Know What Not To Know: Users' Perception of Abstaining Classifiers

Andrea Papenmeier
andrea.papenmeier@uni-due.de
University of Duisburg-Essen
Duisburg, Germany
University of Twente
Enschede, Netherlands

Daniel Hienert
daniel.hienert@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

Yvonne Kammerer
kammerer@hdm-stuttgart.de
Stuttgart Media University
Stuttgart, Germany

Christin Seifert
christin.seifert@uni-due.de
University of Duisburg-Essen
Essen, Germany
University of Marburg
Marburg, Germany

Dagmar Kern
dagmar.kern@gesis.org
GESIS – Leibniz Institute for the Social Sciences
Cologne, Germany

ABSTRACT

Machine learning systems can help humans to make decisions by providing decision suggestions (i.e., a label for a datapoint). However, individual datapoints do not always provide enough clear evidence to make confident suggestions. Although methods exist that enable systems to identify those datapoints and subsequently abstain from suggesting a label, it remains unclear how users would react to such system behavior. This paper presents first findings from a user study on systems that do or do not abstain from labeling ambiguous datapoints. Our results show that label suggestions on ambiguous datapoints bear a high risk of unconsciously influencing the users' decisions, even toward incorrect ones. Furthermore, participants perceived a system that abstains from labeling uncertain datapoints as equally competent and trustworthy as a system that delivers label suggestions for all datapoints. Consequently, if abstaining does not impair a system's credibility, it can be a useful mechanism to increase decision quality.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; **Human computer interaction (HCI)**.

KEYWORDS

Human-Centered Machine Learning; Perception; Abstaining

1 INTRODUCTION

Decision-support systems (DSS) leverage the computational complexity of machine learning to support users in decision tasks in several domains, e.g., for medical diagnosis [9] or shopping [7]. In many use cases, decisions must be made for uncertain or ambiguous datapoints [1, 13] – datapoints for which a system might not be able to make a confident decision suggestion. There are automatic methods to detect ambiguous datapoints [13, 15]. Like humans saying "I don't know", the system can make decisions when it is certain but defer uncertain datapoints to a human annotator. In decision-support tasks, however, all datapoints are typically shown to the user alongside the DSS's suggestion for a decision. Users tend to rely on DSS's suggestions, especially for ambiguous datapoints [17, 20], which might lead users toward incorrect decisions. As an alternative, abstaining systems do not deliver a suggestion on highly uncertain datapoints [4]. Although methods for equipping a DSS with an abstaining mechanism exist, we do not know how such behavior affects how users perceive the system. This led us to the following research question: *How is the users' perception of a DSS influenced by a system that abstains from offering support on ambiguous datapoints?*

To examine this research question, participants of our user study performed a labeling task with the help of a DSS. We varied the DSS's behavior to either abstain or not abstain from suggesting a label for ambiguous datapoints. Our findings show that users are, often unconsciously, influenced by a system's label suggestion on ambiguous datapoints. An abstaining system does not provide label suggestions on ambiguous datapoints and therefore cannot lead the user toward a wrong decision in those cases. Although an abstaining system explicitly discloses the boundaries of its capabilities to the user, our results suggest that it does not impair perceived system performance or credibility. Our research provides first insights from the users' perspective into abstaining as a mechanism for DSSs to deal with ambiguous datapoints.

2 RELATED WORK

The "I don't know" (IDK) mechanism first appeared in the 1990s in the high-risk domain of anomaly detection in power plants [1]. Usually operating without human interference, these systems delegated uncertain cases to human operators. Contrarily, decision-support systems (DSS) predict a decision (e.g., a label in an annotation task) and suggest the prediction to the user. Mol et al. [15] argued that these systems need a safeguard mechanism to prevent displaying uncertain suggestions. Literature provides several approaches for "rejecting" individual datapoints [6, 13], i.e., classifying them as IDK and subsequently abstaining from labeling the instance. Ambiguous datapoints that fall in the IDK category can be excluded from the dataset during training [18] to increase dataset consistency or during testing to increase accuracy [12, 19]. However, in practice, rejected datapoints still need to be handled, e.g., by being

deferred to a human expert [14], possibly with an explanation of why the sample was rejected [22]. However, all works presented above evaluated their algorithms in offline experiments without users.

A DSS helps users make decisions, e.g., in labeling or classification tasks. However, literature shows that people frequently overrely on DSSs and follow incorrect suggestions, although they would have made a correct decision on their own [3, 8]. Especially ambiguous decisions provoke this behavior: Wang et al. found that users' reliance on a DSS increases with reduced confidence in their own decision [20] and Papenmeier et al. [17] mentioned that users might agree with a system when in doubt. Gandouz et al. argue that an abstaining system, i.e., a system that refrains from showing uncertain suggestions, *"better reflects human decision-making"* [6]. Other works similarly call for systems that communicate the boundaries of their capabilities to users [10]. A system that uses abstaining to deal with ambiguous datapoints would adhere to this design guideline. Yet, if a system shows high levels of uncertainty, a DSS's perceived credibility (i.e., competence and trustworthiness [5, 21]), which is an important aspect of users' perception [5], might be impaired [2]. So far, the immediate effects of a system that explicitly and visibly abstains from giving decision suggestions for ambiguous datapoints have not been investigated from the user's perspective.

3 METHOD

As outlined above, DSSs support users in decision tasks by displaying a suggestion for a decision, e.g., labels in a labeling task. However, ambiguous datapoints are difficult to classify because they provide ambiguous (or not enough) evidence for a clear label decision. The DSS might then be unable to provide reliable suggestions. In those cases, an abstaining DSS could reject the sample, i.e., abstain from making a suggestion. To answer our research question, we assessed users' perception of a DSS's performance and credibility after interacting with it in a text labeling task. We focused on the users' perception and used a fictive DSS with simulated output to be able to control its behavior. We employed a between-subjects design and simulated three DSSs, leading to the following conditions:

- **C1** **Correct**: The fictive DSS displays correct label suggestions, i.e., the ground truth, on ambiguous cases.
- **C2** **Abstain**: The fictive DSS does not provide a label suggestion (abstains) on ambiguous cases.
- **C3** **Wrong**: The fictive DSS displays incorrect label suggestions on ambiguous cases.

All three DSSs displayed correct label suggestions for unambiguous cases, which is possible as we control the DSSs' outputs in our setup. The study received clearance from the ethics board of the first author's institution. All materials used in the study (questionnaire, dataset, and anonymized responses) are available online[1].

3.1 Use Case and Dataset

For analyzing users' perception of a DSS in a labeling task, we needed a dataset with ambiguous datapoints. In their work on human perception of classification mistakes, Papenmeier et al. [17] published a dataset with 50 phrases that are either "easy", "difficult", or "impossible" to label for human annotators. The phrases were taken from descriptions formulated by users who described either their wishes for a new laptop or a jacket. For example, the jacket phrase *"a budget one that is durable enough to last a long time"* was identified as "difficult" to label. At the same time, the ground truth labels were collected along with the descriptions. That is, the ground truth labels were not retrospectively annotated and are therefore not affected by subjective interpretations of annotators.

To identify those phrases from the dataset that a DSS should abstain from, we asked 30 crowd workers from Prolific[2] (native English speakers, UK residents, no literacy problems) to annotate the phrases. We offered the label options "laptop", "jacket", and a residual option "don't know / unsure". We reduced the dataset to 30 phrases to avoid fatigue and boredom effects: Based on the annotations of the crowd workers, we selected 10 laptop and 10 jacket phrases that were unambiguous, i.e., that were correctly annotated by all crowd workers (e.g., *"i would need one with lots of memory and ram available to store large file formats"*). Additionally, we selected 5 laptop and 5 jacket phrases that crowd workers found ambiguous, i.e., that had the highest percentage of "don't know / unsure" annotations (e.g., *"quality is most important to me no matter what brand"*, with 63% of the crowd workers having selected "don't know / unsure").

3.2 Task and Procedure

First, participants gave informed consent for participation and provided demographic data (age, gender). They then read the scenario that framed the text labeling task as part of a high-quality dataset collection in collaboration with a retail company. The DSS was introduced as a software provided by the retail company. In C2 (abstaining on ambiguous phrases), we introduced the abstaining mechanism as a result of the software not reaching a final decision for some phrases. Participants were then instructed to label the phrases and read the software's suggestions to provide feedback on its performance after the task. Following a training phase with two phrases, participants entered the main phase with 30 phrases (see Figure 1) in random order, including two attention checks. Subsequently, participants completed a post-task questionnaire about the perceived performance and credibility of the DSS.

3.3 Measures and Analysis

In this experiment, we measured the following three dependent variables:

User performance: We calculated participants' labeling accuracy on ambiguous phrases w.r.t. the ground truth.

Perceived performance: Participants rated the DSS's overall performance after the task on a 7-point scale ("1" = very poor to "7" = excellent).

Perceived credibility: Participants rated the DSS's credibility after the task on 7-point semantic differentials for competence (unqualified - qualified, inexperienced - experienced, incompetent - competent, Cronbach's α = .91) and trustworthiness (dishonest - honest, untrustworthy - trustworthy, Cronbach's α = .86).

We further gathered qualitative insights into participants' decision-making process: In all three conditions (correct, abstaining, wrong

[1] https://git.gesis.org/papenmaa/dis23_perceptionofabstaining

[2] https://www.prolific.co

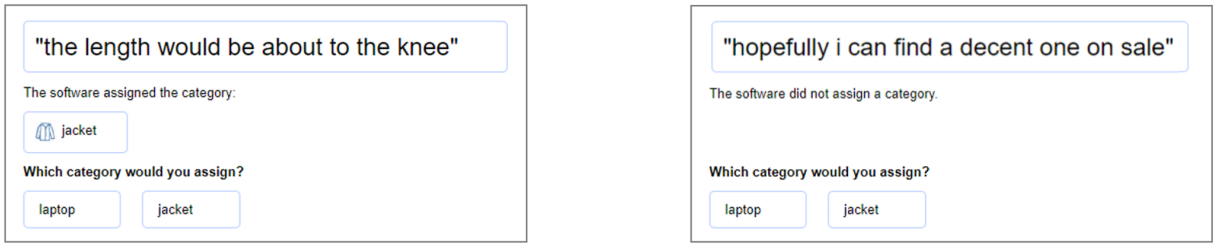

Figure 1: Examples of the task: clear phrase with label suggestion (left) and ambiguous phrase with abstaining system (right).

label suggestions), participants also described how the suggestions influenced their labeling behavior. We performed one-way ANOVAs with two-tailed Mann-Whitney U-Tests for post-hoc comparisons to compare the three groups regarding the dependent variables. We used a significance level of $\alpha = 0.05$ (two-sided) and report p-values after Bonferroni correction to counteract the repeated testing bias.

3.4 Participants

We recruited N = 120 participants on Prolific (English native speakers, UK residents, no literacy difficulties). All participants received a financial allowance of 1.20 GBP (7.20 GBP/h). Twelve responses were excluded from the analysis due to failed attention checks. The 108 participants with valid responses were, on average, M = 43.4 years old (SD = 13.3 years) and approximately balanced between male and female gender (55 female, 52 male, 1 non-binary). Participants in the three conditions did not differ regarding age ($F(2,105) = 0.028$, $p = 0.972$) or gender distribution ($\chi^2(2) = 0.050, p = 0.975$).

4 RESULTS

Users frequently adopt incorrect suggestions of a DSS [3, 8], especially on ambiguous datapoints [17, 20]. In those cases, an abstaining DSS would lead to better outcomes. To confirm that this behavior is also present in our use case of labeling jacket and laptop sentences, we investigated how a DSS influences users' labeling performance (i.e., how often users chose the correct label). Table 1 shows the mean performances, i.e., how often participants chose the correct label on average. We define the "correct" label as the ground truth label from the dataset. The ANOVA showed a significant main effect. The post-hoc test results showed that providing wrong label suggestions led to a significantly poorer label performance than abstaining (C3 vs. C2: $U = 231.0$, $p < .001$) or providing correct label suggestions (C3 vs. C1: $U = 135.5$, $p < .001$). The results did not show a difference between C1 and C2 ($U = 476.0$, $p = .078$). We also asked participants how the system influenced their decisions. Although the label performance is significantly different in C1 than in C3, many participants reported that the label suggestions did not influence their decisions (47% in C1, 56% in C3), e.g., *"It didn't. I used my own judgement of the phrases to decide"* or *"Didn't really influence at all, I went with what I thought about the phras"*.

To understand how users perceive an abstaining system, we investigated the perceived performance and perceived credibility (via competence and trustworthiness) in all three conditions. Table 1 (right) presents the mean ratings of the three perception variables. The ANOVA did not show a significant main effect for any of the three variables.

5 DISCUSSION AND CONCLUSION

In this experiment, we investigated how users perceive different behaviors of decision-support systems (DSS) that either deliver correct (C1), incorrect (C3), or no decision suggestions (C2) for ambiguous datapoints. Our findings show that participants were strongly influenced by the DSS's label suggestion on ambiguous datapoints and often selected the same label as the system. However, participants were, to a large extent, unaware of this influence as the qualitative data reveals.

The difference in user performance between the abstaining system (C2) and the correct labeling system (C1) was small (7% increase) compared to the difference between abstaining (C2) and wrong suggestions (C3) (27% decrease). That is, in our use case, the risk of steering participants toward wrong decisions was higher than the small gain of providing correct labels. As machine learning systems are likely to perform worse on ambiguous datapoints than on clear datapoints (see [12, 19]), abstaining from suggesting a decision for ambiguous datapoints could be advantageous. In the medical domain, for example, a DSS could recommend further tests instead of an uncertain diagnosis when patient data is inconclusive. However, future work should investigate why wrong suggestions had a stronger influence on users than correct suggestions.

A potential downside of an abstaining system could be that users perceive a system as less knowledgeable and less trustworthy when it cannot perform the task it was built for (in our case, provide label suggestions). However, our findings do not support this. There was no difference in perceived system performance, perceived competence, or perceived trustworthiness across conditions. As many real-life datasets contain ambiguity, systems need to be prepared to work with ambiguous data. Our findings suggest that abstaining is a potential mechanism to deal with ambiguous datapoints. Therefore, in our future work, we will take a closer look into abstaining mechanisms from the users' perspective and develop design guidelines for abstaining interactive systems.

In our user study, we used a dataset with clearly ambiguous and clearly unambiguous datapoints and employed fictive DSSs. In reality, datapoints might span the full bandwidth of ambiguity. Practitioners need to identify ambiguous datapoints, for which several methods have been proposed previously [6, 11, 12, 16]. However, as those methods might introduce additional mistakes and DSSs might make classification mistakes, we want to investigate in the future how users perceive incorrect abstaining behavior, e.g., when the

Table 1: Left: Mean user performance on *IDK* phrases in percent with respect to the ground truth. Right: Mean ratings of perceived system performance and credibility and comparison of means with one-way ANOVAs.

	User Performance			Perceived Performance		Perceived Competence		Perceived Trustworthiness	
	M	SD		M	SD	M	SD	M	SD
C1 Correct	76%	16%	**C1**	6.00	1.11	5.75	1.22	5.78	1.33
C2 Abstain	69%	15%	**C2**	5.89	0.85	5.58	1.01	5.72	0.97
C3 Wrong	42%	21%	**C3**	5.75	0.79	5.75	0.96	5.88	0.89
ANOVA	$F(2,105) = 35.13$, $p < .001$		ANOVA	$F(2,105) = 0.629$, $p = .535$		$F(2,105) = 0.301$, $p = .741$		$F(2,105) = 0.181$, $p = .834$	

system abstains from labeling clear datapoints. Moreover, we want to test different ways of communicating the reasons for abstaining, e.g., with explanations or the classifier's confidence score.

REFERENCES

[1] Yair Bartal, Jie Lin, and Robert E Uhrig. 1994. Nuclear power plants transient diagnostics using LVQ or some networks don't know that they don't know. In *Proceedings of 1994 IEEE International Conference on Neural Networks (ICNN'94)*, Vol. 6. IEEE, New York, USA, 3744–3749.

[2] Umang Bhatt, Javier Antorán, Yunfeng Zhang, Q. Vera Liao, Prasanna Sattigeri, Riccardo Fogliato, Gabrielle Melançon, Ranganath Krishnan, Jason Stanley, Omesh Tickoo, Lama Nachman, Rumi Chunara, Madhulika Srikumar, Adrian Weller, and Alice Xiang. 2021. Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (AIES '21)*. ACM, New York, USA, 401–413. https://doi.org/10.1145/3461702.3462571

[3] Adrian Bussone, Simone Stumpf, and Dympna O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *Proceedings of the 2015 International Conference on Healthcare Informatics (ICHI '15)*. IEEE Computer Society, USA, 160–169. https://doi.org/10.1109/ICHI.2015.26

[4] César Ferri and José Hernández-Orallo. 2004. Cautious Classifiers. In *1st workshop on ROC analysis in artificial intelligence (ROCAI-2004)*, José Hernández-Orallo, César Ferri, Nicolas Lachiche, and Peter A. Flach (Eds.). ACM, New York, USA, 27–36.

[5] B. J. Fogg and Hsiang Tseng. 1999. The Elements of Computer Credibility. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '99)*. ACM, New York, USA, 80–87. https://doi.org/10.1145/302979.303001

[6] Mariem Gandouz, Hajo Holzmann, and Dominik Heider. 2021. Machine learning with asymmetric abstention for biomedical decision-making. *BMC medical informatics and decision making* 21, 1 (2021), 1–11.

[7] Nico Herbig, Gerrit Kahl, and Antonio Krüger. 2018. Design Guidelines for Assistance Systems Supporting Sustainable Purchase Decisions. In *Proceedings of the 2018 Designing Interactive Systems Conference (DIS '18)*. ACM, New York, USA, 1333–1344. https://doi.org/10.1145/3196709.3196726

[8] Maia Jacobs, Melanie F Pradier, Thomas H McCoy, Roy H Perlis, Finale Doshi-Velez, and Krzysztof Z Gajos. 2021. How machine-learning recommendations influence clinician treatment selections: the example of antidepressant selection. *Translational psychiatry* 11, 1 (2021), 1–9.

[9] Annika Kaltenhauser, Verena Rheinstädter, Andreas Butz, and Dieter P. Wallach. 2020. "You Have to Piece the Puzzle Together": Implications for Designing Decision Support in Intensive Care. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference (DIS '20)*. ACM, New York, USA, 1509–1522. https://doi.org/10.1145/3357236.3395436

[10] Anna Kawakami, Venkatesh Sivaraman, Hao-Fei Cheng, Logan Stapleton, Yanghuidi Cheng, Diana Qing, Adam Perer, Zhiwei Steven Wu, Haiyi Zhu, and Kenneth Holstein. 2022. Improving Human-AI Partnerships in Child Welfare: Understanding Worker Practices, Challenges, and Desires for Algorithmic Decision Support. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, USA, Article 52, 18 pages. https://doi.org/10.1145/3491102.3517439

[11] Benjamin Kompa, Jasper Snoek, and Andrew L Beam. 2021. Second opinion needed: communicating uncertainty in medical machine learning. *NPJ Digital Medicine* 4, 1 (2021), 1–6.

[12] Alexey Kornaev and Elena Kornaeva. 2020. Room for doubt as a way to improve the accuracy of machine learning algorithms. In *2020 4th Scientific School on Dynamics of Complex Networks and their Application in Intellectual Robotics (DCNAIR)*. IEEE, New York, USA, 135–137.

[13] Max-Heinrich Laves, Sontje Ihler, and Tobias Ortmaier. 2019. Uncertainty Quantification in Computer-Aided Diagnosis: Make Your Model say" I don't know" for Ambiguous Cases. In *Medical Imaging with Deep Learning Conference (MIDL '19)*. MIDL, London, UK, 1–4. https://doi.org/10.48550/arXiv.1908.00792

[14] David Madras, Toni Pitassi, and Richard Zemel. 2018. Predict Responsibly: Improving Fairness and Accuracy by Learning to Defer. In *Advances in Neural Information Processing Systems*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.), Vol. 31. Curran Associates, Inc., Montréal, Canada, 1–11.

[15] Antônio C.de A. Mol, Aquilino S. Martinez, and Roberto Schirru. 2003. A neural model for transient identification in dynamic processes with "don't know" response. *Annals of Nuclear Energy* 30, 13 (2003), 1365–1381. https://doi.org/10.1016/S0306-4549(03)00072-0

[16] Allen Nie, Ashley Zehnder, Rodney L Page, Yuhui Zhang, Arturo Lopez Pineda, Manuel A Rivas, Carlos D Bustamante, and James Zou. 2018. DeepTag: inferring diagnoses from veterinary clinical notes. *NPJ digital medicine* 1 (2018), 60. https://doi.org/10.1038/s41746-018-0067-8

[17] Andrea Papenmeier, Dagmar Kern, Daniel Hienert, Yvonne Kammerer, and Christin Seifert. 2022. How Accurate Does It Feel? – Human Perception of Different Types of Classification Mistakes. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems (CHI '22)*. ACM, New York, USA, Article 180, 13 pages. https://doi.org/10.1145/3491102.3501915

[18] Sunil Thulasidasan, Tanmoy Bhattacharya, Jeff Bilmes, Gopinath Chennupati, and Jamal Mohd-Yusof. 2019. Combating Label Noise in Deep Learning using Abstention. In *Proceedings of the 36th International Conference on Machine Learning*, Kamalika Chaudhuri and Ruslan Salakhutdinov (Eds.), Vol. 97. PMLR, Long Beach, CA, USA, 6234–6243. https://proceedings.mlr.press/v97/thulasidasan19a.html

[19] Thomas P Trappenberg and Andrew D Back. 2000. A classification scheme for applications with ambiguous data. In *Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. Neural Computing: New Challenges and Perspectives for the New Millennium*, Vol. 6. IEEE, New York, USA, 296–301.

[20] Xinru Wang, Zhuoran Lu, and Ming Yin. 2022. Will You Accept the AI Recommendation? Predicting Human Behavior in AI-Assisted Decision Making. In *Proceedings of the ACM Web Conference 2022 (WWW '22)*. ACM, New York, USA, 1697–1708. https://doi.org/10.1145/3485447.3512240

[21] Stephan Winter and Nicole C. Krämer. 2014. A question of credibility – Effects of source cues and recommendations on information selection on news sites and blogs. *Communications* 39, 4 (2014), 435–456. https://doi.org/10.1515/commun-2014-0020

[22] Xiaoge Zhang, Felix T.S. Chan, and Sankaran Mahadevan. 2022. Explainable machine learning in image classification models: An uncertainty quantification perspective. *Knowledge-Based Systems* 243 (2022), 108418. https://doi.org/10.1016/j.knosys.2022.108418