edu (J. Holstein); Michael.Voessing@kit.edu (M. Vössing); Kuehl@uni-bayreuth.de (N. Kühl)

CEUR Workshop Proceedings (CEUR-WS.org)

this is often not the case for humans (e.g., in [10, 22, 23]). Therefore, it remains unclear whether and how, amongst others, this affects humans' perception of difficulty. Second, the difficulty of single instances is assessed differently. For AI agents, the distribution of the softmax outputs is often used to determine its uncertainty [10, 22]. Contrarily, the human's perceived instance difficulty is often measured by observing the distribution of predictions over groups of humans for single instances or by their average performance for an instance [19, 23]. Consequently, individual skills and capabilities of humans are neglected, potentially resulting in poor experiences in human-AI interaction settings [24]. As related literature shows, humans have distinct cognitive styles which can affect their perceived difficulty [25]. Hence, neglecting their individual traits and generalizing their predictions to determine the perceived difficulty can result in poor estimation for individuals.

As we observe inconsistencies in the measurement of the perceived difficulty between human and AI agents, we outline existing metrics to measure their perceived difficulty as a first step. Moreover, we scrutinize methods to compare both agents adequately. Based upon this, we are interested in adequately examining the difference in the perceived difficulty between humans and AI. Therefore, we state the following research question:

RQ: What are the differences in the perceived difficulty of humans and AI for single instances?

To answer this research question, we conduct a literature review to evaluate existing research fields relying on an accurate measurement of the perceived instance difficulty. Furthermore, we present an experimental design that avoids the previously mentioned inconsistencies. Through our experiment, we want to analyze the perceived difficulty of human and AI agents for single instances, using established metrics like confidence [10] and PVI [26] adequately. We support our endeavor to establish adequate methods to consistently measure the perceived instance difficulty of human and AI agents with first empirical results based on an existing, public dataset. Overall, with our experiment, we aim to contribute to a better and more integrated understanding of how to adequately compare human and AI agents' perceived difficulty leading to a thorough understanding of the design of human-AI interaction systems.

2. Related Work

2.1. Human-AI Interaction and Instance Difficulty

With the latest ascent in research on human-AI interaction, the deployment of AI in automated systems is advancing [27, 28]. Hereby, various forms of human-AI interaction rely on estimating an instance's difficulty for effective collaboration. Following, we outline the three forms of human-AI interaction most relevant to our research: human-AI complementarity, curriculum learning, and machine teaching.

In the field of *human-AI complementarity*, recent research studies complementary team performance—exceeding the performance each agent (human or AI) can achieve on their own [3, 5]. In this collaboration, it is crucial to properly delegate tasks to each agent to exploit their complementary capabilities [9]. Steyvers et al. [10] establish a framework to facilitate both human and AI agents' confidence scores to investigate factors that influence complementary capabilities of human-AI collaborations. Lai et al. [11] suggest using uncertainty as a measure to delegate tasks between human and AI agents. In the work of Fügener et al. [29], the authors evaluate different delegation strategies based on the performance of both agents for single instances. They find that humans' perception of task difficulty differs from the actual task difficulty. Lubars and Tan [6] investigate, amongst others, the effect of the difficulty of single instances to delegate tasks.

Curriculum learning denotes another form of human-AI interaction in which the perceived difficulty is relevant to the overall process. This form of learning is based on human learning and incorporates the idea that the order is crucial in which training instances are presented to a learner [12]. A central aspect of curriculum learning is the assertion of difficulty levels of single instances. Wei et al. [13] use the annotator agreement in an image classification task to determine the difficulty of instances.

In the field of *machine teaching*, a human or an AI agent is trained by selecting samples to achieve high learning outcomes [15]. The selection of training instances can be grounded on difficulty estimation. For example, Zhang et al. [16] presents an interactive learning procedure in which crowd workers are trained based on an approximated difficulty for instances. Similarly, Singla et al. [17] select training instances for learners based on an expected uncertainty measured by an AI agent.

2.2. Measuring Perceived Difficulty of Humans and AI

AI's perceived difficulty. In Ståhl et al. [30], the authors evaluate different metrics to compare the uncertainty of deep learning models. One of these metrics is a Bayesian network-based approach using dropout [31]. Further, Xu et al. [32] present a metric that builds on Shannon entropy [33] to compare the difficulty of different datasets. Moreover, Ethayarajh et al. [26] extend this metric, called $\mathcal{V}$-usable information, to apply it to single instances. This metric, the pointwise $\mathcal{V}$-information (PVI), is used to

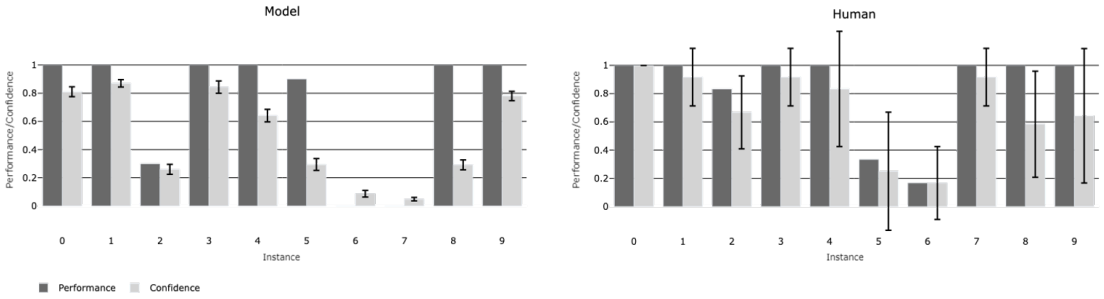

Figure 1: Comparison of performance and confidence for single instances (top: fine-tuned model, bottom: human).

compare the difficulty of single instances with respect to a model family $\mathscr{V}$. According to the authors, PVI, in contrast to related metrics, quantifies the difficulty of single instances accounting for how much information can be extracted beyond the label distribution.

Human's perceived difficulty. Most works focus on estimating the perceived difficulty of humans by aggregating over multiple humans. For example, Peterson et al. [23] asses the disagreement of two decision-makers. In their work, the authors define the difficulty of a single instance by using the disagreement of crowdsourcing annotators. To measure the individual perceived difficulty of instances, Steyvers et al. [10] use a different approach. The authors use the ordinal responses of humans to determine their confidence. Similarly, Bıyık et al. [34] determine human difficulty by asking participants about their perceived task difficulty.

3. Empirical Validation Using Public Datasets

Before our experiment, we examine reports of other studies to investigate the differences in the perceived difficulty of single instances. Therefore, we utilize publicly available datasets, e.g., CIFAR10-H [23], modelvshuman [22], or ImageNet-16H [10]. However, the first two datasets, CIFAR10-H, and modelvshuman do not contain individual human confidence or uncertainty measurements. Instead, the authors of the datasets [22, 23] estimate the instance difficulty by aggregating the performance of multiple human annotators for instances. ImageNet-16H is the only dataset containing human difficulty measurements in the form of self-reported confidence levels, e.g., *low*, *medium*, and *high*. To compare these reported confidence levels with the commonly used technique of average instance performance, we transformed the confidence levels to 0 (low), 0.5 (medium), and 1 (high).

Further, we fine-tune an efficientnet model with the dataset for two epochs and use Monte-Carlo Dropout to receive the perceived confidence of the AI agent. Finally, with the confidence of human and AI agents available, we compare their performance and confidence for single instances.

Figure 1 illustrates and compares instance performance and confidence for ten randomly sampled instances of the ImageNet-16H dataset. The left part represents the AI agent's output, while the right part shows the human's self-reported confidence. Based on this, we can make several observations. First, task performance is not necessarily a reliable factor to determine the perceived difficulty of an instance. For example, instances seven to nine have the same performance but differ greatly in their reported confidence. Second, human and AI agents can perceive different instances as easy, e.g., the AI agent has low confidence for instances seven and eight, while the humans have medium to high confidence. Third, the human self-reported confidence scores differ among participants, as can be seen from the standard deviation of confidence. We argue that these observations represent first evidence in the direction of our hypotheses. More specifically, we can see that the average performance of an instance cannot be used to determine the perceived difficulty of an instance for individual humans. Instead, other metrics need to be considered.

Moreover, the high standard deviation of human confidence for almost all instances indicates that humans differ in their perceived difficulty. Consequently, the diversity of humans must be taken into consideration when designing human-AI interaction systems.

4. Experimental Design

Our experiment is based on a mixed-effects model that combines a between-subject and a within-subject design [35]. We follow the notion of existing works and understand confidence as a proxy for difficulty [36]. More precisely, we measure the difficulty of the human and the AI agent by two metrics: the commonly used confidence [10] and the PVI score [26] as a novel metric that considers the label distribution. We measure the confi-

dence of AI agents by Monte-Carlo Dropout [31] and for humans via probabilities, e.g., using a scale between 0% and 100%. We use a binary classification task to avoid participants having to assign multiple probabilities. The binary classification allows us to observe one probability, e.g., an image showing a cat with a probability of 80%, and calculate the complementary probability, e.g., the complementary probability that the image does not represent a cat is 20%.

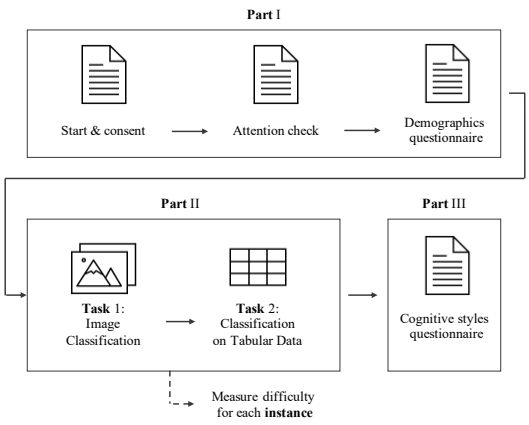

Figure 2: Design of the study.

The preliminary experiment design is illustrated in Figure 2. The experiment is composed of three parts. *Part I* includes consent, instructions, and a demographics questionnaire. Next, *Part II* comprises two binary classification tasks—one visual and one textual—and, finally, *Part III* is a questionnaire on cognitive styles. In both tasks, we measure the perceived difficulty of participants and AI for single instances.

In our experiment, we have two treatments. First, as we want a consistent comparison of the perceived difficulty between humans and AI, we must ensure they have access to the relevant information. However, in contrast to humans, the AI agent has access to the label distribution through its training prior to the task. As we want to examine this effect, we show humans the label distribution before conducting the task in one condition. Thus, we hypothesize:

Hypothesis 1. *Access to the information on label distribution has an impact on humans' perceived difficulty of single instances.*

After providing a consistent way to measure the confidence—as a proxy for the perceived difficulty—of the human and the AI agent, we want to examine the differences in their perceived difficulty of instances. Previous research identified subsets of data on which either human or AI agent has a better performance, e.g., [22]. As the performance of human and AI agents is a consequence of the probabilities they assign to each class and, thus, their uncertainty, we argue that the perceived difficulty for an instance can differ even for instances both agents have classified the same. Thus, we hypothesize:

Hypothesis 2. *There are instances for which human and AI agents make the same prediction but differ in their perceived difficulty.*

Within our experiment, we leverage two datasets for the tasks of Part II to compare the perceived difficulty of human and AI agents. Both conditions comprise the same tasks. We chose two different tasks: one visual classification task and one based on tabular data. Research shows the impact of different cognitive styles on participants' task performance (i.e., [25, 37, 38]). By choosing a visual and a text-based task, we account for participants' different cognitive styles and individual perceptions of difficulty. Accordingly, participants will be asked to conduct a questionnaire in which we determine their cognitive styles.We assess these styles by using the validated items of Kirby et al. [37] (initially presented by Richardson [39]). The items of the cognitive style questionnaire are randomly arranged as suggested by Kirby et al. [37]. All items are measured on a five-point Likert scale. We hypothesize:

Hypothesis 3. *Humans with distinct cognitive styles perceive the difficulty of single instances differently.*

5. Discussion

In this work, we propose an experimental design to investigate the difference in perceived difficulty between human and AI agents for single instances. To build a foundation, we assess related work and common metrics to estimate instance difficulty. Yet, these studies insufficiently scrutinize consistent difficulty estimations between humans and AI. By first examining a related dataset, we show the discrepancies in difficulty estimation by applying conventional approaches. Thus, we propose an experiment design that paves the way for a broad main study in which we: (I) Develop a consistent way to measure the perceived difficulty of instances, (II) Examine the differences in the perceived difficulty of human and AI agents, (III) Investigate a potential cause in varying perceived difficulty of humans.

Through our main study, we expect to contribute to the ongoing discussion on developing automated and reliable AI agents interacting with humans with diverse skills and capabilities. Moreover, our results will provide guidance not only in research but also in practice on designing human-AI interaction systems. A promising field of research lies ahead.

References

[1] S. Pokrivčáková, Preparing teachers for the application of ai-powered technologies in foreign language education, Journal of Language and Cultural Education (2019).

[2] M. Johns, B. Mok, D. Sirkin, N. Gowda, C. Smith, W. Talamonti, W. Ju, Exploring shared control in automated driving, in: 2016 11th ACM/IEEE International Conference on Human-Robot Interaction (HRI), IEEE, 2016, pp. 91–98.

[3] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, 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, 2021, pp. 1–16.

[4] P. Hemmer, M. Schemmer, M. Vössing, N. Kühl, Human-ai complementarity in hybrid intelligence systems: A structured literature review., PACIS (2021) 78.

[5] M. Schemmer, N. Kühl, C. Benz, G. Satzger, On the influence of explainable ai on automation bias, arXiv preprint arXiv:2204.08859 (2022).

[6] B. Lubars, C. Tan, Ask not what ai can do, but what ai should do: Towards a framework of task delegability, Advances in Neural Information Processing Systems 32 (2019).

[7] J. M. Wing, Trustworthy ai, Communications of the ACM 64 (2021) 64–71.

[8] P. Hemmer, M. Schemmer, N. Kühl, M. Vössing, G. Satzger, On the effect of information asymmetry in human-ai teams, arXiv e-prints (2022) arXiv–2205.

[9] A. Fügener, J. Grahl, A. Gupta, W. Ketter, Cognitive challenges in human–artificial intelligence collaboration: Investigating the path toward productive delegation, Information Systems Research 33 (2022) 678–696.

[10] M. Steyvers, H. Tejeda, G. Kerrigan, P. Smyth, Bayesian modeling of human–ai complementarity, Proceedings of the National Academy of Sciences 119 (2022) e2111547119.

[11] V. Lai, S. Carton, R. Bhatnagar, Q. V. Liao, Y. Zhang, C. Tan, Human-ai collaboration via conditional delegation: A case study of content moderation, in: Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems, 2022, pp. 1–18.

[12] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.

[13] J. Wei, A. Suriawinata, B. Ren, X. Liu, M. Lisovsky, L. Vaickus, C. Brown, M. Baker, M. Nasir-Moin, N. Tomita, et al., Learn like a pathologist: curriculum learning by annotator agreement for histopathology image classification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2021, pp. 2473–2483.

[14] R. Lotfian, C. Busso, Curriculum learning for speech emotion recognition from crowdsourced labels, IEEE/ACM Transactions on Audio, Speech, and Language Processing 27 (2019) 815–826.

[15] X. Zhu, Machine teaching: An inverse problem to machine learning and an approach toward optimal education, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 29, 2015, pp. 4083–4087.

[16] J. Zhang, H. Wang, S. Meng, V. S. Sheng, Interactive learning with proactive cognition enhancement for crowd workers, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 540–547.

[17] A. Singla, I. Bogunovic, G. Bartók, A. Karbasi, A. Krause, Near-optimally teaching the crowd to classify, in: International Conference on Machine Learning, PMLR, 2014, pp. 154–162.

[18] P. Spitzer, N. Kühl, M. Goutier, Training novices: The role of human-ai collaboration and knowledge transfer, arXiv preprint arXiv:2207.00497 (2022).

[19] A. Taudien, A. Fügener, A. Gupta, W. Ketter, The effect of ai advice on human confidence in decision-making, in: Proceedings of the 55th Hawaii International Conference on System Sciences, 2022.

[20] A. Pouget, J. Drugowitsch, A. Kepecs, Confidence and certainty: distinct probabilistic quantities for different goals, Nature neuroscience 19 (2016) 366–374.

[21] J. G. Nicholls, A. T. Miller, The differentiation of the concepts of difficulty and ability, Child development (1983) 951–959.

[22] R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F. A. Wichmann, W. Brendel, Partial success in closing the gap between human and machine vision, in: Advances in Neural Information Processing Systems 34, 2021.

[23] J. C. Peterson, R. M. Battleday, T. L. Griffiths, O. Russakovsky, Human uncertainty makes classification more robust, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2019, pp. 9617–9626.

[24] M. Szymanski, M. Millecamp, K. Verbert, Visual, textual or hybrid: the effect of user expertise on different explanations, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 109–119.

[25] D. M. Broverman, Dimensions of cognitive style., Journal of Personality (1960).

[26] K. Ethayarajh, Y. Choi, S. Swayamdipta, Understanding dataset difficulty with *mathcalv*-usable information, in: International Conference on Ma-

chine Learning, PMLR, 2022, pp. 5988–6008.
[27] E. Barboni, J.-F. Ladry, D. Navarre, P. Palanque, M. Winckler, Beyond modelling: an integrated environment supporting co-execution of tasks and systems models, in: Proceedings of the 2nd ACM SIGCHI symposium on Engineering interactive computing systems, 2010, pp. 165–174.
[28] V. Roto, P. Palanque, H. Karvonen, Engaging automation at work–a literature review, in: Human Work Interaction Design. Designing Engaging Automation: 5th IFIP WG 13.6 Working Conference, HWID 2018, Espoo, Finland, August 20-21, 2018, Revised Selected Papers 5, Springer, 2019, pp. 158–172.
[29] A. Fügener, J. Grahl, A. Gupta, W. Ketter, Collaboration and delegation between humans and AI: An experimental investigation of the future of work, Erasmus Research Institute of Management (ERIM), 2019.
[30] N. Ståhl, G. Falkman, A. Karlsson, G. Mathiason, Evaluation of uncertainty quantification in deep learning, in: Information Processing and Management of Uncertainty in Knowledge-Based Systems: 18th International Conference, IPMU 2020, Lisbon, Portugal, June 15–19, 2020, Proceedings, Part I 18, Springer, 2020, pp. 556–568.
[31] Y. Gal, Z. Ghahramani, Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR, 2016, pp. 1050–1059.
[32] Y. Xu, S. Zhao, J. Song, R. Stewart, S. Ermon, A theory of usable information under computational constraints, arXiv preprint arXiv:2002.10689 (2020).
[33] C. E. Shannon, A mathematical theory of communication, ACM SIGMOBILE mobile computing and communications review 5 (2001) 3–55.
[34] E. Bıyık, M. Palan, N. C. Landolfi, D. P. Losey, D. Sadigh, Asking easy questions: A user-friendly approach to active reward learning, arXiv preprint arXiv:1910.04365 (2019).
[35] L. Riefle, C. Benz, T. Tomar, "may i help you?": Exploring the effect of individuals' self-efficacy on the use of conversational agents, ICIS 2022 Proceedings (2022).
[36] B. Kompa, J. Snoek, A. L. Beam, Second opinion needed: communicating uncertainty in medical machine learning, NPJ Digital Medicine 4 (2021) 4.
[37] J. R. Kirby, P. J. Moore, N. J. Schofield, Verbal and visual learning styles, Contemporary educational psychology 13 (1988) 169–184.
[38] L. Riefle, P. Hemmer, C. Benz, M. Vössing, J. Pries, On the influence of cognitive styles on users' understanding of explanations, ICIS 2022 Proceedings (2022).
[39] A. Richardson, Verbalizer-visualizer: a cognitive style dimension., Journal of mental imagery (1977).