Please quote as: Reinhard, P.; Li, M. M.; Dickhaut, E.; Reh, C.; Peters, C.; Leimeister, J. M. (2023). A Conceptual Model for Labeling in Reinforcement Learning Systems: A Value Co-Creation Perspective. In A. Gerber & R. Baskerville (eds.), International Conference on Design Science Research (DESRIST). Pretoria, South Africa.

A Conceptual Model for Labeling in Reinforcement Learning Systems: A Value Co-Creation Perspective

Philipp Reinhard[1[0000-0003-3207-666X]], Mahei Manhai Li[1[0000-0002-7171-9130]], Ernestine Dickhaut[1[0000-0002-1946-4038]], Cornelius Reh[1], Christoph Peters[1,2[0000-0001-8140-1516]] and Jan Marco Leimeister[1,2[0000-0002-1990-2894]]

[1] University of Kassel, Kassel, Germany
{philipp.reinhard, mahei.li, ernestine.dickhaut, christoph.peters, leimeister}@uni-kassel.de
[2] University of St.Gallen, St.Gallen, Switzerland
{christoph.peters,janmarco.leimeister}@unisg.ch

Abstract. Artificial intelligence (AI) possesses the potential to augment customer service employees e.g. via decision support or solution recommendations. Still, its underlying data for training and testing the AI systems is provided by human annotators through human-in-the-loop configurations. However, due to the high effort for annotators and lack of incentives, AI systems face low underlying data quality. That in turn results in low prediction performance and limited acceptance by the targeted user group. Faced with the enormous volume and increasing complexity of service requests, IT service management (ITSM) especially, relies on high data quality for AI systems and incorporating domain-specific knowledge. By analyzing the existing labeling process in that specific case, we design a revised to-be process and develop a conceptual model from a value co-creation perspective. Finally, a functional prototype as an instantiation in the ITSM domain is implemented and evaluated through accuracy metrics and user evaluation. The results show that the new process increases the perceived value of both labeling quality and the perceived prediction quality. Thus, we contribute a conceptual model that supports the systematic design of efficient and interactive labeling processes in diverse applications of reinforcement learning systems.

Keywords: Human-in-the-loop, Interactive labeling, Artificial intelligence, Value co-creation

1 Introduction

Artificial Intelligence (AI) based information systems are becoming a key factor in today's workplaces [1]. Especially knowledge-intensive organizations, such as customer service support [2], aim to augment employees by employing machine learning and deep learning [3]. In the realm of ITSM, augmentation is required to cope with the ever-increasing number of customer problems and the challenge of having high turnover rates [4]. Even before the pandemic, the average annual turnover rate of help desks reached 40%, with ITSM domain showing the highest overall turnover rate [5]. Overall,

the technological capabilities of AI possess a large potential improving workplaces by relieving service employees from monotone and repetitive tasks [6] and supporting problem-solving capabilities through AI-enabled recommender systems [7]. However, implementing intelligent systems and the subsequent adoption of AI-based systems at the workplace comes with two major challenges, which is this paper's DSR focal challenge: The first challenge refers to the lack of data quality caused by limited incentives to label and maintain data [8]. AI and in particular supervised or semi-supervised hybrid intelligence systems typically rely on high-quality data to train their models [9]. Yet, employees are typically not eager to annotate data given the processing effort [10] and the lack of incentives and lack of immediate returns, resulting in subpar data quality [8]. The second challenge refers to a lack of trust and confidence in AI-based systems. For example, experienced support agents are especially skeptical of AI systems, also known as algorithm aversion [11]. A lack of trust may induce resistance to the suggestions of AI [12]. Therefore, designers of human-AI collaboration tools and especially decision support and recommendation systems aim at increasing trust in these AI-based systems by ensuring a high prediction performance and emphasizing the capabilities of the systems [13]. Other approaches address the role of explainability (XAI) and transparency concerning trust [14]. Explanations of performance, functionality, and limitations of AI systems can increase trust in their prediction results [12].

Interactive machine learning approaches [15] try to integrate the human user into the learning mechanism as a so-called human-in-the-loop (HITL) [16] - for example by interactive labeling processes [17]. Prior research on interactive labeling aimed at improving the upfront labeling activities by increasing user engagement [18] or gamifying the annotation tasks [19]. Still, prior approaches have yet not considered the underlying cause of the mentioned challenges. Users are either only incentivized extrinsically or are rewarded with a delay. However, HITL-based labeling processes constitute a value co-creation phenomenon as different actors – here AI-based systems and human users – create value through a value-driven interaction. Thus, by incorporating a value co-creation perspective and a service-dominant logic as our input knowledge [20, 21], we aim to overcome the challenges of data quality and trust and facilitate the co-creation of value in human-ai interaction via labeling activities. Co-creation of value in human-ai systems is going to motivate users to label data during use and show more trust as the realized value-in-use is higher. Therefore, we state the following research question: *How can we design a conceptual model for interactive labeling for reinforcement learning by incorporating a value co-creation perspective?*

By following the DSR process, we propose a conceptual model as a solution for designing value co-creation-based labeling processes in interactive machine learning. With the configured labeling process and the derived design principles, we aim at emphasizing and strengthening the role of humans in the development and application of AI [22]. We focus primarily on the interaction between the AI-based system and the end user. The role of the operator is out of scope. Finally, an instantiation and an evaluation of the derived design conclude the design process.

2 Related Work

2.1 Interactive Machine Learning and Labeling

Machine learning (ML) in complex environments, such as ITSM, is in demand of high amounts of high-quality domain-specific user input [23]. Meza Martínez, Nadj, Maedche [24] differentiate between two possible strategies to overcome the lack of domain-specific knowledge in ML. The first approach suggests ML practitioners learn from domain experts when designing and developing ML models [23]. As users are not involved in the development at all, there is a lack of user engagement and a lack of trust because the systems are being considered a "black box" [25, 26]. Another approach, which represents the underlying theoretical foundation for this research project, is interactive ML. Interactive ML [15, 26], sometimes referred to as hybrid intelligence [9], induces a hybrid of human and machine intelligence by joining both – humans and machines – in a learning mechanism. Thereby, complementary strengths can be leveraged, especially for scaling services where human interaction is key [27]. A common practice among developers involves employing HITL mechanisms to ensure that users are directly involved and that the models learn continuously and iteratively as users provide domain-specific input [25]. Meza Martínez, Nadj, Maedche [24] distinguish between supervised learning, active learning, and reinforcement learning as HITL learning mechanisms. This paper focuses only on a ticket recommender system that relies on reinforcement learning as its underlying learning mechanism [28]. Reinforcement learning is characterized by a self-learning mechanism based on rewards and punishments during use [29].

Interactive labeling is an important part of building high-performance interactive ML systems [17]. Interactive labeling constitutes a field of practice and research, in which the role of the human is especially prevalent for providing domain-specific knowledge for training the models. Whereas other approaches see the role of domain experts in the development of the functionalities, for example in low-code development projects [30]. One of the goals of this approach aims at removing noisy annotated data [31]. Another goal is to inspect the most uncertain label instances [31]. Most of the prior work on interactive labeling concern the number of high-quality labels as the main objective. Therefore, optimizing labeling processes and motivating human users to contribute labeled data is of certain interest in practice as well as in literature [8, 17, 10, 26]. Facilitating labeling processes by providing interactive modes and incentives can contribute to not only increasing the available data quality but also for example incorporating more valuable domain knowledge into other systems and artifacts. Within this research, we emphasized facilitating value co-creation through interactive labeling.

2.2 Value Co-Creation

When applying the service-dominant logic (SDL) as a theoretical lens on interactive ML, the interaction between humans and the AI can be seen as a value co-creation where both actors provide and integrate resources [21]. Human users or annotators contribute domain-specific knowledge and offer feedback while the prediction model learns certain patterns, provides useful recommendations, and thereby augments the

human workplace. A key aspect of SDL is the co-creation of value, which emphasizes the collaborative and interactive creation of value between actors and entities through the mutually beneficial integration of resources [32–34]. Following the theories on value co-creation, customers – here support agents – are not only consuming a service, but they are also active participants in the value creation process through interaction [21]. The creation of value is defined as a process by which the user is made better off in some way [35] – for instance by reducing the time for finding a solution or by receiving high-quality recommendations. The unique nature of SDL is that there is more meaning to the value-in-use than there is to the value-in-exchange [36]. Value-in-use refers to the individually perceived value when using the AI-based system instead of only for example consuming a recommended solution [37, 21]. In contrast to value-in-exchange, value-in-use accumulates over time [20]. Grönroos, Voima [20] differentiate between three spheres in which value is created: The provider sphere (here: AI-based system) acts as a facilitator and enables a potential value. On the other side, within the customer sphere, the support agents create value independently by translating and adjusting recommended solutions to fit the special problem and by communicating with the end user. Between these two spheres exists the joint sphere, in which both – the value facilitator and the value recipient - create real value through direct interaction [20]. Thus, following a value co-creation and value-in-use perspective, we derive design knowledge for a value-driven interactive labeling process for reinforcement learning-based AI systems.

3 Methodology

Following the DSR approach our overall goal is to design a solution for the case of problem-solving tasks in ITSM and infer design knowledge for a generally improved value-driven interactive labeling process in HITL configurations, presented as a conceptual model (Figure 1).

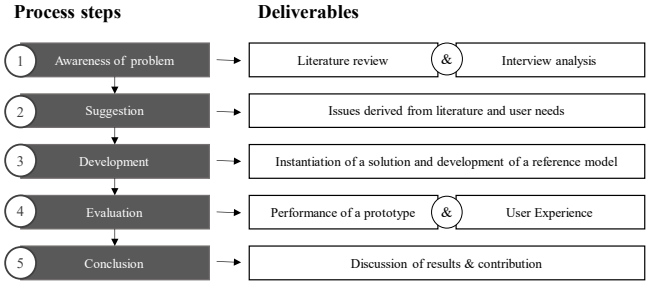

Fig. 1. DSR procedure according to Kuechler & Vaishnavi (2008)

According to Kuechler, Vaishnavi [38], in *Phase 1* (Awareness of problem) we conducted a review of the literature on HITL and interactive labeling. We additionally enriched our fundamental awareness and base of theories by incorporating literature on human-centered AI including trust, control, explainability, and transparency. The

retrieved knowledge and theories ensure the rigor of our research [39]. In addition, value co-creation serves as a theoretical base. To consider practical relevance, we performed semi-structured interviews with eight support agents. The focus was placed on problem-solving capabilities and the interaction with AI-based systems within the service workplace. The interview analysis ensures practical relevance and allows for identifying pressing issues. In *Phase 2* (Suggestion) we aggregated common issues along the existing labeling process, which was derived from the case of ticket recommender systems in the realm of ITSM. Business Process Model and Notation (BPMN) is applied to visualize the current situation and outline the as-is process of currently developed AI systems in a larger DSR project that has been running for 3 years within a consortium. Afterward, in *Phase 3* (Development), we derive design principles from the identified issues and the proposed to-be process. The design principles are then translated into a to-be process model for value-driven interactive labeling [40, 41]. Furthermore, we instantiate the to-be process by developing a multi-armed contextual bandit system for IT support ticket recommendation and conceptualize a conceptual model of value co-creating labeling processes. The to-be process and the corresponding design principles in form of the instantiated AI-based system are evaluated in *Phase 4* (Evaluation) utilizing labeling performance as well as perceived value, sense of control, usability, and overall usefulness. We follow the Framework for Evaluation in Design Science Research (FEDS) [42] to evaluate prototypical instantiation, a technical experiment, and a user evaluation including 11 interviews [41] to provide the necessary validation of the efficiency of the proposed process. Finally, we reflect on the design process and the developed conceptual model and conclude with a discussion.

4 Design & Development

4.1 As-is Labeling Process

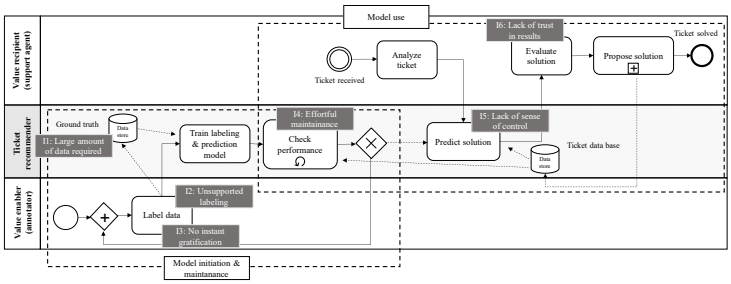

Fig. 2. As-is labeling and model training process for TRS.

To a large extent, literature on developing and operating supervised- or semi-supervised AI models refers to two different roles in terms of model initialization and model use. The group of "annotators" label data and provide the ground truth for training and testing the model during the initialization phase. However, annotators typically do not use the system afterward and are incentivized intrinsically [43]. For the case of supporting service employees in ITSM organizations, we differentiate between the roles of

annotators and users – here support agents using a ticket recommender system. The ticket recommender system augments the problem-solving activities of support agents by recommending already solved problems as possible solutions. We analyzed and visualized the as-is labeling and model use process. As Fig. 2. shows, the ground truth is utilized to train a model, that predicts a recommended solution ticket given an incoming request. The human in the loop then evaluates the prediction and decides how to adapt or reject the solution. Along the as-is process, we identified six major issues justified by practical relevance and literature:

Issue 1: A large amount of data is required upfront. Currently, annotators are meant to label a large amount of data upfront to categorize a ticket (Expert (E) 2) and to train and test the model [23]. This approach is time-intensive and costly (E2, E3, E4, E7) [10, 44]. An improved labeling process should therefore reduce the required volume of ground truth data during the initial incorporation of domain-specific knowledge and generate labeled data during operation.

Issue 2: Unsupported labeling. The upfront labeling efforts are not being supported by the traditional labeling processes and act as an oracle. Typically, samples for labeling are selected by the development team and then forwarded to the annotators uncured. This hampers labeling efficiency, discourages annotators in the long term, and reduces data quality. In practice, a knowledge manager is responsible for providing domain knowledge (E4, E7) and agents are under large pressure (E6). An improved process should therefore aim at augmenting and semi-automate at least a part of the labeling tasks [45].

Issue 3: No intrinsic gratification. Usually, annotators do not benefit from labeling the data [43]. As such the relationship between the task giver and the annotator can be described by the principle-agent theory, where both actors strive for different goals [46]. Accordingly, an improved labeling process should ensure that the HTIL takes over both – the role of a value recipient and a value enabler by auditing data [1] to ensure an intrinsic motivation – "practically get the tickets from a certain period on a certain topic as an extract" (E1, E7). However, even if dedicated experts are involved in the labeling process upfront, they only experience delayed effects of their resource contribution to the prediction model as our process visualization shows. There is a lack of incentives to label data [8]. Given the preference for instant gratification, a to-be process should provide immediate value-in-use for the annotators [47].

Issue 4: Extensive maintenance for continuous model improvement. Because of the increasing number of incoming tickets for IT support, new data is generated during the use rapidly. Therefore "it [the knowledge base] is very high-maintenance, outdated and you can't find anything" (E4). To improve the model continuously and adapt to data drift [48], model operators have to monitor the performance and initiate new labeling phases. In sum, the effort for observing data drift and performing model maintenance is high [49]. For that reason, an interactive labeling process and self-learning system should autonomously improve its performance and its underlying data quality continuously and simultaneously (E4, E8).

Issue 5: Lack of sense of control. The theory of IS identity concerns the impact of AI in workplaces in terms of a lack of control [50]. Employees are being more and more replaced in knowledge-sharing and extraction activities [51]. In the context of human-

ai interaction, it is important to give humans control over the results and the adaption of AI recommendations. We propose that service employees should experience more interaction with AI, where they act as supervisors [52] and verify machine outcomes [9, 53]. In addition, Dietvorst, Simmons, Massey [54] showed that giving users some degree of control can reduce algorithm aversion. In conclusion, a labeling process for HITL configurations should enable humans to control in terms of adjusting, evaluating, and accepting or rejecting results.

Issue 6: Lack of trust in the results. Prior research has examined the phenomenon of algorithm aversion [54] and emphasized the importance of trust in AI systems [12, 14]. Generally, AI systems possess a lack of trust of humans. Especially in complex problem-solving tasks like IT support, trust is limited by the restricted performance and reliance on recommender systems and the difficulty to recommend optimal solutions [55]. Support agents state that they typically need a lot of time to find a certain ticket in the database and thus rely on colleagues instead of using an AI-based system (E1, E5). As AI is perceived as being a "black box", researchers and practitioners aim at increasing transparency and explainability to increase trust [14, 55]. However, simply increasing trust in AI predictions is not ideal, as it could lead to over-reliance. An improved labeling mechanism therefore should increase trust by enhancing the perceived accuracy while at the same time giving users an indication and transparency of when to trust the systems' predictions.

4.2 To-be Process: Value-driven Labeling Process

By addressing the issues, we developed a revised process for value-driven labeling (Figure 3). Starting with the role of the user as the annotator, we removed the bottom lane to illustrate the removement of the dichotomy between value enablers and value recipients. Still, the process requires an initial set of ground truth data for pre-training the system. However, we ensured that the initial required data remains manageable. As in the as-is process, the ground truth data is applied for training a labeling and prediction model. Given an incoming ticket, the system proposes a labeling of the content by highlighting relevant entities. The subsequent prediction model processes the pre-labeled ticket and directly presents solutions to the user. Now, the user can check whether the proposed recommendations are of use and whether the pre-labels are correct. If not, it is possible to adjust the labels and start a new prediction. Finally, the user evaluates the suggested solutions and gives feedback back to the reinforcement learning system.

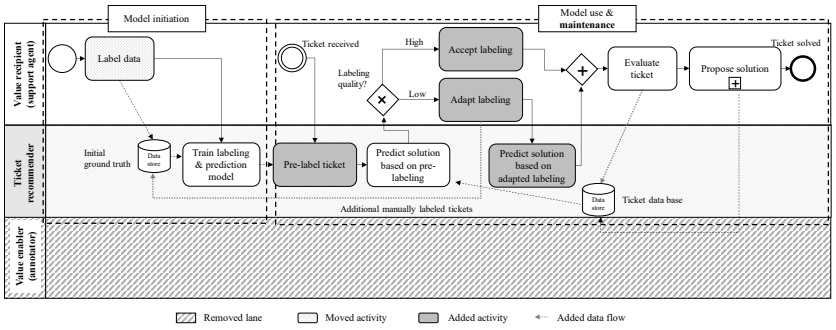

Fig. 3. To-be value-driven and interactive labeling process.

4.3 Conceptual Model for Value-Driven Labeling

From the identified issues of the as-is process and components of the developed to-be process, we derived a conceptual model (Figure 4) that incorporates a value co-creation perspective [20, 21] on interactive labeling and interactive ML. By including the principles of service-dominant logic [21] and value co-creation [20], finally, four generalized design principles for value-driven interactive labeling systems constitute the value co-creation model for labeling:

DP1: Augment the labeling process. The labeling process is augmented by a pre-trained model, that provides automatically generated labels initially. Based on a small size of ground truth data, that requires less effort to generate, a labeling model is being initiated. This initialization phase is required to augment the labeling process more automatically and effectively. Therefore, a traditional annotation phase is placed upfront involving domain experts who at the same time are the users. The automatically retrieved labels themselves should provide a benefit for the users by providing the user additional information or support for the problem-solving or decision task. According to the theory of value co-creation customers are meant to act as value creators [20]. Thus, the design principle proposes that the role of external annotators as value enablers should be dismissed. Instead, the actual user must label voluntarily during the phase of model use. The user thereby decides which labels need to be improved to improve model prediction [31]. This overall results in a removal of the dichotomy between value enabler and value recipient. By minimizing the initial need for data and domain-specific knowledge, the design downsizes the so-called provider sphere [20].

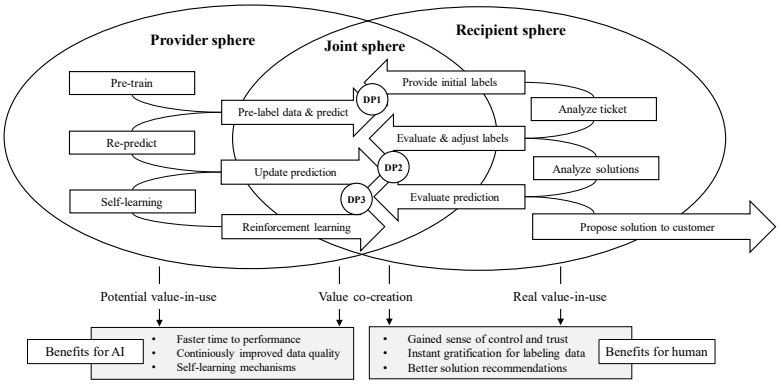

Fig. 4. Conceptual model for co-creation-based labeling based on [20]

DP2: Enable control over value creation through labeling. From a service perspective, the second design principle ensures accumulating value throughout the user's value-creation process [21]. The role of the user as a value creator is emphasized by the gained control and the enablement of co-producing value [20]. After receiving a pre-labeled ticket, the user decides whether to adjust the labels and share knowledge with

the system [47]. In addition, the designed process enables users to evaluate the recommendation results and return feedback (e.g., selection of a ticket, 5-star rating, thumbs up/thumbs down). Our research suggests maintaining "human-in-control" as a key paradigm of ML-based recommender systems [16].

DP3: Provide immediate value-in-use for labeling activities. Due to the high effort [10] and the low immediate incentives of labeling [8], data quality for ML models remains extremely precarious in the complex and context-specific environment. For such complex domains, interactive systems provide particular value [17]. An interactive labeling process should therefore reward labeling activities immediately to motivate and incentivize labeling and thereby increase the perceived relevance for annotators [17]. Such rewards can be better decision support or more personalized recommendations and must be provided immediately after contributing high-quality domain knowledge as part of the value-in-use [21]. The long-term benefit – delayed benefits – of providing the system with labels can only be hardly presented. The conceptual model places the focus on value-in-use and not primarily on potential or prospective use.

The following conceptual model summarizes the key activities of value co-creation in interactive ML systems and the abstracted design principles within that ML value chain. Additionally, the visualization shows how the complementary strengths generate added value above mere solution recommendations for both humans and machines. According to Grönroos, Voima [20], we differentiate between a (1) provider sphere, (2) recipient sphere, and (3) joint sphere. The ML system represents the provider sphere as it supports the human user through recommendations. The ML system itself only generates a potential value-in-use, that is being activated within the joint sphere and realized by the user within the recipient respectively customer sphere.

5 Evaluation

We evaluate our design and its instantiation by applying FEDS [56] to answer our research question within an artificial and a naturalistic summative evaluation. At first, a technical performance evaluation is conducted [42]. We apply common accuracy measures (accuracy, precision, recall, F1-Score) to validate the usefulness of the pre-highlighting model[57]. Afterward, our prototypical instantiation is presented in a naturalistic summative evaluation where 11 users work with the system and provide feedback in a subsequent interview.

5.1 Technical Performance Evaluation

To evaluate the initialization phase, we evaluate the automated labeling mechanism and the prediction model based on 60,000 support tickets from 2021, which were provided by an international manufacturing company and preprocessed in several steps. The results should indicate how well the initial labeling phase performs. We compared different approaches of modern entity taggers to maximize the quality of the pre-labeling. For our underlying database, a BERT-based highlighting model achieves significantly higher accuracy scores than for example a Bi-LSTM. The corresponding confusion matrix reveals that BERT more precisely labels "system", "failure description"

and “service request”. The results (Table 1) showed that transformer-based ML tools like BERT can provide annotators with useful suggestions for highlighting the tickets based on a small database. Thereby augmentation of the labeling process and reduction of the upfront labeling effort takes place (DP1). However, overall, the pre-labeling model possesses a comparatively low performance. This is reasoned by the unstructured and informal character of most of the problem descriptions.

Table 1. Performance metrics of the pre-trained labeling system[1]

	Accuracy	Precision	Recall	F1-Score
BERT-based	0.806	0.769	0.806	0.773
Bi-LSTM	0.249	0.392	0.249	0.299

5.2 User Evaluation

Within a naturalistic user evaluation, we instantiated the process and enabled a walk of the designed interactive labeling process. The goal of this evaluation was to account for the perceived immediate value of manual labeling (DP3) and the perceived sense of control (DP2). For the sake of simplicity, the ticket recommender system only presented 3 solution tickets in each case. The involved participants were instructed to first-rate the recommended solution based on pre-labeled tickets, then manually adjust the labels if necessary and rate the new predictions on a scale of 1 to 5. Solving the tickets took an average of 20 minutes. The interview partners were on average 26 years old and all possess a technical background and experience with IT. Afterward, the 11 interviewees (I) were asked to answer questions regarding their experience during interviews. Finally, we conducted a short survey. In total, the user evaluation lasted about 60 minutes.

We interviewed the participants regarding the overall experience of the value-driven labeling process. Regarding the augmentation of the labeling process (DP1), the interviewed participants stated that the tool intuitively supported the labeling, and annotating the data was perceived as being straightforward. In addition, interviewees mentioned that the labeling supported their cognitive processing of the presented problem cases and understanding of the recommended solutions. The interviewees expected that the recommended labels were correct, and showed signs of reliance and blind trust. The automated labeling could be extended to the presented solutions as well, to enable an easier matching of problem-solution pairs: *“I think next time I would label first and then read the tickets, to be able to match the problem with the recommended tickets easier*” (I3). Overall, the participants did not perceive highlighting the data as being effortful, unnecessary, or meaningless. Nevertheless, to make the process more convenient, the system could give agents information and examples on the different categories and make the highlighting clickable (I7).

Providing the system with domain-specific knowledge in the form of labeling the text phrases, stimulated a sense of control (DP2) (I2, I4, I6). Interestingly, the willingness to contribute to higher data quality through interactive labeling was broadly confirmed and reasoned by the benefit of receiving better recommendations (DP3):

[1] Weighted precsion, recall and F1-score

"*Because then I noticed that with correct labeling I also get immediately meaningful solution possibilities"* (I4). The users understood that labeling the data supported the AI to "*narrow down the problem request*" (I1) or "*filter based on important phrases*" (I2). An interviewee compared the labeling to providing prompts to ChatGPT: "*you have to specify the input to the AI so that it can answer your question exactly – that's similar to this ChatGPT*" (I1). In conclusion, the system motivates users to input their knowledge and justifies the effort of highlighting the text. The immediate adjustment of the recommendation was thus perceived as valuable. Subsequently, trust in the system was raised (DP3). Asking the participants whether bad recommendations were caused by insufficient labeling or the underlying system, they showed more confidence in their labeling activities and blame the machine to be inaccurate (I1). This was confirmed by others that stated that the system presented single recommendations that were extremely unsuitable (I1, I4, I5). During operations, the system must ensure the extraction of such distrust-generating results as trust is strongly influenced by the quality of recommendations. Furthermore, trust was generated by matching the labeled keywords with the solution recommendations: "*So if you see the same keywords again at the bottom. So the same topic then you always think well the results fit*" (I7). The effects should be examined within larger quantitative research as a few interviewees did not deduce a connection between labeling and recommendations. However, this could be caused by single inferior recommendations. Following the interviews, we measure the usability of our system using the system usability scale (SUS) [58]. Given the 11 responses, we achieved a 73.5 SUS score which in the interpretation speaks for "good usability" confirming the qualitative feedback.

6 Implications, Limitations, and Conclusion

We developed a conceptual model for value-driven interactive labeling by incorporating a value co-creation perspective. Along with the research project, we analyzed the as-is process and identified six key issues that restrict the potential of AI systems. After developing a to-be labeling process, we derived a generalized conceptual model which integrates a service perspective [20, 21]. As one of the first papers in the literature stream on interactive ML [15], hybrid intelligence [9], and HITL configurations [1], we show how value co-creation can be the core of self-learning systems and how an integrated interactive labeling process removes the dichotomy between value recipients and value facilitators. With enabling value-in-use in terms of immediate perceived value and providing a sense of control, we contribute novel mechanisms to the knowledge of designing interactive labeling systems [17]. From a practical perspective, the model can be used to improve ML development and operations in terms of efficient initialization, continuous usage, and model maintenance. The approach outlines a way to incentivize users to contribute domain-specific knowledge. The optimized interaction between humans and AI will lead to a higher prediction and subsequent service performance.

Our research comes with limitations and provides room for future research. Given the scope of this research, the role of operators and how they are integrated into the conceptual model remains neglected. One limitation refers to the selected interactive

ML type. Considering aspects of other types such as active learning can provide additional insights into value co-creation-based labeling. For example, a system could only request new labels or feedback when the data is needed to improve the model or the results. Thereby researchers and practitioners could further reduce the demand on users to label data [17]. In addition, our evaluation does not consider labeling quality and performance as key metrics. Calculating the differences between automated labels and manual labels through similarity scores, our evaluation results could have been enhanced by revealing the user's contribution and engagement. Another limitation points to the challenge of bad labels but good recommendations which require additional feedback mechanisms. It is not revealed whether users will first check the labels or check the recommendations as our test setup asked the user to first-rate the pre-predictions. That way we ensured that a comparison of recommendations based on pre-labeled and manually labeled could be drawn. Overall, further research has to be conducted to ensure an accumulation of value and value co-creation in the long term by evaluating the performance of the automated labeling model after adding manual labels. A future large-scale experiment should aim for quantitative analysis that could underline the effects mentioned by the interviewees.

We expect that our resulting conceptual model for value-co-creation-based labeling can be applied to different labeling tasks and different learning mechanisms as well. In conclusion, our evaluation suggests that the perceived immediate value can stimulate the willingness to co-create value in HITL configurations and thus improve data quality and subsequent recommender performance. Thus, we provide novel insights into solving the challenge of data quality in AI.

References

1. Grønsund T, Aanestad M (2020) Augmenting the algorithm: Emerging human-in-the-loop work configurations. The Journal of Strategic Information Systems 29:101614. https://doi.org/10.1016/j.jsis.2020.101614
2. Li MM, Peters C, Leimeister JM (2017) Designing a Peer-Based Support System to Support Shakedown
3. Al-Hawari F, Barham H (2021) A machine learning based help desk system for IT service management. Journal of King Saud University - Computer and Information Sciences 33:702–718. https://doi.org/10.1016/j.jksuci.2019.04.001
4. Dostál M (2022) Service Desk Onboarding Training Environment. Acta Informatica Pragensia 11:265–284
5. Rumburg J (2018) Metric of the Month: Annual Agent Turnover.
6. Schmidt S, Li M, Peters C (2022) Requirements for an IT Support System based on Hybrid Intelligence. In: HICSS
7. Li M, Löfflad D, Reh C et al. (2023) Towards the Design of Hybrid Intelligence Frontline Service Technologies – A Novel Human-in-the-Loop Configuration for Human-Machine Interactions. In: HICSS

8. Kubiak P, Rass S (2018) An Overview of Data-Driven Techniques for IT-Service-Management. IEEE Access 6:63664–63688. https://doi.org/10.1109/ACCESS.2018.2875975
9. Dellermann D, Ebel P, Söllner M et al. (2019) Hybrid Intelligence. Bus Inf Syst Eng 61:637–643
10. Choi M, Park C, Yang S et al. (2019) AILA: Attentive Interactive Labeling Assistant for Document Classification through Attention-Based Deep Neural Networks. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. ACM, New York, NY, USA
11. Luo X, Qin MS, Fang Z et al. (2021) Artificial Intelligence Coaches for Sales Agents: Caveats and Solutions. Journal of Marketing 85:14–32. https://doi.org/10.1177/0022242920956676
12. Kim T, Song H (2022) Communicating the Limitations of AI: The Effect of Message Framing and Ownership on Trust in Artificial Intelligence. International Journal of Human–Computer Interaction:1–11. https://doi.org/10.1080/10447318.2022.2049134
13. Jacovi A, Marasović A, Miller T et al. (2021) Formalizing Trust in Artificial Intelligence. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency. ACM, New York, NY, USA
14. Schmitt A, Wambsganss T, and Janson A (2022) Designing for Conversational System Trustworthiness: The Impact of Model Transparency on Trust and Task Performance. In: ECIS, vol 172
15. Holzinger A (2016) Interactive machine learning for health informatics: when do we need the human-in-the-loop? Brain Inf 3:119–131. https://doi.org/10.1007/s40708-016-0042-6
16. Wiethof C, Bittner E (2021) Hybrid Intelligence - Combining the Human in the Loop with the Computer in the Loop: A Systematic Literature Review. In:
17. Nadj M, Knaeble M, Li MX et al. (2020) Power to the Oracle? Design Principles for Interactive Labeling Systems in Machine Learning. Künstl Intell 34:131–142. https://doi.org/10.1007/s13218-020-00634-1
18. Viana L, Oliveira E, Conte T (2021) An Interface Design Catalog for Interactive Labeling Systems. In: Proceedings of the 23rd International Conference on Enterprise Information Systems. SCITEPRESS - Science and Technology Publications
19. Warsinsky S, Schmidt-Kraepelin M, Thiebes S et al. (2022) Gamified Expert Annotation Systems: Meta-Requirements and Tentative Design. In: Springer, Cham, pp 154–166
20. Grönroos C, Voima P (2013) Critical service logic: making sense of value creation and co-creation. J of the Acad Mark Sci 41:133–150
21. Vargo SL, Lusch RF (2004) The four service marketing myths: remnants of a goods-based, manufacturing model. Journal of Service Research 6:324–335
22. Zanzotto FM (2019) Viewpoint: Human-in-the-loop Artificial Intelligence. jair 64:243–252
23. Porter RB, Theiler JP, Hush DR (2013) Interactive Machine Learning in Data Exploitation. Office of Scientific and Technical Information (OSTI)

24. Meza Martínez MA, Nadj M, Maedche A (2019) Towards an integrative theoretical framework of interactive machine learning systems. In: ECIS
25. Amershi S, Cakmak M, Knox WB et al. (2015) Power to the People: The Role of Humans in Interactive Machine Learning. AIMag 35:105–120. https://doi.org/10.1609/aimag.v35i4.2513
26. Jiang L, Liu S, Chen C (2019) Recent research advances on interactive machine learning. J Vis 22:401–417. https://doi.org/10.1007/s12650-018-0531-1
27. Kleinschmidt S, Peters C, Leimeister JM (2020) How to scale up contact-intensive services: ICT-enabled service innovation. JOSM 31:793–814. https://doi.org/10.1108/JOSM-12-2017-0349
28. Afsar MM, Crump T, Far B (2021) Reinforcement learning based recommender systems: A survey. CoRR
29. Kaelbling LP, Littman ML, Moore AW (1996) Reinforcement Learning: A Survey. jair 4:237–285. https://doi.org/10.1613/jair.301
30. Elshan E, Ebel PA, Söllner M et al. (2022) Leveraging Low Code Development of Smart Personal Assistants: An Integrated Design Approach with the SPADE Method. Journal of Management Information Systems (JMIS)
31. Bernard J, Hutter M, Zeppelzauer M et al. (2018) Comparing Visual-Interactive Labeling with Active Learning: An Experimental Study. IEEE Trans Visual Comput Graphics 24:298–308. https://doi.org/10.1109/tvcg.2017.2744818
32. Schüritz R, Farrell K, Wixom B et al. (2019) Value Co-Creation in Data-Driven Services: Towards a Deeper Understanding of the Joint Sphere. In:
33. Blaschke M, Riss U, Haki K et al. (2019) Design principles for digital value co-creation networks: a service-dominant logic perspective. Electron Markets 29:443–472. https://doi.org/10.1007/s12525-019-00356-9
34. Peters C (2020) Designing Work and Service Systems. Doctroral Dissertation
35. Grönroos C (2008) Service logic revisited: who creates value? And who co-creates? European Business Review 20:298–314. https://doi.org/10.1108/09555340810886585
36. Vargo SL, Lusch RF (2008) Service-dominant logic: continuing the evolution. J of the Acad Mark Sci 36:1–10. https://doi.org/10.1007/s11747-007-0069-6
37. Grönroos C (2011) Value co-creation in service logic: A critical analysis. Marketing Theory 11:279–301. https://doi.org/10.1177/1470593111408177
38. Kuechler B, Vaishnavi V (2008) On theory development in design science research: anatomy of a research project. European Journal of Information Systems 17:489–504
39. Hevner, March, Park et al. (2004) Design Science in Information Systems Research. MIS Quarterly 28:75. https://doi.org/10.2307/25148625
40. Winter R (2008) Design science research in Europe. Eur J Inf Syst 17:470–475. https://doi.org/10.1057/ejis.2008.44
41. Peffers K, Rothenberger M, Tuunanen T et al. (2012) Design Science Research Evaluation. In: Springer, Berlin, Heidelberg, pp 398–410
42. Venable J, Pries-Heje J, Baskerville R (2016) FEDS: a Framework for Evaluation in Design Science Research. Eur J Inf Syst 25:77–89. https://doi.org/10.1057/ejis.2014.36

43. Cao H-A, Wijaya TK, Aberer K et al. (2015) A collaborative framework for annotating energy datasets. In: 2015 IEEE International Conference on Big Data (Big Data). IEEE
44. Yan, Jie Yang, Hauptmann (2003) Automatically labeling video data using multi-class active learning. In: Proceedings Ninth IEEE International Conference on Computer Vision. IEEE
45. Desmond M, Duesterwald E, Brimijoin K et al. (2021) Semi-Automated Data Labeling. NeurIPS 2020 Competition and Demonstration Track:156–169
46. Eisenhardt KM (1989) Agency Theory: An Assessment and Review. AMR 14:57–74. https://doi.org/10.5465/amr.1989.4279003
47. Ranjan KR, Read S (2016) Value co-creation: concept and measurement. J of the Acad Mark Sci 44:290–315. https://doi.org/10.1007/s11747-014-0397-2
48. Mallick A, Hsieh K, Arzani B et al. (2022) Matchmaker: Data Drift Mitigation in Machine Learning for Large-Scale Systems. In: Proceedings of Machine Learning and Systems, vol 4, pp 77–94
49. Pianykh OS, Langs G, Dewey M et al. (2020) Continuous Learning AI in Radiology: Implementation Principles and Early Applications. Radiology 297:6–14. https://doi.org/10.1148/radiol.2020200038
50. Mirbabaie M, Brünker F, Möllmann Frick, Nicholas R. J. et al. (2022) The rise of artificial intelligence – understanding the AI identity threat at the workplace. Electron Markets 32:73–99. https://doi.org/10.1007/s12525-021-00496-x
51. Vorobeva D, El Fassi Y, Costa Pinto D et al. (2022) Thinking Skills Don't Protect Service Workers from Replacement by Artificial Intelligence. Journal of Service Research 25:601–613. https://doi.org/10.1177/10946705221104312
52. Braun M, Greve M, Riquel J et al. (2022) MEET YOUR NEW COLLE (AI) GUE–EXPLORING THE IMPACT OF HUMAN-AI INTERACTION DESIGNS ON USER PERFORMANCE. In: ECIS
53. Hemmer P, Schemmer M, Riefle L et al. (2022) Factors that influence the adoption of human-AI collaboration in clinical decision-making. In: ECIS
54. Dietvorst BJ, Simmons JP, Massey C (2018) Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. Management Science 64:1155–1170
55. Lockey S, Gillespie N, Holm D et al. (2021) A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. In: HICSS
56. Venable J, Pries-Heje J, Baskerville R (2012) A Comprehensive Framework for Evaluation in Design Science Research. In: Springer, Berlin, Heidelberg, pp 423–438
57. Shani G, Gunawardana A (2011) Evaluating Recommendation Systems. In: Recommender Systems Handbook. Springer, Boston, MA, pp 257–297
58. Brooke J (1996) SUS: A 'Quick and Dirty' Usability Scale. In: Usability Evaluation In Industry. CRC Press, pp 207–212