The Impact of Concept Explanations and Interventions on Human-Machine Collaboration

Jack Furby[1(✉)], Dan Cunnington[2], Dave Braines[3], and Alun Preece[1]

[1] Cardiff University, Cardiff, UK
furbyjl@cardiff.ac.uk
[2] Imperial College London, London, UK
[3] IBM Research Europe, Hursley, UK

Abstract. Deep Neural Networks (DNNs) are often considered black boxes due to their opaque decision-making processes. To reduce their opacity Concept Models (CMs), such as Concept Bottleneck Models (CBMs), were introduced to predict human-defined concepts as an intermediate step before predicting task labels. This enhances the interpretability of DNNs. In a human-machine setting greater interpretability enables humans to improve their understanding and build trust in a DNN. In the introduction of CBMs, the models demonstrated increased task accuracy as incorrect concept predictions were replaced with their ground truth values, known as intervening on the concept predictions. In a collaborative setting, if the model task accuracy improves from interventions, trust in a model and the human-machine task accuracy may increase. However, the result showing an increase in model task accuracy was produced without human evaluation and thus it remains unknown if the findings can be applied in a collaborative setting. In this paper, we ran the first human studies using CBMs to evaluate their human interaction in collaborative task settings. Our findings show that CBMs improve interpretability compared to standard DNNs, leading to increased human-machine alignment. However, this increased alignment did not translate to a significant increase in task accuracy. Understanding the model's decision-making process required multiple interactions, and misalignment between the model's and human decision-making processes could undermine interpretability and model effectiveness.

Keywords: Concept Models · Human study · Alignment · Interpretability · XAI

1 Introduction

Concept Model (CMs), such as Concept Bottleneck Model (CBMs) [13], is a class of Deep Neural Network (DNN) which aims to improve the interpretability of model predictions by structuring predictions around human-understandable components, called concepts. These concepts often correspond to intermediate

R. Guidotti et al. (Eds.): xAI 2025, CCIS 2576, pp. 255–280, 2026.
https://doi.org/10.1007/978-3-032-08317-3_12

attributes of tasks, effectively "splitting" the prediction process into sub-tasks. For instance, a CM that predicts types of birds, might predict concepts for the wing colour and beak shape.

After a CM makes a prediction a human collaborating with the model will be able to inspect concept predictions, known as *concept explanations*, to help understand the model's decision-making process. In domains such as healthcare this may be used to answer why a downstream task was predicted. With some CM architectures, such as CBMs, the concept explanations also introduce the capability for a human to intervene in the concept predictions. As CBMs predict concepts in the range of 0 to 1 (not present to present) with a 0.5 threshold, concept outputs can be replaced with new values within this range. Interventions may be made to correct mistakes the model made when predicting concepts, or otherwise query the model to see what task prediction would be made with a different set of concept predictions. In these scenarios the model's concept vector provides counterfactual explanations [13]. In a collaborative setting, we may consider interventions as a way to increase trust in a model as interventions may reveal the sensitivity a model has to concept values, and thus any bias the model has to concepts.

The authors of CBMs [13] presented results using automated metrics demonstrating improved model accuracy as incorrect concept predictions were intervened and replace with their ground truth values. In addition, model predictions have been shown as a preferred method to identify bias [1]. As the intervention metric was automated it remains unknown if the findings can be applied in a collaborative setting.

The authors of CBMs also made claims of improved human-machine collaboration [13], but human studies to show this are limited and instead compare CBMs to other model architectures [6,12], or complete tasks such as selecting the concepts participants believe the model detected [28]. Only a few studies analyse the class of CMs with collaborative tasks [21,22].

This paper presents two human studies where we analysed CBMs in a collaborative setting. We answer the research question: *Do Concept Models improve task accuracy and model interpretability in a human-machine setting?* We have broken this question down into the following sub-questions:

1. Do test-time interventions improve human-machine task and concept accuracy?
2. Do interventions increase the interpretability of CBMs?
3. Are CBMs trusted?

The main contributions of this paper are as follows:

– We perform the first human studies using CBMs in a joint human-machine task setting which analyses the interaction between humans and the CBM. We find interventions often increased trust in a model but this trust was sometimes misplaced. In addition, the CBM decision-making process is not aligned with that of the humans.

– We show the initial promise of interpretability from high-level concepts is upheld. However, understanding the model's decision-making process requires participants to actively interact with the model. Additionally, providing concept predictions without the capability to intervene has a similar effect on task accuracy.

Although we used CBMs, our findings also apply to other CMs with intervention capabilities and that predict task labels in the same feed-forward fashion from input to concepts, to task label. Namely these are Concept Embedding Models [31], Sidecar CBMs [18], and hybrid CBMs [19].

2 Related Work

Several studies have analysed CMs with human participants. These can be placed into several categories; human concept preference [2,23], concept explanations [6,11,12,27], human-in-the-loop [21,22] and bias discovery [20,30].

In studies on concept preference, one study found that concepts humans identified in samples varied widely and performed worse when used by a CM on downstream tasks compared to those identified by the model [2]. Separately, Participants have also been found to prefer to see 32 or fewer concepts [23] instead of all concepts a model uses for task predictions, if greater than 32. This is consistent with balancing a models completeness [15] to keep participants engaged [14].

Next, studies exploring concept explanations preference have demonstrated a mixed response where concept explanations are favoured in some studies [11], while not in others [6,12,27]. Out of these studies [6,27] evaluated explanation types for bias discovery and model task prediction. As concept explanations underperformed in this area it raises the question of whether CM's concept explanations improve the interpretability of models such that they can aid in human-machine collaboration.

Some studies have investigated human-machine collaboration with CMs. A CBM inspired recommender system was introduced that combined user-provided and automatically generated concepts to suggest relevant text documents [21]. Participants could intervene in the concepts by editing a concept value, leading to improvements of 20–47% in recommendation accuracy compared to initial concept values. A separate study asked participants to label images [22]. Participants either had fixed model predictions to help them or could select parts of the input image for the model to focus on, finding little difference in performance (73.57% vs. 72.68% respectively). Additionally, participants often agreed with the model's predictions regardless of whether the model was correct or incorrect. Both of these studies look at humans updating a model's prediction, similar to interventions with CBMs. As the recommender system [21] is explicitly based on CBMs, their findings suggest similarities could be observed in an image modality.

For bias discovery, [30] (and repeated in the study [20]) used CBM-like architectures to study human-guided pruning on a model where input samples had

shifted (e.g. the correlation of concepts co-occurring is changed after training). Participants selected concepts to prune based on input samples and model predictions, outperforming random pruning and only slightly less effective than fine-tuning or greedy performance. This demonstrates that the concept explanations are effective at aiding a human-in-the-loop understanding of bias in a model.

From these studies, we have identified no papers which look at CBMs or similar model architectures that evaluate the capabilities of CBMs in real-world tasks. Most importantly it has not been shown if human performed interventions improve joint task performance, and whether these models are more interpretable than standard DNNs.

Table 1. Participants in the expert study were split into two groups, both with access to the same model. Participants in the lay-person study were split into eight groups where the model used and explanations provided were varied.

Expert study	Lay-person study	
Participant groups	Participant groups	
	Accurate model	Inaccurate model
CExp+Int	NoExp	NoExp
CExp+Int+SMap	CExp	CExp
	CExp+Int	CExp+Int
	CExp+Int+SMap	CExp+Int+SMap

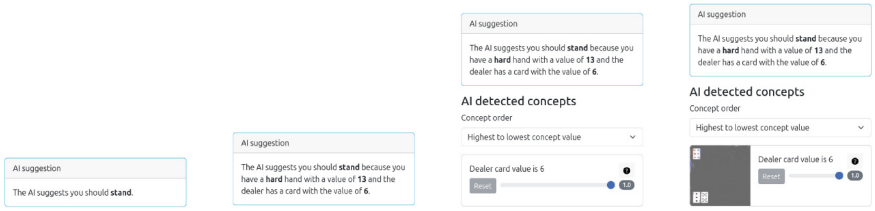
3 Methods

We ran two human studies: (1) An expert study where participants had extensive knowledge about the task domain (skin disease diagnosis) where the model acted as a second opinion. (2) A lay-person study with a general task (Playing games of Blackjack), involving participants with experience levels ranging from novices to skilled individuals, but none being professionals. The model also acted as a second opinion, but could also serve as a guide for participants with less experience. Both studies received favourable ethical opinion from the School of Computer Science and Informatics at Cardiff University.

By running two studies we were able to compare results in similar, but distinct settings where participants can interact with an AI agent to assist them. Following the taxonomy by [5], our lay-person study is human-grounded as we do not use expert participants and use a simulated task, while the expert study is application-grounded as we use both expert participants and a real-world task.

We split participants into two groups in the expert study, and eight groups in the lay-person study, detailed in Table 1, and example explanations are shown

in Fig. 1. As our sub-questions required us to analyse the use of interventions, interpretability, and trust of CBMs, both groups for the expert study included participant access to interventions. As we did not have the same limitation in the lay-person study we also included groups that had access to a different model and included groups with no model explanations and just concept explanations. 12 participants took part in the expert study who were either doctors, consultants or trainees with expertise in dermatology. 104 participants took part in the lay-person study where most were either university staff or students.



(a) Only concept task predictions (NoExp)

(b) Task predictions and top three concept predictions (CExp)

(c) Task predictions, concept predictions and interventions (CExp+Int)

(d) Task predictions, concept predictions, interventions and saliency maps (CExp+Int+SMap)

Fig. 1. Model output variations.

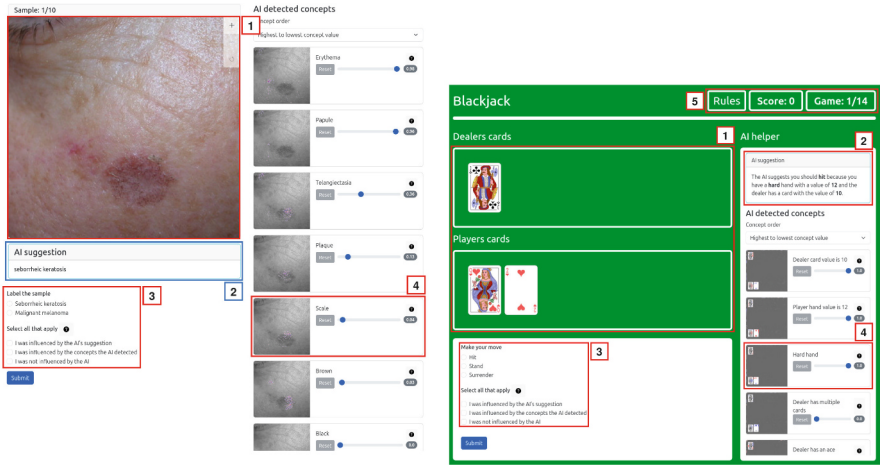
We use the following acronyms to separate each participant group:

Accurate model (Acc) Accurate model (lay-person study only).
Inaccurate model (Inacc) Inaccurate model (lay-person study only).
NoExp No explanations (lay-person study only).
CExp Predicted task label and concept explanations (lay-person study only).
CExp+Int CExp plus Intervention capability.
CExp+Int+SMap CExp+Int plus Saliency maps.
With Interventions (WithInt) Participants who not perform interventions or samples where interventions were performed.
No Interventions (NoInt) Participants who did not perform interventions or samples where no interventions were performed.

Acc and Inacc are placed before the model output and feature capabilities. WithInt and NoInt are placed after model output and feature capabilities. For example, participants using the accurate Blackjack model with concept explanations and interventions, and who performed interventions would be referred to as Acc-CExp+Int-WithInt.

3.1 Human Study Design

Expert study participants were asked to diagnose skin conditions in 10 images from the Skincon dataset [4]. We excluded images that were out of focus and limited images to those with the label "malignant melanoma" and "seborrhoeic keratosis" as a dermatologist would typically look to diagnose a patient. The images were shown in a random order.



(a) Expert study (b) Lay-person study

1: Sample image
2: AI agent output class and concept labels
3: Label selection and AI use options for the participant to select
4: Concept outputs, salience map, and intervention slider
5: Game rules, participant score, and game counter (lay-person study)

Fig. 2. Study interfaces with key components labeled.

For the lay-person study, each participant played 15 games of Blackjack, where the first game was without the model enabled while the other 14 included model predictions. Like with the expert study, the model suggested actions for participants to take. Each game had a maximum between 1 and 7 moves (depending on the cards dealt) with cards drawn from a single deck of cards. We removed betting and added a score which increased or decreased based on the number of wins and losses. Participants could select one of three moves: hit, stand, or surrender.

For each sample labelled/move made participants selected how they used the model. These options were: (1) I was influenced by the AI's suggestion, (2) I was influenced by the concepts the AI detected, and (3) I was not influenced by the

AI. These were designed to capture whether participants selected labels based on the model's outputs or disregarded them.

Depending on the explanation group, participants were provided with the model's task and concept predictions, saliency maps, and an intervention slider for each concept. Adjusting any intervention slider automatically updated the model's predicted task label. An example of the interfaces is shown in Fig. 2.

At the end of the study, participants completed a closing survey where they were asked to complete questions asking about the model explanations.

4 Experiment Set-Up

The studies share a number of similarities including interfaces and model capabilities. The studies also apply findings from [23] by reducing the number of concepts initially shown to participants by placing them into a scrollable list.

Both studies start with a short demographic survey asking participants for their age, gender, computer science experience and skin disease identification/blackjack experience. Computer science experience and skin disease identification/blackjack experience are recorded using a Likert scale [17]. Next, participants were briefed on how the model works at a high level and followed a tutorial so they know how to participate in the study and interact with the model. Following this participants compete the study, and finally complete a closing survey.

4.1 Datasets and Models

Expert Study[1]: The expert study uses a model trained on the Skincon dataset [4], a combination of Fitzpatrick 17k [8] and DDI [3]. This is a real-world dataset with 48 clinical concepts, of which we have kept 22 that occur 50 or more times. Concepts were selected by two dermatologists using standard descriptive terms such as "plaque" and "scale". We have provided an example sample with concept annotation in Fig. 3a . For task labels, we used the malignant label. We kept the original dataset splits with 10 samples removed from the training and validation splits for the study. In total, we used 2574 train samples, 644 validation samples and 656 test samples.

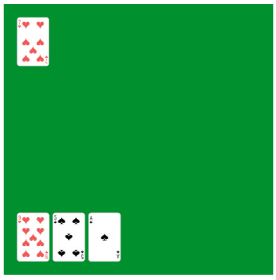
The Skincon model used a Densenet121 architecture [10] for the concept encoder which was initialised with pre-trained weights from ImageNet, and two linear layers with a ReLU activation function for the task predictor which was not pre-trained. The concept encoder was trained to maximise the AUC of concept predictions, while the task predictor was trained to minimise task loss. The concept encoder was trained with a learning rate of 0.00053, a SGD optimiser and trained for 100 epochs. The task predictor was trained with a learning rate of 0.0593, an Adam optimiser and trained for 100 epochs. The resulting model had a concept accuracy of 91.235% and a task accuracy of 88.474%. For the samples included in the study, the model was 70% accurate.

[1] Expert Study: https://github.com/JackFurby/skin-cbm-study.

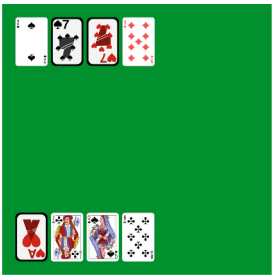
Lay-Person Study[2]: For the lay-person study we created the dataset Blackjack[3] which is similar to Playing cards [7]. Concepts represent the sum of card values in the player's hand, whether the player has an "Ace" card with the value 11, the dealer's first card, and if the dealer has multiple cards. Task labels represent the best move available to the player according to the single deck strategy guide [25]. These labels are *hit* (player gets another card), *stand* (player ends the game with their current cards), *surrender* (player forfeits their hand for a smaller loss), and *bust* (player's cards sums to over 21).

We created two versions *standard Blackjack*, and *mixed Blackjack*. Standard Blackjack uses one style of playing cards, whereas mixed Blackjack uses a different style for all "Ace" and "Seven" cards. This allowed us to artificially reduce the accuracy of a model trained on mixed Blackjack if tested on standard Blackjack samples. Each dataset variation has 10,000 samples which are split into training samples and test samples with a 70%-30% split respectively. Example samples can be seen in Figs. 3b and 3c .

(a) Skincon sample with the present concepts "Plaque" and "Patch"



(b) Standard Blackjack sample



(c) Mixed Blackjack sample

Fig. 3. Example samples from the datasets.

Blackjack models used a VGG-11 architecture with batch normalisation [26] for the concept encoder and two linear layers with a ReLU activation function for the task predictor. Blackjack models were trained to minimise the concept and task loss. The concept encoder was trained with a learning rate of 0.02 using a SGD optimiser. The task predictors had a learning rate of 0.01, used used an Adam optimiser. The models were trained for 200 epochs. The standard Blackjack model achieved a concept accuracy of 99.818% and a task accuracy of 98.874%. The mixed Blackjack model achieved a concept accuracy of 96.434% and a task accuracy of 81.306%.

[2] Lay-person study: https://github.com/JackFurby/blackjack-cbm-study.
[3] Blackjack dataset: https://huggingface.co/datasets/JackFurby/blackjack.

4.2 Evaluation Methodology

In our studies, we analysed interventions, trust, interpretability, and human-machine performance. To understand when interventions are made we have classified them into two categories: *error correction* and *feature adjustment*. Error correction interventions are concepts that are intervened a maximum of once per sample where the intervened concept value $\bar{c}$ is in the range $0 \leq \bar{c} \leq 0.1$ or $0.9 \leq \bar{c} \leq 1$. Feature adjustments are all other interventions, including concepts that are intervened more than once in a sample, or where the intervened concept value is in the range $0.1 < \bar{c} < 0.9$. Feature adjustment interventions are when the participant is not certain the model has incorrectly predicted the presence of a concept, or where they are inspecting how concepts change task label predictions.

We have also assigned the following labels to understand how the concept value changes with interventions:

- Binary change: A concept changes from present to not present, or vice versa.
- Changed model task label: An intervention changes the predicted task label.
- Magnitude: How much interventions changes concept values by.
- Cumulative Change: The total difference between model predicted concept values to the final intervened concept values.
- Reversal: Whether final intervened concept values are close to initial concept values.

We also tracked interventions over time to evaluate if the rate of interventions increased or decreased. For trust, we evaluated participant and model task label alignment. If the model and human task labels are the same for a large proportion of samples we can argue the human participants trust the model [16,29]. We also analysed team performance in comparison to ground truth labels from the dataset.

We repeated the test-time intervention metric [13] to measure the change in task accuracy and concepts after interventions. This metric measures the change in model task accuracy when concepts are intervened on. Unlike the original evaluation with CBM [13], our results used human-performed interventions.

At the end of the study participants answered SCS [9] questions to measure quality for AI explanations. Specifically these questions use a Likert scale and ask participants how they perceived the models ability to answer "why" it made task and concept predictions.

5 Results

The expert study demographic survey showed participants either agreed or strongly agreed that they can diagnose skin diseases from images. Computer experience was evenly distributed between strongly disagree and agree. In the lay-person study participant demographic showed computer experience increased with blackjack experience.

In the expert study, 120 samples were labelled with 93 interventions performed by 7 out of the 12 participants. Meanwhile, in the lay-person study, 1,456 games of blackjack were played with model move suggestions, and 104 games were played without a model output. 243 interventions were performed by 23 participants out of 52 who had the capability to do so. 65.4% of interventions were performed on the inaccurate model, and 34.6% of interventions performed on the accurate model.

5.1 Intervention Classification

Table 2. Breakdown of interventions performed in the studies.

Data subset	Total interventions (count)	Error correction (count)	Feature adjustment (count)	Interventions per sample (count)	Concept intervened per sample (count)	Binary (count)	Changed model task label (count)	Reversal (count)	Mean intervention magnitude (normalized value)	Mean cumulative magnitude (normalized value)
Expert study										
All	93	48	45	2.91	2.50	44	14	11	0.48	0.5
CExp+Int	82	48	34	3.04	2.78	41	12	6	0.49	0.52
CExp+Int+SMap	11	0	11	2.20	1	3	2	5	0.39	0.11
Lay-person study										
All	243	83	160	4.05	2.42	152	29	60	0.58	0.53
Acc-CExp+Int	55	11	44	3.93	2.71	38	8	7	0.59	0.57
Inacc-CExp+Int	73	47	26	3.48	2.57	41	8	20	0.56	0.52
Acc-CExp+Int+SMap	29	0	29	4.83	1.50	17	5	8	0.52	0.11
Inacc-CExp+Int+SMap	86	25	61	4.53	2.32	56	8	25	0.62	0.58

Table 2 summarises interventions. Expert study CExp+Int participants performed 58.5% error correction interventions and 41.5% feature adjustment interventions. 17.6% of interventions were reversed. 2.78 concepts were intervened per

sample with at least one intervention with each intervention changing a concept value by approximately 0.5. Half of all interventions change a concept's presence. 12 interventions changed the model's task prediction, suggesting the model is not sensitive to the concepts participants intervened on.

CExp+Int+SMap participants performed fewer interventions with all being feature adjustments and almost all interventions reversed. This indicates all interventions were performed to explore a model's concept sensitivity.

Lay-person study participants demonstrated similar trends with CExp+Int+SMap participants performing more feature adjustment interventions while CExp+Int participants performed a mix of both intervention types. Those with an accurate model primarily performed feature adjustments, aligning with model sensitivity exploration, while those with an inaccurate model split interventions nearly evenly between error correction (47.2%) and feature adjustments (52.7%).

CExp+Int+SMap participants increased intervention frequency (4.53–4.83 vs. 3.48–3.93 concept intervened), though reduced the number of interventions per sample (1–2 vs. 2–3). Binary interventions were common with inaccurate models, likely due to increased error corrections (47 vs. 11 concepts intervened). High reversal rates for inaccurate models suggest participants required additional interventions to refine their mental model, possibly indicating a misalignment between participants and the model's decision processes. E.g. if the participant plays with a different strategy.

5.2 Human-Machine Task Alignment

We measured the alignment between participant-selected task labels and model predicted task labels to determine if participants trusted the model or not. To further reinforce if alignment is helping the human-machine team, we also compare team accuracy to determine if trust is justified.

Table 3 shows expert study alignment. As there is no guarantee the final model task prediction is the model prediction that a participant aligns with, the overall alignment includes agreement with initial and post-intervention task labels. Alignment for this ranges from 60% to 81.8%. Initial alignment reflects the model's original label and is consistently higher for participants without interventions, final alignment only includes the model's last task label after interventions, and intermediate alignment includes model labels between the initial model task prediction and final task prediction.

A slight decline in alignment is observed with WithInt participants (78.1% vs. 81.8%) and CExp+Int+SMap participants (80% vs. 81.7%). These findings suggest interventions influence participants' labelling decisions, reducing agreement with the model. In fact, initial alignment with interventions (65.6%) is closer to the model's actual accuracy (70%). This indicates interventions help participants calibrate trust to align with the model's true accuracy. Meanwhile, NoInt participants appear to over-trust the model. A one-tailed t-test reveals statistical significance with a p-value of 0.03, below the 0.05 threshold, confirming that the absence of interventions led to increased alignment in this study.

Table 3. Expert study human-machine task alignment.

Data subset	Overall (%)	Initial model task prediction (%)	Intermediate model task prediction (%)	Final model task prediction (%)
All	80.8 (±3.6)	77.5 (±3.8)	77.8 (±8.2)	65.6 (±8.5)
CExp+Int	**81.7 (±5.0)**	76.7 (±5.5)	81.8 (±8.4)	70.4 (±9.0)
CExp+Int+SMap	80.0 (±5.2)	78.3 (±5.4)	60.0 (±24.5)	40.0 (±24.5)
NoInt	**81.8 (±4.1)**	81.8 (±4.1)	–	–
WithInt	78.1 (±7.4)	65.6 (±8.5)	77.8 (±8.2)	65.6 (±8.5)
CExp+Int-NoInt	**81.8 (±6.8)**	81.8 (±6.8)	–	–
CExp+Int-WithInt	81.5 (±7.6)	70.4 (±9.0)	81.8 (±8.4)	70.4 (±9.0)
CExp+Int+SMap-NoInt	**81.8 (±5.2)**	81.8 (±5.2)	–	–
CExp+Int+SMap-WithInt	60.0 (±24.5)	40.0 (±24.5)	60.0 (±24.5)	40.0 (±24.5)

However, the difference in alignment between CExp+Int and CExp+Int+SMap participants was not statistically significant (p-value of 0.59), suggesting that saliency maps had no meaningful impact on alignment. However, a larger study may confirm otherwise.

Alignment for the lay-person study is shown in Table 4 averages to 77.3%, which is lower than the model task accuracy (99.8% for the accurate model and 96.4% for the inaccurate model). However, participants may have playing strategies misaligned with the training data. Alignment in this study differs from the expert study as interventions increase alignment (86.7% vs. 77.1%). Further, participants using the accurate model consistently had a higher alignment than the inaccurate model.

Across all participant groups, interventions improve human-machine alignment. Alignment also increases from the initial model prediction to the final model prediction for most participant subsets. A one-tailed t-test supports this by rejecting the null hypothesis (no alignment increase) and accepting the alternative (interventions improve alignment), with p-values of 0.041 for all participants and 0.04 for those capable of performing interventions, both below the 0.05

Table 4. Lay-person study human-machine task alignment.

Data subset	Overall (%)	Initial model task prediction (%)	Intermediate model task prediction (%)	Final model task prediction (%)
All	77.3 (±0.8)	77.1 (±0.8)	81.5 (±5.3)	83.3 (±4.9)
NoInt	77.1 (±0.8)	77.1 (±0.8)	–	–
WithInt	**86.7 (±4.4)**	76.7 (±5.5)	81.5 (±5.3)	83.3 (±4.9)
Acc	**80.1 (±1.1)**	80.0 (±1.1)	93.8 (±6.2)	90.0 (±6.9)
Inacc	74.6 (±1.2)	74.3 (±1.2)	76.3 (±7.0)	80.0 (±6.4)
Acc-NoExp	**79.8 (±2.2)**	79.8 (±2.2)	–	–
Inacc-NoExp	70.4 (±2.6)	70.4 (±2.6)	–	–
Acc-CExp	**84.5 (±2.0)**	84.5 (±2.0)	–	–
Inacc-CExp	73.9 (±2.5)	73.9 (±2.5)	–	–
Acc-CExp+Int-NoInt	78.8 (±2.4)	78.8 (±2.4)	–	–
Acc-CExp+Int-WithInt	**92.9 (±7.1)**	78.6 (±11.4)	90.0 (±10.0)	92.9 (±7.1)
Inacc-CExp+Int-NoInt	74.8 (±2.5)	74.8 (±2.5)	–	–
Inacc-CExp+Int-WithInt	**76.2 (±9.5)**	66.7 (±10.5)	73.7 (±10.4)	71.4 (±10.1)
Acc-CExp+Int+SMap-NoInt	76.0 (±2.5)	76.0 (±2.5)	–	–
Acc-CExp+Int+SMap-WithInt	**100 (±0.0)**	100 (±0.0)	100 (±0.0)	83.3 (±16.7)
Inacc-CExp+Int+SMap-NoInt	78.5 (±2.4)	78.5 (±2.4)	–	–
Inacc-CExp+Int+SMap-WithInt	**89.5 (±7.2)**	78.9 (±9.6)	78.9 (±9.6)	89.5 (±7.2)

significance threshold. Additionally, comparing CExp participants to NoExp participants resulted in a p-value of 0.036, showing concept explanations also result in a higher human-machine alignment. These results show both interventions and concept explanations increase human-machine task alignment.

If appropriate trust is given to the model, we should expect the human-machine accuracy to be higher than the model alone. Table 5 shows human-machine accuracy for the expert study. Accuracy is averaged by participants, assuming that they build a mental model of the model over time. Even if an indi-

Table 5. Expert study human-machine task accuracy.

Data Subset	Overall Accuracy (%)	Malignant Melanoma(%)	Seborrheic Keratosis (%)
All	78.3 (±2.4)	88.3 (±4.6)	68.3 (±3.9)
CExp+Int	75.0 (±3.4)	83.3 (±8.0)	66.7 (±4.2)
CExp+Int+SMap	**81.7 (±3.1)**	93.3 (±4.2)	70.0 (±6.8)
NoInt	78.0 (±3.7)	88.0 (±8.0)	68.0 (±8.0)
WithInt	**78.6 (±3.4)**	88.6 (±5.9)	68.6 (±4.0)
CExp+Int-NoInt	75.0 (±5.0)	80.0 (±20.0)	70.0 (±10.0)
CExp+Int-WithInt	75.0 (±5.0)	85.0 (±9.6)	65.0 (±5.0)
CExp+Int+SMap-NoInt	80.0 (±5.8)	93.3 (±6.7)	66.7 (±13.3)
CExp+Int+SMap-WithInt	**83.3 (±3.3)**	93.3 (±6.7)	73.3 (±6.7)

vidual prediction is ignored, they may still influence participants. For instance, participants may recognise when the model is incorrect without the need to perform interventions.

In the expert study CExp+Int participants achieved an accuracy of 75% and CExp+Int+SMap participants achieved an accuracy of 81.7%, indicating that the additional information provided by saliency maps aids decision-making. WithInt Participants had a slightly higher accuracy than NoInt participants (78.6% vs. 78%), suggesting interventions either match or slightly enhance participant performance. However, the expert study lacks the sample size to show statistical significance. A one-tailed t-test resulted in a p-value of 0.09 for accuracy being higher if participants had saliency maps, and a p-value of 0.46 if participants performed interventions.

In the lay-person study (Table 6) the largest increase in task accuracy came from participants using the accurate model compared to the inaccurate model. Interventions only increased human-machine accuracy for participants using the inaccurate model. This suggests participants are over-trusting the model, expectedly considering interventions increased alignment as just discussed in Table 4. Despite this, the lowest task accuracy was achieved by Inacc-CExp and Inacc-NoExp participants and participants with the AI disabled. Therefore interventions are still showing signs of increasing human-machine accuracy.

A one-tailed t-test resulted in a p-value of 0.042 showed CExp improved task accuracy compared to NoExp. In contrast, participants who performed inter-

Table 6. lay-person study human-machine task accuracy averaged by participant.

Data Subset	Accuracy (%)
AI disabled	74.4 (±3.9)
All	83.6 (±0.9)
WithInt	83.3 (±1.1)
NoInt	**84.7 (±1.8)**
Acc	**84.6 (±2.7)**
Inacc	78.1 (±3.2)
Acc-NoExp	**84.6 (±2.7)**
Inacc-NoExp	78.1 (±3.2)
Acc-CExp	**91.0 (±2.4)**
Inacc-CExp	81.4 (±1.6)
Acc-CExp+Int-NoInt	**86.6 (±3.3)**
Acc-CExp+Int-WithInt	83.8 (±2.9)
Inacc-CExp+Int-NoInt	75.7 (±4.6)
Inacc-CExp+Int-WithInt	**84.3 (±1.8)**
Acc-CExp+Int+SMap-NoInt	**83.4 (±2.4)**
Acc-CExp+Int+SMap-WithInt	**83.4 (±6.4)**
Inacc-CExp+Int+SMap-NoInt	83.5 (±2.9)
Inacc-CExp+Int+SMap-WithInt	**86.6 (±3.6)**

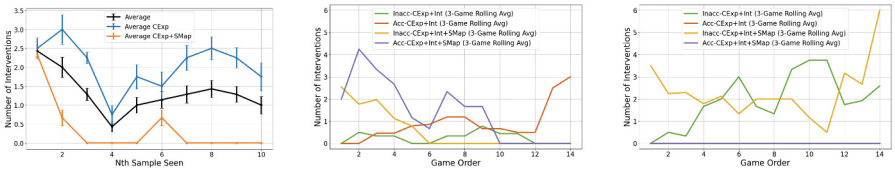
ventions did not achieve a statistically significant improvement in task accuracy compared to those who did not, achieving a p-value of 0.27 when compared to all NoExp and CExp participant groups and 0.195 when compared to participants who did not perform interventions but had the capability to do so. While we observed a trend of higher accuracy among participants using interventions, we cannot conclude that interventions directly improve task accuracy.

Overall, our findings indicate that concepts are beneficial for improving human-machine alignment and human-machine task accuracy. Interventions are beneficial for increasing human-machine task alignment but do not result in a statistically significant increase in task accuracy and can result in over-trust.

5.3 Interventions Over Time

We may expect the number of interventions to decrease over time as participants learn about a model's sensitivity to concepts. We found this is the case in both the expert study (Fig. 4a) where we show we show the average number of interventions per sample with the standard error, and the layperson study when the model correctly predicts concepts (Fig. 4b) where we show the decline with a rolling average. An exception to the decline is seen with Acc-CExp+Int participants in the lay-person study where there is a spike of interventions performed

on game 13 with an average of 7 and a standard error of ±2. As this occurs once, this spike is not representative of all participants.



(a) Expert study (b) Lay-person study correctly predicted concepts (c) Lay-person study incorrectly predicted concepts

Fig. 4. Interventions performed declined over time except for incorrectly predicted concepts in the lay-person study where the number of interventions performed remains constant.

In the expert study, CExp+Int+SMap participants see a sharp decline in interventions, while CExp+Int participants see an initial decline which recovers for later samples. This suggests saliency map explanations provide additional insights into the model's concept predictions. CExp+Int participants appear to be incentivised to use interventions as a means of understanding the model's decision-making process.

For incorrect concept predictions in the lay-person study, participants consistently performed around 2 interventions per sample. These results demonstrate participants identify concepts that need to be intervened on while ignoring the concepts that are correctly predicted by the models.

Overall, these results show that participants initially explore the model's capabilities and sensitivity to concept values before developing a mental model and reducing the number of interventions performed to where it is required.

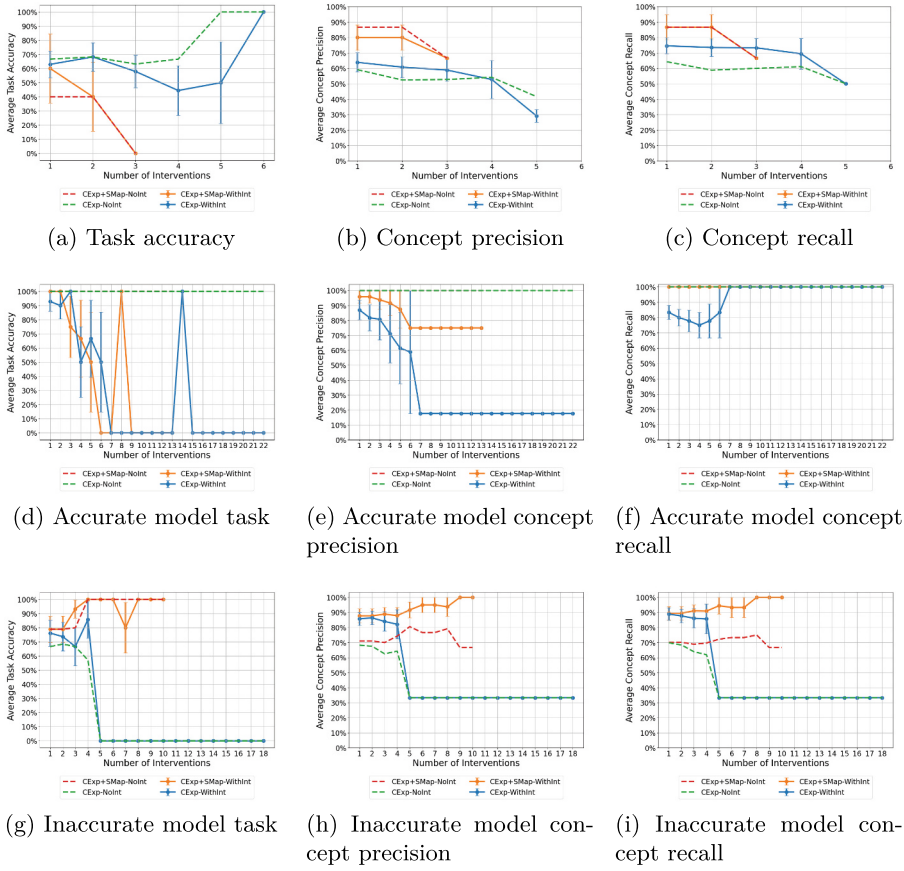
5.4 Test-Time Intervention and Concept Accuracy

The authors of CBMs showed results using the metric test-time intervention where interventions updating concept predictions with ground truth concept values improved model task performance [13]. However, it remains unknown how interventions improve model task performance when interventions are made by humans.

In the related work section, we discussed [2] where they found CBMs may not apply the same weight to concepts for task labels as humans would. Therefore we hypothesise interventions performed by humans may not see an improvement in task accuracy which would show a misalignment between humans and the model's sensitivity to concepts.

Test-time intervention results are shown in Fig. 5a for the expert study and Figs. 5d and 5g for the lay-person study. Task accuracy is averaged by participant and explanation groups. We only included the same samples between with

interventions and without. For example, if participants intervened on concepts for samples 1 and 3 but not 2, we only work out the task accuracy for samples 1 and 3. As participants performed 2–3 interventions on average in each study conclusions for concept and task accuracy past these intervention counts cannot always be made.



(a) Task accuracy (b) Concept precision (c) Concept recall

(d) Accurate model task (e) Accurate model concept precision (f) Accurate model concept recall

(g) Inaccurate model task (h) Inaccurate model concept precision (i) Inaccurate model concept recall

Fig. 5. Interventions decrease model task accuracy in the expert study and the layperson study accurate model while increasing model task accuracy with the lay-person study inaccurate model. Concept precision and recall increase with interventions in most cases.

Task accuracy, for the most part, does not improve with interventions compared to no interventions. In the expert study between 1–3 interventions task accuracy is close to matching the accuracy of the model with no interventions, and outperforms the model with 1 intervention for CExp+Int+SMap participants. In the lay-person study task accuracy initially declines slightly for the accurate model before sharply declining, although with large error bars. The

exception to declining task accuracy is observed for the lay-person study inaccurate model where interventions increase or match the model's initial task accuracy.

In addition to task accuracy, we have also measured the change in concept accuracy. Figure 5b and Fig. 5c shows the precision and recall for the expert study, Fig. 5e and Fig. 5f for the lay-person study accurate model, and Fig. 5h and Fig. 5i for the inaccurate model. Notably, in the expert study and lay-person study with the inaccurate model, interventions lead to an increase or matching the precision and recall. In the lay-person study with the accurate model, precision is lower than the model concept prediction, while recall initially declines before rising the match the model. Most of these results also have no overlapping error bars between the model-predicated concepts and intervened concepts.

Table 7. Likert Scores for SCS questions.

Question	All	CExp+Int	CExp+Int+SMap	WithInt	NoInt	Skin Experience Agree	Skin Experience Strongly Agree
Factors in data	3.09	3.00	3.20	3.00	3.20	2.80	**3.33**
Understood	3.73	3.33	4.20	3.67	3.80	3.60	**3.83**
Change detail level	3.18	**3.67**	2.60	3.50	2.80	2.80	3.50
Need support	3.64	3.33	**4.00**	3.50	3.80	**4.00**	3.33
Understanding causality	3.00	3.17	2.80	**3.33**	2.60	2.80	3.17
Use with knowledge	3.45	3.50	3.40	3.50	3.40	3.00	**3.83**
No inconsistencies	3.00	3.17	2.80	3.00	3.00	2.60	**3.33**
Learn to understand	3.55	3.50	3.60	3.50	3.60	3.20	**3.83**
Needs references	3.73	3.67	3.80	3.50	4.00	3.60	**3.83**
Efficient	3.45	**3.67**	3.20	**3.67**	3.20	3.40	3.50
Overall score	0.68	0.68	0.67	0.68	0.67	0.64	**0.71**

As previously discussed. We expect participants to explore a model's sensitivity to concept values. In the lay-person study accurate model we observe the clearest sign of this. In particular with concept recall where recall initially falls before increasing from 4 interventions. This shows participants appear to

be exploring the concept speck before correcting concept values and making a move.

When combining all test-time intervention results, it becomes clear that CBMs are mostly not aligned to the concepts participants are adjusting. Although interventions often make concept vectors more accurate, task accuracy does not reflect this improvement. This aligns with the findings in [2].

5.5 System Causability Scale

We used the SCS [9] to get a subjective rating of explanation suitability with expert study results presented in Table 7. The overall score, computed as the average of participants' summed responses normalised by the maximum possible score. This score is between 0 and 1 where 0.68 indicates an average response [9]. Almost all overall scores are either 0.68 or slightly below. The sub-section of participants who's overall score exceeded this are participants who selected "strongly agree" as their experience at classifying skin diseases in the demographic survey, with a score of 0.71.

For individual questions, most averaged to be between high 2 and high 3 (Likart options "disagree" and "neutral"). A few questions stand out. Starting with *change detail level* (*I could change the level of detail on demand*) was rated higher for CExp+Int participants, WithInt participants, and participants who self-rated their experience at skin disease identification as "strongly agree" (of which 57% performed interventions). This suggests that if participants perform interventions they understand the information it provides.

Need support (*I did not need support to understand the explanations*) averaged to 4 for both CExp+Int+SMap participants and participants who answered their skin disease identification experience as "agree". For CExp+Int+SMap participants these results indicates the potential benefit saliency maps provide to help participants interpret the model's concept predictions. For skin experience agree participants, 60% of which used interventions (with two of these participants performing almost 60 interventions), suggests their increased interaction with the model improved their understanding of the model.

Finally, *efficient* (*I received the explanations in a timely and efficient manner*) was also consistently rated slightly over 3. WithIntParticipants answered this question with a slightly higher score than NoInt participants.

Responses from the SCS questions for the lay-person study are shown in Table 8. All overall scores are 0.70 or above. The highest score was 0.78 for Acc-CExp participants. Surprisingly, Acc-NoExp participants score was 0.74 which matches some participant groups who had access to concepts and interventions. As each participant only answered questions for one version of the explanations instead of ranking each, this may be attributed to the similar scores.

The SCS results suggest that incorporating concepts improves participants' understanding of causality as shown by the questions *Understanding causality* (*I found the explanations helped me to understand causality*) scoring higher with the inclusion of explanation techniques, although the differences between participant groups was small. Additionally, These results varied between studies where the

Table 8. Lay-person study Likert scores for SCS questions.

Question	All Participants	Acc-NoExp	Inacc-NoExp	Acc-CExp	Inacc-CExp	Acc-CExp+Int	Inacc-CExp+Int	Acc-CExp+Int+SMap	Inacc-CExp+Int+SMap	CExp+Int-WithInt and CExp+Int+SMap-WithInt	CExp+Int-NoInt and CExp+Int+SMap-NoInt
Factors in data	3.39	3.08	2.85	**3.92**	3.69	3.62	3.50	3.23	3.25	2.87	3.25
Understood	4.18	4.08	4.15	**4.31**	**4.31**	4.23	4.17	4.08	4.08	4.13	4.08
Change detail level	2.91	2.69	2.31	2.77	2.38	3.38	**3.50**	3.31	3.00	3.13	3.00
Need support	3.77	3.92	3.92	3.77	3.62	3.54	**4.17**	3.54	3.75	3.61	3.75
Understanding causality	3.51	3.15	3.38	3.62	3.46	3.31	3.75	3.46	**4.00**	3.43	**4.00**
Use with knowledge	3.99	4.15	3.92	4.23	3.62	3.92	3.92	3.85	**4.33**	3.91	**4.33**
No inconsistencies	3.59	3.77	3.38	**4.15**	3.54	3.54	3.08	3.77	3.42	2.96	3.42
Learn to understand	4.06	**4.31**	4.15	4.00	4.15	3.92	4.00	3.92	4.00	3.74	4.00
Needs references	3.64	**4.00**	3.77	3.85	3.46	3.54	3.33	3.54	3.58	3.13	3.58
Efficient	4.24	4.08	3.85	4.15	**4.62**	4.15	4.42	4.08	4.58	4.26	4.58
Overall score	0.75	0.74	0.71	**0.78**	0.74	0.74	0.76	0.74	0.76	0.70	0.76

expert study answered this question inline with the lay-person study with no model explanations. Model outputs were generally well understood as shown by the results for *Understood* (*I understood the explanations within the context of my work.*), and *learn to understand* (*I think that most people would learn to understand the explanations very quickly*).

Overall, while model outputs were generally well understood, reflected in the scores for Understood ("I understood the explanations within the context of my work") and Learn to understand ("I think that most people would learn to understand the explanations very quickly"), there are indications of a possible mismatch between human and machine decision-making. *Understood*, *Learn to understand*, *Use with knowledge*, *No inconsistencies* ("I did not find inconsisten-

cies between explanations"), and *Need references* ("I did not need more references in the explanations: e.g., medical guidelines, regulations") were all scored lower if participants made interventions. Although *no inconsistencies* was also low for the inaccurate model it also shows interventions may be causing confusion over how they relate to task predictions or how they are used instead of aiding in completing the task.

6 Discussion

Before beginning the discussion, we have detailed several limitations in our studies. The expert study had a small sample size, which may limit the generalisability of our findings. To address this, we conducted a larger lay-person study and drew parallels between the two to provide additional context. Additionally, participants in the expert study lacked access to patient history, high-quality diagnostic images, and multiple images of each sample, which may be expected in clinical settings. To mitigate this, we simplified the task to distinguish between "malignant melanoma" and "seborrhoeic keratosis" which have clear visual differences. Finally, in the lay-person study, participants played Blackjack, meaning game success depended partly on luck, though our evaluation focused on optimal moves rather than overall game outcomes.

From our human studies evaluating CBMs, we observed mixed results regarding interpretability and task performance. While our findings reinforce CBMs interpretability with participants who utilised concepts and interventions to explore the concept space and inspect task predictions, task accuracy improvements were inconsistent. Notably, while concept accuracy increased, task accuracy mostly decreased.

6.1 Do Test-Time Interventions Improve Human-Machine Task and Concept Accuracy?

Test-time interventions found mixed results across models. In most cases, task accuracy with interventions matched or underperformed the model's accuracy with no interventions, with further declines in task accuracy as the number of interventions increased. The only notable increases in model task accuracy was seen in the lay-person study with the inaccurate model. Following our test-time intervention results, it suggests interventions have the risk of leading to decreased task accuracy if humans follow model task predictions after interventions are performed.

For concept accuracy, interventions increased accuracy. Despite this not being consistent across all models, decreases in some situations (e.g. accurate blackjack model) were expected to account for participants learning the model's sensitivity to concepts.

Similar to [2], our findings suggest that CBMs task predictions may use different concepts than humans use. Future research should explore methods to align CBM decision-making with human decision-making.

6.2 Do Interventions Increase the Interpretability of Concept Bottleneck Models?

In the expert study, interventions were almost evenly split between error correction and feature adjustments. Further, intervention decreased over time, suggesting that participants relied on interventions less as they developed a mental model of the model's behaviour. This aligns with the idea that CBMs improve interpretability.

Saliency maps decreased the number of interventions performed suggesting they provide sufficient insight into the model's behaviour. However, participants also reported they placed little weight on the model's predictions, implying that participants preferred their own intuition than relying on the model.

Regarding the lay-person study, participants using the accurate models primarily performed feature adjustments, with nearly 75% of interventions involving changes to concept presence. Combined with the eventual increase in concept recall, this indicates participants used the interpretability of concepts to improve their understanding of the model. More interventions, including reversals, were performed with the inaccurate model.

While we observe a decline in interventions over time in both studies, part of this decline may be attributed to the novelty of interventions, with engagement naturally decreasing as participants became more familiar with the task. Although we cannot entirely rule out this effect, the fact that the decline is not uniform, particularly in the lay-person study, where interventions remained higher when concepts were incorrectly predicted suggests that participants were not merely losing interest but actively leveraging interventions to improve their understanding of the model.

Our findings support the claim that CBMs improve interpretability by allowing users to interactively query and adjust concept predictions. However, we have identified this process is limiting as humans are required to seek explanations and iteratively probe the model's concept sensitivity, which may not be practical or obvious for all users or applications. Further, our study does not look at the role of the interface in engaging participants to interact with the model. We suggest future research should look at the delivery of concept explanations to ensure they are efficiently delivered.

6.3 Are Concept Bottleneck Models Trusted?

We chose to use alignment as a proxy for trust [24]. In the expert study, NoInt participants aligned to the model's predictions 81% of the time, which is 11% higher than the model's accuracy on the samples in the study suggesting overtrust. In contrast, WithInt participants were aligned to the model's initial task prediction 66% of the time, 4% lower than the model's accuracy. Alignment then increased by almost 13% after interventions. In addition, accuracy was higher for WithInt participants compared to NoInt participants. This shows interventions increased trust, which itself was better justified compared to participants who did not use interventions.

For the lay-person study, alignment was lower than the model's accuracy before interventions. When participants used interventions, alignment increased significantly across all participant groups. However, this increase in alignment only increased task accuracy for participants using the inaccurate model. This shows the potential for interventions to lead to over-trust. In addition, providing just concept explanations lead to an increase in alignment while also increasing task accuracy.

The trends of alignment and joint task accuracy are conflicting between the studies. We hypothesis this is because in the expert study the model outputs were used purely as a second opinion as the participants would have sufficient expertise in the task domain. As this is not guaranteed in the lay-person study we believe participants may follow the model if they are unsure themselves. This is a concerning point if these models are deployed in situations where humans are not domain experts.

7 Conclusion

In this paper, we ran the first studies to evaluate how humans use CBMs in a collaborative setting. We focused on how concepts are interacted with and the interpretability of these models. In particular, we evaluate (1) if concept interventions increase the model's task accuracy, (2) do concept interventions increase model interpretability, and (3) are CBMs are trusted. We find CBMs do not translate to increased model task accuracy in a human-machine setting, but this model architecture and other CMs are shown to increase both the interpretability and trust with the model's task label predictions.

We drew three main conclusions from our studies:

Firstly, interventions significantly improved concept accuracy but had limited impact on task accuracy. This suggests a misalignment between the concepts humans use and the concepts the models use to label samples. Addressing this misalignment is critical to improving the effectiveness of CMs human-machine teams.

Next, we show the initial promise of interpretability from high-level concepts and interpretability is upheld with CMs. However, as this required participants to engage in interventions, we highlight a need for CBMs to present their decision-making process proactively, reducing the cognitive effort required from humans. In addition, much of the interpretability can be provided by just providing concept predictions.

Finally, using alignment as a proxy for trust, we found that interventions led to higher trust. This did not always lead to increased task accuracy and, as shown in the lay-person study, can result in overtrust. This highlights the importance of interpretable models that are evaluated with human participants to enable the creation of trust that is suitably applied to a model.

Acknowledgments. This research was funded by the UK Engineering and Physical Sciences Research Council (EPSRC) and IBM UK via an Industrial CASE (ICASE) award.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Adebayo, J., Muelly, M., Liccardi, I., Kim, B.: Debugging tests for model explanations. In: Proceedings of the 34th International Conference on Neural Information Processing Systems. NIPS '20, Curran Associates Inc. (2020)
2. Barker, M., Collins, K.M., Dvijotham, K., Weller, A., Bhatt, U.: Selective concept models: Permitting stakeholder customisation at test-time (2023), https://arxiv.org/abs/2306.08424
3. Daneshjou, R., Vodrahalli, K., Novoa, R.A., Jenkins, M., Liang, W., Rotemberg, V., Ko, J., Swetter, S.M., Bailey, E.E., Gevaert, O., Mukherjee, P., Phung, M., Yekrang, K., Fong, B., Sahasrabudhe, R., Allerup, J.A.C., Okata-Karigane, U., Zou, J., Chiou, A.S.: Disparities in dermatology ai performance on a diverse, curated clinical image set. Science Advances **8**(32), eabq6147 (2022). https://doi.org/10.1126/sciadv.abq6147
4. Daneshjou, R., Yuksekgonul, M., Cai, Z.R., Novoa, R., Zou, J.Y.: Skincon: A skin disease dataset densely annotated by domain experts for fine-grained debugging and analysis. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 18157–18167. Curran Associates, Inc. (2022)
5. Doshi-Velez, F., Kim, B.: Towards a rigorous science of interpretable machine learning (2017), https://arxiv.org/abs/1702.08608
6. Dubey, A., Radenovic, F., Mahajan, D.: Scalable interpretability via polynomials. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022), https://openreview.net/forum?id=TwuColwZAVj
7. Furby, J., Cunnington, D., Braines, D., Preece, A.: Can we constrain concept bottleneck models to learn semantically meaningful input features? (2024), https://arxiv.org/abs/2402.00912
8. Groh, M., Harris, C., Soenksen, L., Lau, F., Han, R., Kim, A., Koochek, A., Badri, O.: Evaluating Deep Neural Networks Trained on Clinical Images in Dermatology with the Fitzpatrick 17k Dataset . In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1820–1828. IEEE Computer Society (Jun 2021https://doi.org/10.1109/CVPRW53098.2021.00201
9. Holzinger, A., Carrington, A., Müller, H.: Measuring the Quality of Explanations: The System Causability Scale (SCS). KI - Künstliche Intelligenz **34**(2), 193–198 (2020). https://doi.org/10.1007/s13218-020-00636-z
10. Huang, G., Liu, Z., van der Maaten, L., Weinberger, K.Q.: Densely connected convolutional networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (July 2017).https://doi.org/10.1109/CVPR.2017.243
11. Jeyakumar, J.V., Dickens, L., Cheng, Y.H., Noor, J., Garcia, L.A., Echavarria, D.R., Russo, A., Kaplan, L.M., Srivastava, M.: Automatic concept extraction for concept bottleneck-based video classification (2022), https://openreview.net/forum?id=66kgCIYQW3

12. Jeyakumar, J.V., Sarker, A., Garcia, L.A., Srivastava, M.: X-char: A concept-based explainable complex human activity recognition model. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **7**(1) (mar 2023).https://doi.org/10.1145/3580804
13. Koh, P.W., Nguyen, T., Tang, Y.S., Mussmann, S., Pierson, E., Kim, B., Liang, P.: Concept bottleneck models. In: III, H.D., Singh, A. (eds.) Proceedings of the 37th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 119, pp. 5338–5348. PMLR (13–18 Jul 2020), http://proceedings.mlr.press/v119/koh20a.html
14. Kulesza, T., Burnett, M., Wong, W.K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces. p. 126–137. IUI '15, Association for Computing Machinery (2015).https://doi.org/10.1145/2678025.2701399
15. Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., Wong, W.K.: Too much, too little, or just right? ways explanations impact end users' mental models. In: 2013 IEEE Symposium on Visual Languages and Human Centric Computing. pp. 3–10 (2013https://doi.org/10.1109/VLHCC.2013.6645235
16. Lai, V., Tan, C.: On human predictions with explanations and predictions of machine learning models: A case study on deception detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency. p. 29–38. FAT* '19, Association for Computing Machinery (2019).https://doi.org/10.1145/3287560.3287590
17. Likert, R.: A technique for the measurement of attitudes. Archives of Psychology (1932)
18. Lockhart, J., Magazzeni, D., Veloso, M.: Learn to explain yourself, when you can: Equipping concept bottleneck models with the ability to abstain on their concept predictions (2022), https://arxiv.org/abs/2211.11690
19. Mahinpei, A., Clark, J., Lage, I., Doshi-Velez, F., WeiWei, P.: Promises and pitfalls of black-box concept learning models. In: proceeding at the International Conference on Machine Learning: Workshop on Theoretic Foundation, Criticism, and Application Trend of Explainable AI,. vol. 1, pp. 1–13 (2021)
20. Midavaine, N., Go, G.H.T., Canez, D., Simion, I., Chatterji, S.: [re] on the reproducibility of post-hoc concept bottleneck models. Transactions on Machine Learning Research (2024)
21. Mysore, S., Jasim, M., Mccallum, A., Zamani, H.: Editable user profiles for controllable text recommendations. In: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval. p. 993–1003. SIGIR '23, Association for Computing Machinery (2023).https://doi.org/10.1145/3539618.3591677
22. Nguyen, G., Taesiri, M.R., Kim, S.S.Y., Nguyen, A.: Allowing humans to interactively guide machines where to look does not always improve human-ai team's classification accuracy. In: The 3rd Explainable AI for Computer Vision (XAI4CV) Workshop at CVPR 2024 (2024)
23. Ramaswamy, V.V., Kim, S.S.Y., Fong, R., Russakovsky, O.: Overlooked factors in concept-based explanations: Dataset choice, concept learnability, and human capability. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10932–10941 (June 2023)
24. Rong, Y., Leemann, T., Nguyen, T., Fiedler, L., Qian, P., Unhelkar, V., Seidel, T., Kasneci, G., Kasneci, E.: Towards human-centered explainable ai: A survey of user studies for model explanations. IEEE Transactions on Pattern Analysis and Machine Intelligence **46**(04), 2104–2122 (apr 2024).https://doi.org/10.1109/TPAMI.2023.3331846

25. Shackleford, M.: Blackjack Single Desk Strategy (10 2023), https://wizardofodds.com/games/blackjack/strategy/1-deck/
26. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations (2015)
27. Sixt, L., Schuessler, M., Popescu, O.I., Weiß, P., Landgraf, T.: Do users benefit from interpretable vision? a user study, baseline, and dataset. In: International Conference on Learning Representations (2022), https://openreview.net/forum?id=v6s3HVjPerv
28. Wang, B., Li, L., Nakashima, Y., Nagahara, H.: Learning bottleneck concepts in image classification. In: 2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10962–10971. IEEE Computer Society (jun 2023https://doi.org/10.1109/CVPR52729.2023.01055
29. Wang, X., Yin, M.: Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In: Proceedings of the 26th International Conference on Intelligent User Interfaces. p. 318–328. IUI '21, Association for Computing Machinery (2021).https://doi.org/10.1145/3397481.3450650
30. Yuksekgonul, M., Wang, M., Zou, J.: Post-hoc concept bottleneck models. In: ICLR 2022 Workshop on PAIR2Struct: Privacy, Accountability, Interpretability, Robustness, Reasoning on Structured Data (2022)
31. Zarlenga, M.E., Barbiero, P., Ciravegna, G., Marra, G., Giannini, F., Diligenti, M., Shams, Z., Precioso, F., Melacci, S., Weller, A., Lio, P., Jamnik, M.: Concept embedding models: beyond the accuracy-explainability trade-off. In: Proceedings of the 36th International Conference on Neural Information Processing Systems. NIPS '22, Curran Associates Inc. (2024)