nature machine intelligence

Article

https://doi.org/10.1038/s42256-023-00612-w

A typology for exploring the mitigation of shortcut behaviour

Received: 10 June 2022

Accepted: 9 January 2023

Published online: 09 March 2023

Check for updates

Felix Friedrich [1,2] ✉, Wolfgang Stammer [1,2], Patrick Schramowski [1,2,3,4] & Kristian Kersting [1,2,4,5]

As machine learning models become larger, and are increasingly trained on large and uncurated datasets in weakly supervised mode, it becomes important to establish mechanisms for inspecting, interacting with and revising models. These are necessary to mitigate shortcut learning effects and to guarantee that the model's learned knowledge is aligned with human knowledge. Recently, several explanatory interactive machine learning methods have been developed for this purpose, but each has different motivations and methodological details. In this work, we provide a unification of various explanatory interactive machine learning methods into a single typology by establishing a common set of basic modules. We discuss benchmarks and other measures for evaluating the overall abilities of explanatory interactive machine learning methods. With this extensive toolbox, we systematically and quantitatively compare several explanatory interactive machine learning methods. In our evaluations, all methods are shown to improve machine learning models in terms of accuracy and explainability. However, we found remarkable differences in individual benchmark tasks, which reveal valuable application-relevant aspects for the integration of these benchmarks in the development of future methods.

Trust is considered to be 'reliance on the integrity, strength, ability, surety, etc., of a person or thing'[1], but how reliable are machine learning (ML) models and do they in fact base their decisions on correct reasoning? These questions emerge as ML becomes more present in our daily lives and, importantly, high-stakes environments (for example, for disease detection[2]), making it more and more necessary for humans to rely on such machines. However, deep neural networks (DNNs), which are considered state-of-the-art models for many tasks, in particular show an inherent lack of transparency regarding the underlying decision process for their predictions as well as fundamental issues concerning robustness (for example, slight input perturbations can lead to very different model predictions). Solutions to these issues are considered integral components for trust development in current and future artificial intelligence (AI) systems[3]. In particular, this first issue becomes ever more important for identifying shortcut behaviour[4] as the latest trend of DNNs—large-scale pretrained models, such as GPT-3 and DALL·E 2 (refs. [5,6])—employ huge numbers of unfiltered data that contain biases and can lead to negative societal impacts if left unchecked[7,8].

Consequently, explainable AI (XAI) was introduced to address this lack of transparency[9,10]. Via such explainer methods proposed by XAI research, recent works have revealed that DNNs can show Clever Hans behaviour—making use of confounders—due to spurious correlations in the data[11]. However, only making such models explainable can be insufficient for properly building trust as well as for the overall deployability of a model, as this does not offer the possibility to revise incorrect and hazardous behaviour. For this reason, the explanatory interactive machine learning (XIL) framework[12] was proposed to promote a more fruitful approach to communication between humans and machines, allowing for a more complementary approach.

[1]Computer Science Department, Artificial Intelligence and Machine Learning Lab, Technical University of Darmstadt, Darmstadt, Germany. [2]Hessian Center for Artificial Intelligence (hessian.AI), Darmstadt, Germany. [3]LAION, Hamburg, Germany. [4]German Center for Artificial Intelligence (DFKI), Darmstadt, Germany. [5]Centre for Cognitive Science, Technical University of Darmstadt, Darmstadt, Germany. ✉e-mail: friedrich@cs.tu-darmstadt.de

https://doi.org/10.1038/s42256-023-00612-w

Specifically, in XIL, a model makes a prediction and presents its corresponding explanation to the user, who responds by providing corrective feedback, if necessary, on the prediction and explanation. It has been shown that XIL can improve performance and explanations, that is, help overcome Clever Hans behaviour and improve generalization to unseen data[13]. Moreover, interaction through explanations is considered a natural form of bidirectional communication between human experts, making XIL methods effective protocols to open black boxes. In this way, XIL methods may fill the trust gap between ML systems and human users[14].

Unfortunately, existing XIL methods[12,13,15–18] were developed independently and often with different motivations. In these works, evaluations of the effectiveness of a method often reverted to qualitative explanation evaluations and test accuracy on separate known confounded datasets. However, these evaluation measurements do not unveil essential methodological characteristics, which are particularly important for the practical use case. Furthermore, currently, no study exists that covers a comprehensive comparison of relevant XIL methods. Therefore, in this work, we provide a typology for XIL and propose that existing methods can in fact be summarized via a common underlying terminology. Hereby a method's individual differences correspond to specific instantiations of the basic modules. We additionally propose an extensive set of evaluation criteria, consisting of new measures and tasks, for extensively benchmarking current and future XIL methods based on our typology. This includes the robustness of a method to faulty user feedback and its efficiency in terms of the number of required interactions. Thus, in this work, we provide an extensive study of six recent XIL methods based on these various criteria.

In summary, our main contributions are (1) unifying existing XIL methods into our typology with a single common terminology, (2) extending the typology by introducing measures and tasks to benchmark XIL approaches, (3) evaluating existing methods based on these various criteria that are of great relevance for real-world applicability and (4) identifying unresolved issues to motivate future research.

Algorithm 1. XIL takes as input sets of annotated examples A and non-annotated examples N, and iteration budget T.

$f \leftarrow \text{Fit}(A)$
repeat
  $X \leftarrow \text{SELECT}(f, A)$
  $\hat{y} \leftarrow f(X)$
  $\hat{E} \leftarrow \text{EXPLAIN}(f, X, \hat{y})$
  Present X, $\hat{y}$ and $\hat{E}$ to the user
  $\bar{y}, \bar{C} \leftarrow \text{OBTAIN}(X, \hat{y}, \hat{E})$
  $A \leftarrow A \cup \{(X, \bar{y}, \bar{C})\}$
  $f \leftarrow \text{REVISE}(A)$
  $N \leftarrow N \setminus \{X\}$
until budget T is exhausted or f is good enough
return f

Explanatory interactive machine learning

To examine XIL, in the following, we present a typology for XIL (Fig. 1) based on Algorithm 1. We describe its modules in detail and use them as a foundation for our evaluations. Moreover, we use our typology to examine present XIL methods and thereby investigate the different modules to uncover limitations and suggest avenues for future work.

A unified XIL typology

In general, XIL combines explanation methods (XAI) with user supervision (active learning) of the model's explanations to revise the model's learning process interactively. Notably, XIL can be considered a subfield of AI methods that leverage explanations into the learning process (for example, see ref. [19] for a comprehensive overview of methods that leverage explanations into interactive ML). The conceptual function can be described as follows: XAI focuses on generating explanations from the model, whereas XIL aims to reverse the flow and inserts user feedback on these explanations back into the model. The goal is to establish trust in the model's predictions not only by revealing false and potentially harmful behaviour of a model's reasoning process via the model's explanations but also to give the user the possibility to correct this behaviour via corrections of these explanations.

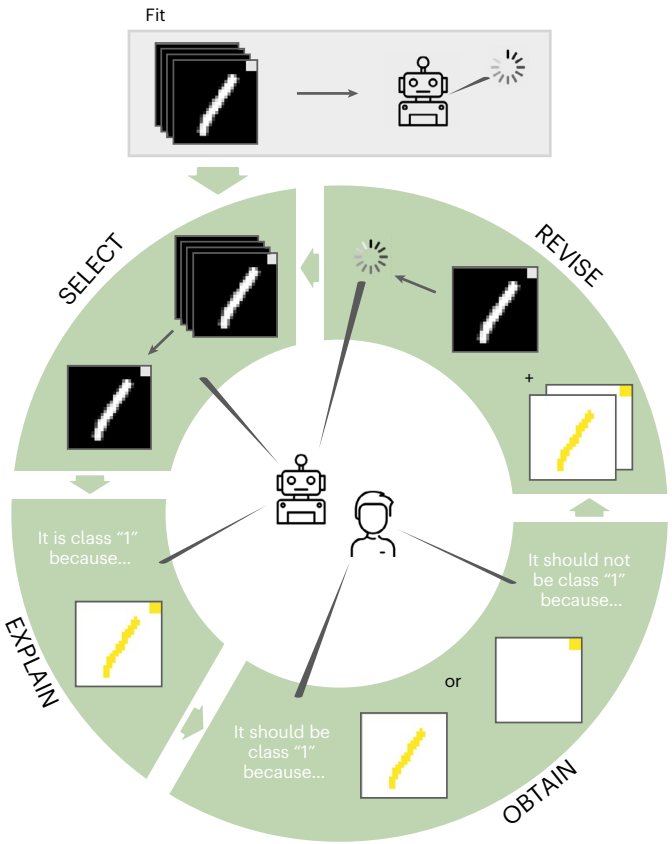

Fig. 1 | The XIL typology. Flowchart visualizing Algorithm 1. SELECT describes how samples X from N are selected in XIL. EXPLAIN depicts how the model provides insights into its reasoning process to the teacher. With OBTAIN, the teacher, in turn, observes whether the learner's prediction is right or wrong, and especially if it is based on the right or wrong reason, and returns corrective feedback, if necessary. The (explanatory) corrections obtained are redirected back into the model's learning process with the REVISE module to correct the model behaviour regarding the user.

Algorithm 1 describes the XIL setting in pseudocode. It uses a set of annotated examples A, a set of non-annotated examples N and an iteration budget T. The annotation comprises both the classification label y and explanation E, that is, a non-annotated example is missing one or both. In general, the procedure is illustrated in Fig. 1 and can be compared to a teacher–learner setting. Active learning is a learning protocol in which the model sequentially presents non-annotated examples (SELECT) from a data pool to an oracle (for example, human annotator) that labels these instances (OBTAIN). Accordingly, this setting allows the user to influence the learning process actively (REVISE). Although the active learning setting enables simplistic interaction between the model and a user, it does not promote trust if explanations do not accompany predictions[12]. The lack of explanations in active learning, however, makes it difficult for the user to comprehend the model's decision process and provide corrections. Therefore, the XIL typology extends the learning pipeline with XAI (EXPLAIN). Consequently, the explanations and potential user corrections are

processed simultaneously with the annotated labels. The necessary modules of this interactive learning loop (Fig. 1) are each described in detail below.

Selection (SELECT). SELECT describes how samples X are selected from a set of non-annotated examples N. These examples are used for the model to perform a predictive task (for example, predict a class label y), with which the user, in turn, has to interact. The selection can be carried out in different ways: manually, randomly or with a specific strategy. One strategy in this regard is to find influential examples: for example, via a model's certainty in a prediction. This can also enable selecting only a subset of examples to apply XIL on. Hence, SELECT also describes how many examples need to be selected to revise a model through explanatory interactions.

Explaining (EXPLAIN). In comparison to active learning, XIL approaches consider standard input–output pairs (for example, $(X, \hat{y})$), insufficient to (1) understand the underlying decision process of a model and (2) provide necessary feedback solely on the predicted labels, denoted as $\bar{y}$. Such feedback, $\bar{y}$, can only correct the model if the model's initial prediction, $\hat{y}$, is incorrect, that is, wrong answer. Due to, for example, shortcut learning[4], deeper insights into a model are required. Hence, in XIL, the model also provides explanations that help the user inspect the reasoning behind a prediction. This, in turn, enables a user to check if the decision is based on right or wrong reasons. Therefore, EXPLAIN is an essential element of an XIL method to revise a model.

In our proposed typology, the learner f (for example, a convolutional neural network, CNN) predicts $\hat{y}$ for an input X. Additionally, the learner explains its prediction to the teacher (for example, user) via an explainer (for example, local interpretable model-agnostic explanations, LIME) and provides an explanation $\hat{E}$. In this way, EXPLAIN depicts how the model provides insights into its reasoning process to the teacher.

There are various ways to provide an explanation. Common explanation methods in XIL works provide attribution maps that highlight important features in the input space, such as input gradients (IG[20]), gradient-weighted class activation maps (Grad-CAM[21]) and LIME[22].

EXPLAIN also describes the capability of an XIL method to facilitate the use of various explainer methods, that is, whether an XIL method depends on a specific explainer method. Whereas some XIL methods can handle arbitrary explainer methods (for example, CounterExamples, CE), it is the defining component for other XIL methods and thus constrains other components of the method as well (for example, feedback types).

Analogous to the view of the explainers, the model flexibility describes the capability of an XIL method to facilitate the use of different model types for EXPLAIN. Depending on the used model, only specific XAI methods can be applied: for example, whereas LIME can be applied to any ML model, IG can only be applied to differentiable ones (for example neural networks), and Grad-CAM only to CNNs. In turn, this means that an XIL method can be model specific or model agnostic. However, the model specificity is linked to the explainer specificity, as an explainer may be only available for certain model types.

Obtain feedback (OBTAIN). Not only does the model have to explain its decision, but also the users have to provide explanatory feedback to the model. This feedback has to be processed in such a way that the model can cope with it. As a result, the model can generate corrections based on user feedback to revise the model. The correction $\bar{C}$ depends on the specific XIL method and model type. Specifically, the user's feedback $\bar{C}$, with respect to the explanation $\hat{E}$, has to potentially be converted to an input space that the model can process. For instance, in the case of counterexamples, the user feedback $\bar{E}$ is on the same level as the explanation: for example, an attribution map. However, correction $\bar{C}$ depicts one or multiple counterexamples, such that $\bar{E}$ must be converted.

In our set-up, the teacher gives feedback based on the model's input X, prediction $\hat{y}$ and explanation $\hat{E}$. Specifically, within OBTAIN, the teacher produces a corresponding explanation, $\bar{E}$, which, however, is transformed into a feedback representation, $\bar{C}$, that corresponds to a representation that can be fed back to the learner. This enables the teacher to observe whether the learner's prediction is right or wrong, but more importantly also to check if the prediction is based on the right or wrong reasoning.

Moreover, OBTAIN determines which feedback types an XIL method can handle. The standard way to provide feedback, partly restricted by using attribution maps in XAI, is to highlight important (right) and/or unimportant (wrong) features in the input. However, other types of feedback are also possible, for example in the form of semantic description: for example, 'Never base the decision on the shape of object X'[23].

Model revision (REVISE). Once the corrections are obtained, they must be redirected back into the model's learning process. Depending on the feedback type and the user's knowledge about what is right or wrong, there are two aspects to consider to revise a model.

The first aspect is how to reinforce user feedback. As indicated in OBTAIN, the REVISE strategy depends on the feedback obtained from the user. On the one hand, the user can penalize wrong explanations, that is, remove confounding factors but not necessarily guide the model towards the right reason. On the other hand, the user can reward the right explanations. Intuitively, it is harder to know the right reason than the wrong reason (for example, on average, the subspace of relevant image regions is much smaller than the space of irrelevant ones). Additionally, rewarding does not necessarily ensure avoiding a confounder's influence. In general, there therefore seems to be an imbalance between knowing what is right and wrong, which needs to be considered.

The second aspect is how to update the model, that is, incorporate the feedback. One common approach is to augment the loss function and backpropagate the feedback information through the loss objective. The other is to augment the dataset with (counter)examples and remove the confounder influence through a diminished presence in the training data.

After the teacher gives feedback to the learner, the corrections are fed back to the learner to revise it. To do so, set A is extended by the processed user annotations, that is, the prediction $\bar{y}$ and the correction $\bar{C}$ for the respective input X. The optimization objective can now incorporate the user feedback to extend the purely data-driven approach and thereby revise (fit) the model f. Finally, N is updated, that is, the annotated instances X are removed from N.

No free lunch in XIL

We hypothesize that there is no single best XIL method. Changing a module has costs such that a modification may improve the performance in one criterion but at the expense of another. Hence, we investigate the different modules with various experiments to verify our hypothesis. Moreover, we showcase our typology and the corresponding evaluation criteria by benchmarking existing XIL methods and their modules. By providing a comprehensive evaluation of these methods, we also reveal some undiscovered limitations to encourage future research.

To this end, we investigate the following questions. (Q1) How well do the existing methods revise a model? (Q2) Are they robust to feedback quality variations? (Q3) Does the revision still work with a reduced number of interactions? (Q4) Can XIL revise a strongly confounded model?

XIL methods, measures and benchmarks. We focus our evaluations on computer vision datasets, where confounders are well known and an active area of research[24]. In the relevant datasets, a confounder is a visual region in an image (for example coloured corner) correlating with the image class but is not a causal factor for determining the true class of the image. The confounder fools the model and constitutes a shortcut learning rule[4]. In the standard set-up, we train an ML model on a confounded training set and run tests on the non-confounded test set. Our goal is to guide the model to ignore the confounder. To account for different facets of XIL, we chose two benchmark datasets, Decoy(F)MNIST, and one scientific dataset, ISIC19. For these datasets, a confounder is visual (in the sense that they are spatially separable) to provide a controlled environment for evaluation.

In the following, we evaluate the XIL methods (1) RRR (right for the right reasons[15]), (2) CDEP (contextual decomposition explanation penalization[17]), (3) HINT (Human Importance-Aware Network Tuning[18]) and (4) CE[12] and analyse the influence of different explainers on the same method, namely (5) the influence function (IF) referred to as RBR (right for better reasons[16]) and (6) Grad-CAM, in the following called RRR-G (right for the right reasons Grad-CAM[13]). We summarize all investigated methods with their respective implementation of each component in Table 1. We set up measures and benchmarks in Methods to provide detailed insights into the versatile facets of an XIL method. Besides standard measures such as accuracy, we provide a new measure, WR, to investigate a model's focus on the wrong reason(s). Furthermore, we provide extensive benchmarks such as the feedback robustness, the interaction efficiency and a 'switch XIL on' experiment. These benchmarks help examine various essential aspects of an XIL method beyond simple accuracy scores. More details of these and the experimental protocol can be found in Methods.

(Q1) Accuracy revision. To investigate the general ability of an XIL method to revise a model (REVISE), we evaluate the accuracy score on each test set (Table 2) of the datasets DecoyMNIST, DecoyFMNIST and ISIC19. To give an impression of the confounder impact in each dataset, we provide a baseline by evaluating each model on the dataset without decoy squares. This is not available for ISIC19 as the confounders are not artificially added. The vanilla model represents the performance of a model without revision via XIL. The confounder causes the models to be fooled, causing low accuracy scores compared with the baseline without the decoy.

In contrast, training via the examined XIL methods generally helps a model overcome the confounder, as the baseline test accuracy is recovered. RBR performs best on DecoyMNIST and RRR on DecoyFMNIST. For DecoyFMNIST, HINT achieves a low accuracy score on par with the fooled vanilla model, indicating that here it cannot correct the Clever Hans behaviour. We assume that its reward strategy does not suffice to overcome the confounder and, in turn, for XIL to function properly. For the ISIC19 dataset, no XIL method helps a model improve the accuracy on the test set. Therefore, we cannot answer (Q1) affirmatively for ISIC19 purely based on the accuracy, thus motivating the need for further measures beyond accuracy.

However, summarized, our experiments answer (Q1), that is the XIL methods have the general ability to revise a model but may have difficulties with increasing data complexity.

(Q1) Wrong reason revision. For the ability to revise wrong reasons, we conduct quantitative (Table 3) and qualitative (Fig. 2) experiments to inspect EXPLAIN.

On one hand, we have the quantitative WR score. It measures the activation in the confounder area and hence automates the visual inspection of explanations. The vanilla model (without XIL) achieves high WR scores, that is, high activation in the confounder region. Again, the XIL methods help a model lower the WR score, reducing the confounder impact. Table 3 further shows that the XIL methods overfit on the internally used explainer in terms of reducing its attention to the confounding region (cf. RRR with IG explanations or RRR-G with Grad-CAM explanations). This is expected, as an XIL method exactly optimizes for the explanation method used. Interestingly, an XIL method also reduces the WR score for other explainers that are not internally used (cf. RRR with Grad-CAM explanations or RRR-G with IG explanations). Consequently, XIL's impact is beyond its internally used explainer and not restricted to it. However, LIME attribution maps are always highly activated, albeit reduced, as it is never internally used as an explainer.

Furthermore, we can see that CDEP and HINT do not remarkably reduce the WR score when compared with the baseline. As HINT works with rewarding instead of penalizing and is thus not explicitly trained to avoid confounders, we do not necessarily expect it to score low WR values. CDEP also does not achieve low WR values and does not overcome the confounder, despite using a penalty. We previously

Table 1 | Overview of the XIL method set-up in our experiments: RRR[15], RRR-G[13], RBR[16], CDEP[17], HINT[18] and CE[12]

Module	RRR	RRR-G	RBR	CDEP	HINT	CE
SELECT	random					
EXPLAIN	IG	Grad-CAM	IF	CD	Grad-CAM	LIME
OBTAIN	attribution mask					
REVISE	penalty	penalty	penalty	penalty	reward	dataset

Table 2 | Mean accuracy scores as percentages; best values are bold; cross-validated on five runs with s.d

	DecoyMNIST		DecoyFMNIST		ISIC19	
XIL	Train	Test	Train	Test	Train	Test
W/o decoy	99.8±0.1	98.8±0.1	98.7±0.3	89.1±0.5	—	—
Vanilla	99.9±0.0	78.9±1.1	99.5±0.2	58.3±2.5	100±0.0	88.4±0.5
RRR	99.9±0.1	98.8±0.1	98.7±0.3	**89.4**±0.4	**100**±0.0	88.1±0.4
RRR-G	99.7±0.2	97.4±0.7	90.2±1.6	78.6±4.0	**100**±0.0	**88.4**±0.5
RBR	**100**±0.0	**99.1**±0.1	96.6±2.3	87.6±0.8	92.6±5.3	80.3±5.6
CDEP	99.3±0.0	97.1±0.7	89.8±2.7	76.7±3.5	**100**±0.0	87.9±0.5
HINT	97.6±0.3	96.6±0.4	99.0±0.9	58.2±2.3	**100**±0.0	87.7±0.5
CE	99.9±0.0	98.9±0.2	**99.1**±0.2	87.7±0.8	**100**±0.0	87.5±0.5

The first row shows performance on a dataset without decoy squares (not available for ISIC19). The next row shows that the vanilla model (no XIL) is fooled, indicated by low test accuracy. Except for HINT on FMNIST, all methods recover test accuracy. On ISIC19, no accuracy improvement can be observed; higher is better.

https://doi.org/10.1038/s42256-023-00612-w

Table 3 | Mean WR scores as percentages; best values are bold; cross-validated on five runs with s.d

	DecoyMNIST			DecoyFMNIST			ISIC19		
XIL	**IG**	**Grad-CAM**	**LIME**	**IG**	**Grad-CAM**	**LIME**	**IG**	**Grad-CAM**	**LIME**
Vanilla	23.1±3.8	38.7±4.6	59.8±2.0	25.0±1.9	34.8±1.4	57.6±0.8	33.2±0.2	35.2±0.8	63.6±0.7
RRR	**0.0**±0.0	13.3±2.0	**32.1**±0.4	**0.0**±0.0	24.2±4.1	**27.4**±0.7	16.6±8.7	27.4±4.3	58.9±1.4
RRR-G	11.9±2.1	**1.5**±0.8	33.3±2.8	2.1±0.4	**4.6**±0.9	38.1±4.5	**11.9**±0.9	**0.9**±0.1	**34.7**±0.8
RBR	2.0±1.3	15.2±3.8	37.7±3.0	5.97±1.4	16.0±4.8	34.9±1.4	17.2±1.4	28.5±22.7	58.0±0.6
CDEP	15.0±1.5	27.8±3.8	37.9±3.7	15.9±4.5	39.1±1.7	40.2±6.5	25.5±0.2	5.4±0.2	67.4±0.0
HINT	11.9±3.1	46.8±1.1	53.8±2.0	29.4±3.3	27.8±2.9	51.4±3.5	31.1±0.1	22.0±0.2	60.9±0.0
CE	7.3±1.4	14.7±2.9	36.9±0.6	8.1±0.4	24.4±0.9	31.1±0.6	32.7±0.0	36.6±0.1	61.5±0.9

XIL reduces WR scores for all methods on all datasets, even for ISIC19; lower is better.

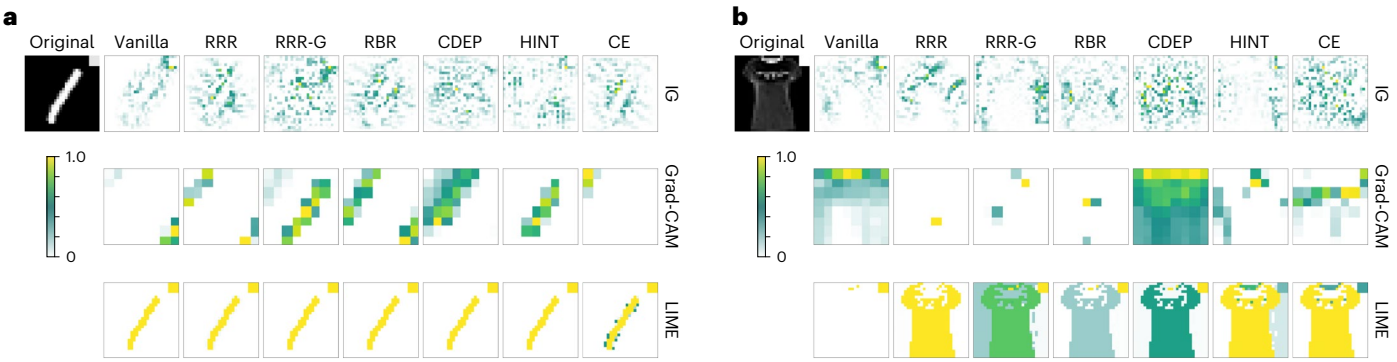

Fig. 2 | Qualitative inspection of explanations. a,**b**, The first column in each panel shows the original image, the second column shows the vanilla model (no XIL) attribution maps and the remaining columns show the attribution maps of a model with each XIL method. Each row represents an explanation method to visualize the model prediction. The colour bar indicates the activation of attribution maps (yellow (1) represents maximum activation and white (0) minimum activation). Results for DecoyMNIST (**a**) and DecoyFMNIST (**b**).

found that XIL could not overcome the test performance of the fooled vanilla model on the ISIC19 dataset. However, the WR score surprisingly indicates a clear reduction. This indicates that, although XIL might not revise a model in terms of accuracy, it can indeed improve the explanations (lower the WR score), proving its function and usefulness. Notably, these findings showcase the importance of additional quantitative measures, such as WR, for evaluating XIL methods.

Apart from this, we manually inspected 200 randomly generated attribution maps for each method–explainer combination. We exemplify the findings for a representative example on each benchmark dataset. Figure 2 shows explanation attribution maps for Decoy(F)MNIST in panel **b**. A high activation in or around the top right corner indicates Clever Hans behaviour of a model with respect to the confounder. For the vanilla model, each explanation method highlights the confounder region, except for the Grad-CAM explanation on MNIST (Discussion). The top row shows activation attribution maps generated with the IG explainer. Here, we can see the previously found overfitting once again. For example, the RRR-revised model shows no activation in the confounding top right corner while RRR-G still has high activation around the corner. Consequently, XIL functions reliably only on the internally used explainer. The qualitative findings here confirm the quantitative findings of the WR score and demonstrate that it is a suitable method to evaluate the performance of XIL methods.

Moreover, in our evaluation, we also find that penalizing wrong reasons does not enforce predictions based on the right reasons (cf. RRR-G with IG attribution maps on MNIST). Second, attribution maps generated via Grad-CAM require upsampling and hence prevent a clear and precise interpretation. Although the right reason is sometimes highlighted, there remains some uncertainty in its precision. Third, the explanation methods provided visually contradict each other. The RRR column, for instance, indicates that XIL, and XAI in general, must be handled with care: the performance values may show overcoming a confounder while the visual explanations (attribution maps) indicate otherwise.

Overall, our evaluation gives more insight into EXPLAIN of XIL and extends the previous findings for (Q1). Although it may not become entirely apparent that the considered XIL methods remove all Clever Hans behaviour when only considering accuracy, we observed that the methods and, in this way, XIL in general do in fact improve a model's explanations and can therefore effectively be used to revise a model.

We note here that we found possibly additional confounding factors in the ISIC19 dataset and hence focus further evaluations on the remaining two datasets.

(Q2) Robustness to feedback quality variations. As previous research focused only on providing correct (ground-truth) feedback, we additionally provide insights into the feedback quality and the robustness of an XIL method towards quality changes. In this experiment, the objective is to gather more knowledge about OBTAIN.

Table 4 compares the impact of different feedback types with a fooled vanilla model. The values for correct feedback are taken from Table 2. Correct feedback demonstrates how XIL improves the accuracy, that is, removes the confounder impact, compared with the vanilla model. Moreover, we can clearly see that incomplete and correct feedback are nearly on par for all methods in improving the test accuracy. This emphasizes XIL's robustness towards user feedback of varying quality and suggests real-world usability of XIL, considering that human user feedback is prone to errors. Note that the performance of CE for

https://doi.org/10.1038/s42256-023-00612-w

Table 4 | Feedback robustness evaluation on Decoy(F)MNIST for arbitrary and incomplete, compared with correct, feedback masks the mean difference in test accuracy (%) compared with the confounded vanilla model is given

	DecoyMNIST						DecoyFMNIST					
Feedback	RRR	RRR-G	RBR	CDEP	HINT	CE	RRR	RRR-G	RBR	CDEP	HINT	CE
Arbitrary —	+3.3	−4.2	−22.1	+17.8	+4.1	+0.3	+1.4	+7.9	−37.3	+12.2	−3.2	−1.2
Incomplete ↑	+19.6	+9.5	+17.2	+17.9	+17.7	+6.7	+24.2	+12.4	+16	+16.8	+21	+3.4
Correct ↑	+19.9	+18.5	+20.2	+18.2	+17.7	+20	+31.1	+20.3	+29.3	+18.4	−0.1	+29.4

For arbitrary feedback unchanged is better, and for incomplete and correct feedback higher is better. Incomplete feedback is on par with correct feedback. The values are cross-validated on five runs (cf. Supplementary Table 5 for s.d. values).

incomplete feedback is worse due to the strategy of augmenting the dataset. While all confounded images remain in the data, the images added to revise the model also still contain part of the confounder. This way, the confounder impact is still quite high and thus not as easily removed by adding images where only part of the confounding square is removed. However, our results indicate that this still suffices to achieve limited revision.

In contrast, for the case of arbitrary feedback, robustness expresses that a method does not remarkably change performance. However, we can see a remarkable increase (decrease) for CDEP (RBR), especially compared with the correct feedback improvement. This suggests that CDEP improves performance no matter what the feedback quality. Consequently, we presume that CDEP does not pass the sanity check, leaving some concerns about its reliability. If it is irrelevant what the user feedback looks like to correct the model, the rationale behind the XIL method is questionable and its usage worrisome for users. For RBR, we presume that arbitrary feedback leads to a collapse of the model's learning process, that is, random guessing, revealing a lack of robustness. A method sensitive to wrong feedback—humans are not perfect—can be assessed as worse than a robust method.

All in all, however, the considered XIL methods prove general robustness for different feedback quality types, thus answering (Q2) affirmatively and providing evidence for XIL's effectiveness in more practical use cases.

(Q3) Interaction efficiency. Obtaining precise and correct feedback annotations is costly and time consuming, making interaction efficiency a crucial property for XIL methods. Therefore, we examine how many explanatory interactions suffice to overcome a known confounder. A method that utilizes annotations more efficiently, that is, requires fewer interactions to revise a model, is preferable. In the previous experiments, every training image was accompanied by its corresponding feedback mask to correct the confounder. In contrast, now we randomly sample a subset of k annotations before training and evaluate each model with different-sized feedback sets, that is, number of explanatory interactions. By doing so, we target SELECT as we investigate how the selection affects the model revision.

Figure 3a shows increasing test accuracy for an increasing number of available feedback masks for all XIL methods, that is, the more feedback available, the better. Moreover, the figure shows that only a tiny fraction of feedback masks is required to revise a model properly. Although there is a remarkable difference between the XIL methods' efficiencies, our obtained results illustrate that XIL utilizes feedback efficiently and can already deal with a few feedback annotations. Note that the methods achieve different test accuracies with all available feedback, such that they do not all converge at the same level; cf. Table 2 for test accuracy with full feedback set. Interestingly, RRR, for example, needs only a few dozen interactions to overcome the confounder, while CE requires considerably more interactions. Summarized, the examined XIL methods can efficiently utilize user feedback, solving (Q3).

(Q4) Revising a strongly corrupted model. To further evaluate the real-world usability of XIL, we conduct a switch XIL on experiment, where we integrate an XIL method (switch XIL on) after a model has solely been trained in a baseline setting (for example standard classification) and shows strong Clever Hans behaviour. Figure 3b shows the test performance of a model during training. First, the model is fooled by the decoy squares. After 50 epochs, the XIL augmentation is switched on (that is, either the loss or dataset is augmented with XIL). As we can see, all methods, except CDEP and RBR, can recover the test accuracy and overcome the confounder. RRR shows a striking curve with the explanation loss sharply increasing, and hence the accuracy drops before it sharply increases again. Most likely, it requires more hyperparameter tuning to avoid this leap. For RBR, we assume the same, as it is difficult to tune the loss accordingly.

Also, (Q4) is thus answered affirmatively by this experiment as it overall shows that XIL can 'cure' already confounded models, which is an important property for real-world applications.

Discussion

The previous sections demonstrated that modifying XIL modules is no free lunch in the sense that modifying one module does not guarantee improvements in all criteria. In the following, we wish to discuss some additional points.

As pointed out initially, it is often easier to state a wrong reason than a right reason[13]. However, penalizing wrong reasons may not be enough to revise a model to be right for all the right reasons. Avoiding one wrong reason but using another wrong one instead is not desirable. The provided attribution maps for ISIC19 (cf. Fig. 4) illustrate this trade-off. As we can observe from the attribution maps, the reward strategy (HINT) visually outperforms the penalty strategy. The penalty method, exemplified here via RRR, does to a certain degree point towards the right reason, but not as reliably as rewarding via HINT. In general, however, a reward strategy cannot guarantee that confounders are avoided.

Throughout our work, we encountered ambiguities between different explainer methods. When an XIL method is applied and a sample is visualized with a different explainer method than was used for optimization, we find contradicting attribution maps (cf. RRR columns in Fig. 2). In fact, the analysis of attribution maps shows remarkable differences between IG, Grad-CAM and LIME. In some cases, we can even observe opposing explanations. Moreover, the Grad-CAM explanation for the vanilla model in Fig. 2a does not show a confounder activation although the scores in Table 3 clearly pinpoint a shortcut behaviour of the model. This consequently raises concerns about how reliable and faithful the explainer methods are. At this point, we note that to investigate XIL we make a general assumption about the correctness of the explanation methods, which is still an open topic in the field[25,26]. Although it is the explicit goal of XIL to improve explanations, this can only work if an explainer method does not inherently fail at producing meaningful and coherent explanations. In that case, the overall objective of increasing user trust is already undermined before XIL enters the game. One of the main challenges of XIL is real-world application.

https://doi.org/10.1038/s42256-023-00612-w

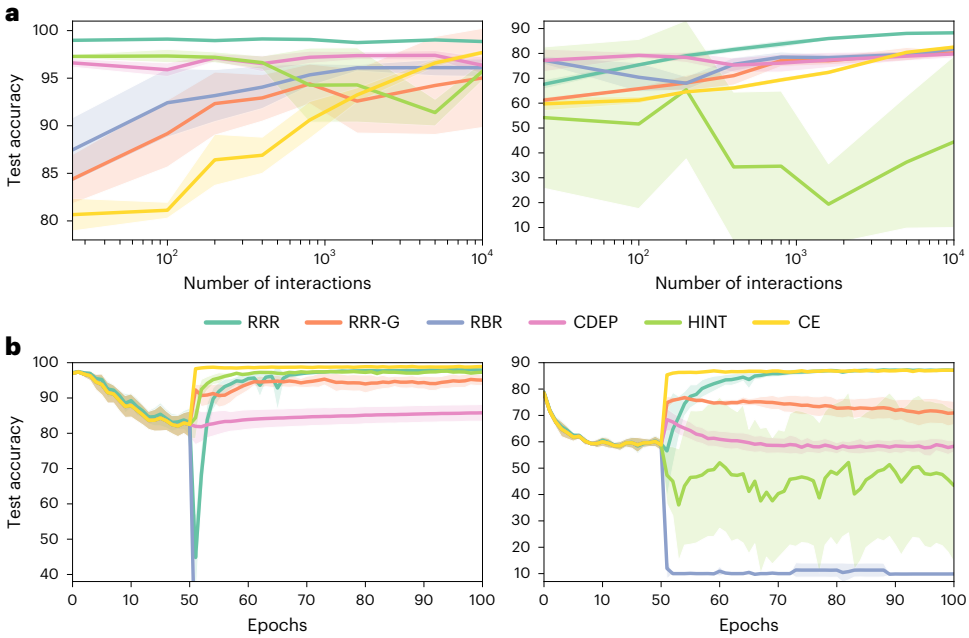

Fig. 3 | Evaluation of interaction efficiency and performance in unconfounding a pretrained model. a,**b**, Each task is evaluated on the DecoyMNIST (left) and DecoyFMNIST (right) datasets. **a**, Test accuracy (%) with different numbers of feedback interactions used. The greater the number of interactions, the better the performance. However, a smaller number of interactions already suffices. **b**, Test accuracy (%) over time after XIL is applied to an already fooled model. All methods, except CDEP and RBR, can recover the test performance and overcome the confounder. Each value represents the mean performance cross-validated on five runs; the given confidence intervals represent the s.d. (left in **b**) the blue curve (RBR) drops to random performance.

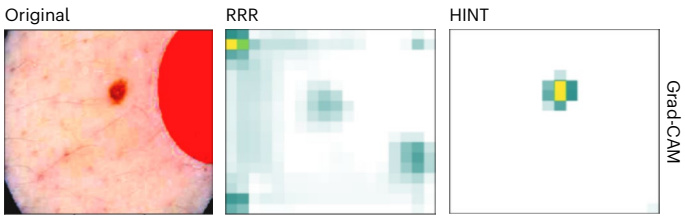

Fig. 4 | An RRR-revised model and a HINT-revised model generate explanations for an ISIC19 image with confounder. Left: ISIC19 image with confounder (red patch). Middle: RRR-revised model. Right: HINT-revised model. The explanations are visualized with Grad-CAM. RRR helps discover unknown confounders (dark corners), and HINT reveals the potential of the reward strategy.

Revising a model must be easy for an average ML practitioner or any human user. If the resource demand is too high, the methods are difficult to use. This is specifically a problem for state-of-the-art, large-scale pretrained models. One example is RBR, which uses IFs, that is, second-order derivatives, to correct Clever Hans behaviour. In our evaluations, we found that IFs induce a huge resource demand, making XIL slower and more challenging to optimize—loss clipping was necessary to avoid exploding gradients.

In terms of architecture choice and design, we also encountered several obstacles. Our typology description has already pointed out that not every XIL method is applicable to every model or explainer method: for example, Grad-CAM-based XIL methods can only be applied to CNNs. We argue that a flexible XIL method is preferable such that various models and explainer methods can be applied.

From our experimental evaluations considering the number of required interactions, we observed that CE, with the dataset augmentation strategy, requires the largest amount of user feedback. Especially for large-scale models, the number of interactions required can be a limiting factor. In practical use cases, often only a limiting amount of explanatory feedback is available. Another aspect here is trustworthiness, in that a user might not trust a model as much if the feedback they have provided is not directly incorporated by the model and should suffice to revise a model. Furthermore, we noticed that CE is less robust to incomplete feedback, possibly compromising this approach alone for real-world application. Hence, combining CE with a loss-based XIL approach could be advantageous.

Finally, a very noteworthy potential of XIL could be observed in the qualitative evaluations of ISIC19 attribution maps. In fact, by applying XIL on one confounder, we could identify further unknown confounders (shortcuts) to the user: in this case, the dark corners found in the images (cf. Fig. 4 (middle)). These findings further demonstrate the importance of a strong level of human–machine interactions via explanations. Particularly in such a setting, each can learn from the other in a stronger bidirectional, discourse-like manner in more than just the unidirectional way of communication provided by XAI alone. To this end, we refer to the theory of embodied intelligence, in which interaction and manipulation of the environment allow for information enrichment to obtain intelligent systems[27].

Conclusions

In summary, this work presents a comprehensive study in the rising field of XIL. Specifically, we have proposed the XIL typology to unify terminology and categorize XIL methods concisely. On the basis of this typology we have introduced benchmarking tasks, each targeting specific aspects of the typology, for properly evaluating XIL methods beyond common accuracy measures. These cover the performance in model revision, robustness under changing feedback quality, interaction efficiency and real-world applicability. In addition, we introduced a WR measure for our experiments to quantify an average confounder activation in a model's explanations. Finally, we showcased the typology and benchmark criteria in an empirical comparison of six currently proposed XIL methods.

Our typology and evaluations showed that XIL methods allow us to revise a model in terms of not only accuracy but also explanations. However, we also observed overfitting to the individual explainer method

being used. Avenues for future research are a mixture of explainers that may account for uncertainty on the right reasons. Moreover, we should combine feedback on what the right explanation is with feedback on what it is not, rather than focusing on only one of these two feedback semantics. The weighting of each feedback on the basis of the knowledge and feedback certainty of the user goes one step further. Moreover, of course, application to large-scale pretrained models is an exciting avenue for future work. Most importantly, existing XIL approaches just follow a linear SELECT, EXPLAIN, OBTAIN and REVISE. As in daily human-to-human communication, machines should also follow more flexible policies such as an EXPLAIN & OBTAIN subloop, pushing for what might be called explanatory cooperative AI[28].

Overall, in this work, apart from a typology we have also introduced a set of measures and benchmarking tasks for differentiating and evaluating current and future XIL methods. This toolbox is by no means conclusive and should act as a starting point for a more standardized means of evaluation. Additional or improved measures and tasks should be investigated in further research: for example, the benefits of mutual information.

Methods

In this section, we present existing XIL methods and the measures and benchmarks we use and at the same time propose to evaluate.

XIL methods

The fundamental task of XIL is to integrate the user's feedback on the model's explanations to revise its learning process. To tackle this core task, several XIL methods have been recently proposed. Below we describe these methods in detail, dividing them on the basis of two revision strategies: revising via (1) a loss term or (2) dataset augmentation. Both strategies rely on local explanations.

Loss augmentation. Strategy (1) can be summarized as optimizing equation (1)[17], where X denotes the input, y ground-truth labels and f a model parameterized by θ. We optimize

$$\min_\theta \underbrace{L_{\text{pred}}(f_\theta(X), y)}_{\text{Prediction error}} + \lambda \underbrace{L_{\text{exp}}(\text{expl}_\theta(X), \text{expl}_X)}_{\text{Explanation error}}, \tag{1}$$

where L_{pred} is a standard prediction loss, for example cross-entropy, guiding the model to predict the right answers, whereas L_{exp} ensures the right reasons, that is right explanations, scaled by the regularization rate λ.

Right for the right reasons (RRR) In the work of Ross et al.[15], the objective is to train a differentiable model to be right for the right reasons by explicitly penalizing wrong reasons, that is, irrelevant components in the explanation. This means that REVISE enforces a penalty strategy. To this end, this approach generates gradient-based explanations $\text{expl}_\theta(X)$ and restricts them by constraining gradients of irrelevant parts of the input. For a model $f(X|\theta) = \hat{y} \in \mathbb{R}^{N\times K}$ and inputs $X \in \mathbb{R}^{N\times D}$ we obtain

$$L_{\text{exp}} = \sum_{n=1}^{N} \left(M_n \text{expl}_\theta(X_n)\right)^2, \tag{2}$$

where N is the number of observations, K is the number of classes and D is the dimension of the input. With this loss term, the user's explanation feedback $M_n = \text{expl}_X$, indicating which input regions are irrelevant, is propagated back to the model in the optimization phase. The loss prevents the model from focusing on the masked region by penalizing large values in this region. According to the authors, L_{pred} and L_{exp} should have the same order of magnitude when setting a suitable λ in equation (1).

Ross et al.[15] implement EXPLAIN with IG by generating explanations based on first-order derivatives, that is, $\text{expl}_\theta(X) = \text{IG}(X)$. However, RRR's EXPLAIN is not limited to this explainer. Schramowski et al.[13] propose RRR-G, generating explanations via $\text{expl}_\theta(X) = \text{Grad-CAM}(X)$, and Shao et al.[16] propose RBR with second-order derivatives, that is, $\text{expl}_\theta(X) = \text{IF}(X)$. We describe further mathematical details in Supplementary Section D. To penalize wrong reasons, OBTAIN in this case expects feedback in the following form. A user annotation mask is given as $\text{expl}_X = M \in \{0,1\}^{N\times D}$, with ones indicating wrong reasons. For example, in this work, we consider confounding pixels as wrong reasons.

Contextual decomposition explanation penalization (CDEP) Compared with the others, CDEP[17] uses a different explainer method, CD; that is, its EXPLAIN module is restricted to this explainer method, $\text{expl}_\theta(X) = \text{CD}(X)$. The CD algorithm measures the layer-wise attribution of a marked feature, here image region, to the output. It decomposes the influence on the prediction between the marked image region and the remaining image. This enables only focusing on the influence of the marked image region and, in this case, penalizing it. Hence, REVISE is implemented again with the penalty strategy. The user mask M penalizes the model explanation via

$$L_{\text{exp}} = \sum_{n=1}^{N} \left\| \text{expl}_\theta(X_n) - M_n \right\|_1. \tag{3}$$

Human importance-aware network tuning (HINT) In contrast to previous methods, HINT[18] explicitly teaches a model to focus on right reasons instead of not focusing on wrong reasons. In other words, HINT rewards activation in regions on which to base the prediction, whereas the previous methods penalize activation in regions on which not to base the prediction. Thus, REVISE is carried out with the reward strategy. EXPLAIN can take any gradient-based explainer, but the authors implemented it with Grad-CAM, that is $\text{expl}_\theta(X) = \text{Grad-CAM}(X)$. Finally, a distance, for example via mean squared error, is computed between the network importance score, that is, generated explanation, and the user annotation mask, resulting in

$$L_{\text{exp}} = \frac{1}{N} \sum_{n=1}^{N} \left(\text{expl}_\theta(X_n) - M_n\right)^2. \tag{4}$$

Importantly, OBTAIN differs from previous methods in that ones in M mark right reasons, not wrong reasons. We define relevant pixels (components) as the right reasons for our evaluations.

Dataset augmentation. In contrast to the XIL methods, which add a loss term to revise the model, that is to implement REVISE, further XIL methods exist which augment the training dataset by adding new (counter)examples to the training data[12]. Where the previous approaches directly influence the model's internal representations, this approach indirectly revises a model by forcing it to generalize to additional training examples, specifically tailored to remove wrong features of the input space. This augmentation can, for example, help prevent a model from focusing on confounding shortcuts.

CounterExamples (CE) Teso and Kersting[12] introduce CE, a method where users can mark the confounder, that is, wrong reason, region in an image from the training data and add a corrected image, that is, one in which an identified confounder is removed, to the training data.

In comparison with strategy (1), this strategy is model and explainer agnostic, that is, EXPLAIN can be implemented with any explainer method as user feedback is not processed directly via the model's explanations. Specifically, OBTAIN takes user annotation masks that mark the components in the explanation that are incorrectly considered relevant. In this case, the explanation corrections are defined by $C = \{j: |w_j| > 0 \wedge j\text{th component marked by user as irrelevant}\}$, where w_j denotes the jth weight component in the attribution map. These explanation corrections are transformed into counterexamples to make the feedback applicable to the model. A counterexample is defined as $j \in C : \{(\bar{X}, \bar{y})\}$, where $\bar{y}_i$ is the corrected label, if needed, and $\bar{X}_i$ is the identical input, except the

previously marked component. This component is either (1) randomized, (2) changed to an alternative value or (3) substituted with the value of the jth component appearing in other training examples of the same class. The counterexamples are added to the training dataset. Moreover, it is also possible to provide multiple counterexamples per correction: for example, different strategies. In our case, where the input is an image, the user's explanation correction is a binary mask, and a counterexample is an original image with the marked pixels corrected.

Instead of using noise to augment an example, Lang et al.[29] present an attractive alternative that generates new realistic examples from a style space learned with a GAN-based approach.

Evaluating XIL is more than just accuracy

Although a variety of works on XIL exist, there remains a research gap due to the lack of an in-depth comparison between these. Moreover, XIL methods are usually only compared, if at all, in terms of accuracy on benchmark confounded datasets. This essentially only measures if an XIL method successfully helps overcome the confounder in terms of predictive power. However, the goal of XIL goes beyond overcoming confounders and also includes improving explanations overall, for example outside a confounding setting. Hence, a profound examination that focuses on different aspects of the typology is crucial for a sound analysis of current and future research and for filling this research gap. We therefore extend our typology by proposing additional measures and benchmarking tasks for a thorough evaluation of an XIL method, and clarify these in the following sections.

Measures for benchmarking. In the following, we present existing and introduce new quantitative and qualitative approaches to evaluate XIL methods.

Accuracy Many previous works on XIL revert to measuring prediction accuracy as a standard measure to evaluate a method's performance. This mainly works by using a confounded dataset in which the predictive accuracy on a non-confounded test set serves as a proxy for 'right reasons'. However, this measure can only be used to evaluate XIL on datasets with known confounders and test sets that do not contain the confounding factor. Otherwise, unknown confounders may still fool the model and prevent an accurate evaluation of an XIL method. This is particularly important as XIL aims not only to improve the predictive power, but also to improve the quality of the model's explanations regarding the preferences and knowledge of the human user. We note that in all further mentions of 'test accuracy' we are considering a model's performance based on the dataset classification rate on an unseen test set.

Qualitative explanation analysis Another approach to evaluating the effectiveness of XIL methods is to qualitatively inspect a model's explanations (for example, attribution maps) before and after revisions. Next to the previously mentioned test accuracy, this approach to quality assessment is another popular measure on which many previous works focus their evaluations. Some recent techniques for (semi-)automatic explanation analysis exist: for example, for detecting Clever Hans strategies. For example, Spectral Relevance Analysis inspects and clusters similar explanations[11,30].

Wrong reason measure Besides standard measures such as accuracy, we therefore propose an intuitive measure, termed the wrong reason measure (WR), to measure how wrong a model's explanation for a specific prediction is, given ground-truth (user-defined) wrong reasons. In contrast to the qualitative evaluation (manual, visual inspection) of attribution maps, our WR measure provides a quantitative complement.

In detail, given an input sample X, for example an image, a model f with parameters θ, an explainer expl and ground-truth annotation mask M, we quantify WR as

$$\mathrm{WR}(X, M) = \frac{\mathrm{sum}(b_\alpha(\mathrm{norm}^+(\mathrm{expl}_\theta(X))) \circ M)}{\mathrm{sum}(M)}, \quad (5)$$

where $\circ$ is the Hadamard product, and norm^+ normalizes the attribution values of the explanation to $[0, 1]$, while only taking positive values into account by setting all negative values to zero. b_α binarizes the explanation ($\mathrm{expl}_{ij} > \alpha \Rightarrow 1$ else 0) and the threshold α can be determined by the pixel-level mean of all explanation attribution maps in the test set beforehand.

Depending on the explainer expl, it might be necessary to scale (down/up) the dimensions of the explanation to match the dimension d of M. In short, the measure calculates to what extent the wrong reason area is activated. WR = 1 translates to 100% activation of the wrong reason region, indicating that the model is fooled by the wrong reason and spuriously uses it as an informative feature. If this behaviour co-occurs with high predictive performance this will imply Clever Hans behaviour and reasoning based on a wrong reason. In contrast, WR = 0 signals that 0% of the wrong reason area is activated. However, it is worth noting that one cannot, in principle, claim that the model's reasoning is based on the right reason from being not wrong.

Comparing the WR scores of a vanilla model with those of an XIL-extended model allows us to estimate the effectiveness of a specific XIL method. As one objective of XIL is to overcome the influence of the wrong reason area, the WR score should at least be smaller than the score for the vanilla model.

New benchmarking tasks. In the following, we introduce further relevant benchmarking tasks for evaluating XIL methods.

Feedback robustness An important aspect of the usability of an XIL approach is its robustness to the completeness of and quality variations within the user feedback. This task is vital, as user feedback in the real world is error prone. To provide a benchmark that is comparable between different datasets and can be efficiently evaluated, we propose to simulate and model robustness via a proxy task for all dataset–model combinations. In the spirit of Doshi-Velez and Kim[31], this task is, therefore, a functionally grounded evaluation, with no human in the loop.

Two compelling cases to examine are cases of arbitrary and incomplete feedback. Arbitrary feedback can also be viewed as a sanity check of an XIL method since it should not change the performance in any direction. In other words, a model should not produce worse or better predictive performance, as the feedback is neither useful nor detrimental. On the other hand, incomplete feedback imitates real-world feedback by providing only partially valuable feedback. For instance, in the case of the DecoyMNIST (for details see Experimental protocol), two scenarios can be modelled as follows.

1. Arbitrary feedback: 5×3 rectangle pixel region in the middle sections of the top or bottom rows of an image, thus neither on relevant digit feature regions nor on any parts of confounder squares, that is $M \neq C$.
2. Incomplete feedback: with subregion S (here top half) of relevant components C. Thus, $M = \mathbf{1}_S : C$.

A feedback mask is again denoted by M and the set of (ir)relevant components by C. In the case of correct user feedback $M = C$.

CE uses manipulated copies of the original images instead of binary masks. There are different CE strategies to manipulate (we chose the CE strategy randomize). The manipulated images are added to the training set. Exemplary feedback masks for this experiment are illustrated in Supplementary Fig. 1, visualizing the feedback types, and further details and examples can be found in Supplementary Section B.1.

Interaction efficiency. In many previous applications and evaluations of XIL methods, every training sample was accompanied by

corresponding explanatory feedback. Unfortunately, feedback, for example in the form of input masks, can be costly to obtain and potentially only available to a limited extent. A very relevant evaluation, particularly for a method's practical usability, is how many feedback interactions are required for a human user to revise a model via a specific XIL method. In other words, we propose to investigate the interaction efficiency of a method as an additional benchmark task.

To simulate a reduced feedback size, we propose to randomly sample a subset of k annotations before training and evaluate each model with the different-sized feedback sets. Different values for k, that is number of explanatory interactions, enable a broad insight into the capability of an XIL method to revise a model efficiently. The effect of the reduced set size is measured with accuracy. Thus, this evaluation task reduces the feedback set size and investigates its impact on the overall effectiveness of an XIL method.

Switch XIL on A further benchmark task that we propose is called switch XIL on and is motivated in two ways. First, it complements previous works, which often only simulated interaction with a model from scratch and not from a strongly confounded model, which would grant more insight into the effectiveness and function of XIL. Second, Algorithm 1 shows that a model is usually fitted to the given data beforehand, and XIL is applied to the confounded model after, for example, Clever Hans behaviour is detected. This contrasts with the other evaluation tasks, where often a model is optimized via an XIL method from scratch. In addition to related work investigating the real-world applicability of XIL with a more real-world dataset[32], we want to propose methods for the same purpose instead. This property is essential, as completely retraining a model can be very costly or even infeasible: for example, for large-scale pretrained models. Hence, it would be very valuable for an XIL method if it can successfully be applied in revising an already corrupted model.

This evaluation task targets correcting a pretrained, strongly corrupted model, that is, a model already strongly biased towards Clever Hans behaviour. To this end, a vanilla model is trained on the confounded training set for several epochs. Subsequently, the XIL loss is switched on (for CE, the training set is augmented).

Experimental protocol

For our experiments, we use two different models: a simple CNN for the benchmark and a VGG16 for the scientific dataset. We use RRR with different explainer methods (IG (RRR), Grad-CAM (RRR-G) and IF (RBR)) to not only compare different XIL methods but also investigate the impact of different explainer methods on the same XIL method. For simplicity, we only investigate one XIL method with different explainers (Table 1). In this work, we optimize our models with Adam[33] and a learning rate of 0.001 for 50 epochs. For the standard experiments, the right reason loss is applied from the beginning. For Decoy(F)MNIST we use the standard train–test split (60,000–10,000), and for the ISIC19 dataset we use an 80–20 split. We set the batch size to 256 for the benchmark datasets and to 16 for ISIC19. Further experimental details can be found in Supplementary Sections A and B.

The DecoyMNIST dataset[15] is a modified version of the MNIST dataset, where the training set introduces decoy squares. Specifically, training images contain 4×4 grey squares in randomly chosen corners, whose shades are functions of their digits. These grey-scale colours are randomized in the test set. The binary feedback masks M mark confounders for the penalty strategy, while the masks mark the digits for the reward strategy.

Fashion-MNIST (FMNIST) is an emendation of MNIST, as it is overused in research and limited in complexity. FMNIST consists of images from ten fashion article classes. The DecoyFMNIST dataset introduces the same confounding squares as DecoyMNIST.

The ISIC (International Skin Imaging Collaboration) Skin Cancer 2019 dataset[34–36] consists of high-resolution dermoscopic images of skin lesions, having either a benign or malignant cancer diagnosis. In contrast to the benchmark datasets and in addition to related work on a medical toy dataset[32], this dataset is substantially more complex and covers a real-world high-stakes scenario. The main difference is that the confounders are not added artificially, and we only know of one confounder, while there can still exist unknown confounders. The known confounders are coloured patches next to a skin lesion. We adjust the original test set as it contains images with both known and unknown confounders. We exclude the images with the known confounder (patches) to ensure a more non-confounded test set, which is essential to measure the confounder influence. Note that the dataset only contains images of Europeans with lighter skin tones, representing the well known skin colour problem, and therefore results cannot be generalized to other skin tones.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All datasets are publicly available. We have two benchmark datasets in which we add a confounder/shortcut, that is we adjust the original dataset and a scientific dataset where the confounder is not artificially added but inherently present in the images. The MNIST dataset is available at http://yann.lecun.com/exdb/mnist/ and the code to generate its decoy version at https://github.com/dtak/rrr/blob/master/experiments/Decoy%20MNIST.ipynb. The FMNIST dataset is available at https://github.com/zalandoresearch/fashion-mnist and the code to generate its decoy version at https://github.com/ml-research/A-Typology-to-Explore-the-Mitigation-of-Shortcut-Behavior/blob/main/data_store/rawdata/load_decoy_mnist.py. The scientific ISIC dataset and its segmentation masks to highlight the confounders are both available at https://isic-archive.com/api/v1/.

Code availability

All the code[37–41] to reproduce the figures and results of this article can be found at https://github.com/ml-research/A-Typology-to-Explore-the-Mitigation-of-Shortcut-Behavior (archived at https://doi.org/10.5281/zenodo.6781501). The CD algorithm is implemented at https://github.com/csinva/hierarchical-dnn-interpretations. Furthermore, other implementations of the evaluated XIL algorithms can be found in the following repositories: RRR at https://github.com/dtak/rrr, CDEP at https://github.com/laura-rieger/deep-explanation-penalization and CE at https://github.com/stefanoteso/calimocho.

References

1. Trust; Definition and Meaning of trust. *Random House Unabridged Dictionary* (2022); https://www.dictionary.com/browse/trust
2. Obermeyer, Z., Powers, B., Vogeli, C. & Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **366**, 447–453 (2019).
3. Holzinger, A. The next frontier: AI we can really trust. In *Machine Learning and Principles and Practice of Knowledge Discovery in Databases—Proc. International Workshops of ECML PKDD 2021* (eds Kamp, M. et al) 427–440 (Springer, 2021).
4. Geirhos, R. et al. Shortcut learning in deep neural networks. *Nat. Mach. Intell.* **2**, 665–673 (2020).
5. Brown, T. et al. Language models are few-shot learners. In *Adv. Neural Inf. Process. Syst* vol. 33. (eds Larochelle, H. et al) 1877–1901 (Curran Associates, Inc., 2020).
6. Ramesh, A., Dhariwal, P., Nichol, A., Chu, C. & Chen, M. Hierarchical text-conditional image generation with CLIP latents. Preprint at *arXiv* https://arxiv.org/abs/2204.06125 (2022).
7. Bender, E. M., Gebru, T., McMillan-Major, A. & Shmitchell, S. On the dangers of stochastic parrots: can language models be

too big? In *Conference on Fairness, Accountability, and Transparency (FAccT)* (eds Elish, M. C. et al.) 610–623 (Association for Computing Machinery, 2021).
8. Angerschmid, A., Zhou, J., Theuermann, K., Chen, F. & Holzinger, A. Fairness and explanation in AI-informed decision making. *Mach. Learn. Knowl. Extr.* **4**, 556–579 (2022).
9. Belinkov, Y. & Glass, J. Analysis methods in neural language processing: a survey. *Trans. Assoc. Comput. Linguist.* **7**, 49–72 (2019).
10. Atanasova, P., Simonsen, J. G., Lioma, C. & Augenstein, I. A diagnostic study of explainability techniques for text classification. In *Proc. Conference on Empirical Methods in Natural Language Processing (EMNLP)* (eds Webber, B. et al) 3256–3274 (Association for Computational Linguistics, 2020).
11. Lapuschkin, S. et al. Unmasking Clever Hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1096 (2019).
12. Teso, S. & Kersting, K. Explanatory interactive machine learning. In *Proc. AAAI/ACM Conference on AI, Ethics, and Society (AIES)* (eds Conitzer, V., et al) 239–245 (Association for Computing Machinery, 2019).
13. Schramowski, P. et al. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.* **2**, 476–486 (2020).
14. Popordanoska, T., Kumar, M. & Teso, S. Machine guides, human supervises: interactive learning with global explanations. Preprint at *arXiv* https://arxiv.org/abs/2009.09723 (2020).
15. Ross, A. S., Hughes, M. C. & Doshi-Velez, F. Right for the right reasons: training differentiable models by constraining their explanations. In *Proc. 26th International Joint Conference on Artificial Intelligence (IJCAI)* (ed Sierra, C.) 2662–2670 (AAAI Press, 2017).
16. Shao, X., Skryagin, A., Schramowski, P., Stammer, W. & Kersting, K. Right for better reasons: training differentiable models by constraining their influence function. In *Proc. 35th Conference on Artificial Intelligence (AAAI)* (eds Honavar, V. & Spaan, M.) 9533-9540 (AAAI, 2021).
17. Rieger, L., Singh, C., Murdoch, W. & Yu, B. Interpretations are useful: penalizing explanations to align neural networks with prior knowledge. In *Proc. International Conference on Machine Learning (ICML)* (eds Daumé, H. & Singh, A.) 8116–8126 (PMLR, 2020).
18. Selvaraju, R. R. et al. Taking a HINT: leveraging explanations to make vision and language models more grounded. In *Proc. IEEE/CVF International Conference on Computer Vision (ICCV)* (ed O'Conner, L.) 2591–2600 (The Institute of Electrical and Electronics Engineers, Inc., 2019).
19. Teso, S., Alkan, Ö., Stammer, W. & Daly, E. Leveraging explanations in interactive machine learning: an overview. Preprint at *arXiv* https://arxiv.org/abs/2207.14526 (2022).
20. Hechtlinger, Y. Interpretation of prediction models using the input gradient. Preprint at *arXiv* https://arxiv.org/abs/1611.07634v1 (2016).
21. Selvaraju, R. R. et al. Grad-CAM: visual explanations from deep networks via gradient-based localization. In *Proc. IEEE International Conference on Computer Vision (ICCV)* (ed O'Conner, L.) 618–626 (The Institute of Electrical and Electronics Engineers, Inc., 2017).
22. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should I trust you?': explaining the predictions of any classifier. In *Proc. 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations* (eds Bansal, M. & Rush, A. M.) 97–101 (Association for Computing Machinery, 2016).
23. Stammer, W., Schramowski, P. & Kersting, K. Right for the right concept: revising neuro-symbolic concepts by interacting with their explanations. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (ed O'Conner, L.) 3618–3628 (The Institute of Electrical and Electronics Engineers, Inc., 2021).
24. Zhong, Y. & Ettinger, G. Enlightening deep neural networks with knowledge of confounding factors. In *Proc. IEEE International Conference on Computer Vision Workshops (ICCVW)* (ed O'Conner, L.) 1077–1086 (The Institute of Electrical and Electronics Engineers, Inc., 2017).
25. Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B. Sanity checks for saliency maps. *Adv. Neural Inf. Process. Syst.* 9505–9515 (2018).
26. Krishna, S. et al. The disagreement problem in explainable machine learning: a practitioner's perspective. Preprint at *arXiv* https://arxiv.org/abs/2202.01602v3 (2022).
27. Tan, A. H., Carpenter, G. A. & Grossberg, S. Intelligence through interaction: towards a unified theory for learning. In *Advances in Neural Networks: International Symposium on Neural Networks (ISNN)* (eds Derong, L. et al) 1094–1103 (Springer, 2007).
28. Dafoe, A. et al. Cooperative AI: machines must learn to find common ground. *Nature* **593** 33–36 (2021).
29. Lang, O. et al. Training a GAN to explain a classifier in StyleSpace. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (ed O'Conner, L.) 673–682 (The Institute of Electrical and Electronics Engineers, Inc., 2021).
30. Anders, C. J. et al. Analyzing ImageNet with spectral relevance analysis: towards ImageNet un-Hans'ed. Preprint at *arXiv* https://arxiv.org/abs/1912.11425v1 (2019).
31. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. Preprint at *arXiv* https://arxiv.org/abs/1702.08608 (2017).
32. Slany, E., Ott, Y., Scheele, S., Paulus, J. & Schmid, U. CAIPI in practice: towards explainable interactive medical image classification. In *AIAI Workshops* (eds Maglogiannis, L. I. et al) 389-400 (Springer, 2022).
33. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *Proc. Third International Conference on Learning Representations (ICLR)* (eds Bengio, Y. & LeCun, B.) (2015).
34. Codella, N. et al. Skin lesion analysis toward melanoma detection: a challenge at the 2017 International Symposium on Biomedical Imaging (ISBI), hosted by the International Skin Imaging Collaboration (ISIC). In *15th International Symposium on Biomedical Imaging (ISBI)* (eds Egan, G. & Salvado, O.) 32–36 (The Institute of Electrical and Electronics Engineers, Inc., 2017).
35. Combalia, M. et al. BCN20000: dermoscopic lesions in the wild. Preprint at *arXiv* https://arxiv.org/abs/1908.02288 (2019).
36. Tschandl, P., Rosendahl, C. & Kittler, H. The HAM10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Sci. Data* **5**, 180161 (2018).
37. Friedrich, F., Stammer, W., Schramowski, P. & Kersting, K. A typology to explore the mitigation of shortcut behavior. *GitHub* https://github.com/ml-research/A-Typology-to-Explore-the-Mitigation-of-Shortcut-Behavior (2022).
38. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proc. International Conference on Learning Representations (ICLR)* (eds Bengio, Y. & LeCun, Y.) 1–14 (2015).
39. Deng, J. et al. Imagenet: a large-scale hierarchical image database. In *Proc. IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 248–255 (The Institute of Electrical and Electronics Engineers, Inc., 2009).
40. Paszke, A. et al. PyTorch: an imperative style, high-performance deep learning library. *Adv. Neural Inf. Process. Syst.* **32**, 8024–8035 (2019).
41. Xiao, H., Rasul, K. & Vollgraf, R. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. Preprint at *arXiv* https://arxiv.org/abs/1708.07747 (2017).

Acknowledgements
We thank L. Meister for preliminary results and insights on this research. This work benefited from the Hessian Ministry of Science and the Arts (HMWK) projects 'The Third Wave of Artificial Intelligence—3AI', hessian.AI (F.F., W.S., P.S., K.K.) and 'The Adaptive Mind' (K.K.), the ICT-48 Network of AI Research Excellence Centre 'TAILOR' (EU Horizon 2020, GA No 952215) (K.K.), the Hessian research priority program LOEWE within the project WhiteBox (K.K.), and from the German Center for Artificial Intelligence (DFKI) project 'SAINT' (P.S., K.K.).

Author contributions
F.F., W.S. and P.S. designed the experiments. F.F. conducted the experiments. F.F., W.S., P.S. and K.K. interpreted the data and drafted the manuscript. K.K. directed the research and gave initial input. All authors read and approved the final manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-023-00612-w.

Correspondence and requests for materials should be addressed to Felix Friedrich.

Peer review information *Nature Machine Intelligence* thanks Mengnan Du and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature research

Corresponding author(s): Felix Friedrich

Last updated by author(s): Dec 7, 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	☐	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We used no software to collect data, we used data from the repositories (cf. data availability statement). We adjusted the benchmark datasets (MNIST and FashionMNIST) with a confounder, here decoy square. The code to generate the decoy version for MNIST (DecoyMNIST) is available at \url{https://github.com/dtak/rrr/blob/master/experiments/Decoy\%20MNIST.ipynb} and for FashionMNIST (DecoyFashionMNIST) at \url{https://github.com/ml-research/A-Typology-to-Explore-the-Mitigation-of-Shortcut-Behavior/blob/main/data_store/rawdata/load_decoy_mnist.py}.
Data analysis	All the code to reproduce the figures and results of this article can be found at \url{https://github.com/ml-research/A-Typology-to-Explore-the-Mitigation-of-Shortcut-Behavior} (archived at \url{https://doi.org/10.5281/zenodo.6781501}). The CD Algorithm is implemented at \url{https://github.com/csinva/hierarchical-dnn-interpretations}. Furthermore, other implementations of the evaluated XIL algorithms can be found in the following repositories: RRR at \url{https://github.com/dtak/rrr}, CDEP at \url{https://github.com/laura-rieger/deep-explanation-penalization}, and CE at \url{https://github.com/stefanoteso/calimocho}.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All datasets are publicly available. We have two benchmark datasets, in which we add a confounder/ shortcut. The MNIST dataset is available at \url{http://yann.lecun.com/exdb/mnist/} and the FashionMNIST dataset is available at \url{https://github.com/zalandoresearch/fashion-mnist}. The scientific ISIC dataset with its segmentation masks to highlight the confounders are both available at \url{https://isic-archive.com/api/v1/}.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	*Briefly describe the study type including whether data are quantitative, qualitative, or mixed-methods (e.g. qualitative cross-sectional, quantitative experimental, mixed-methods case study).*
Research sample	*State the research sample (e.g. Harvard university undergraduates, villagers in rural India) and provide relevant demographic information (e.g. age, sex) and indicate whether the sample is representative. Provide a rationale for the study sample chosen. For studies involving existing datasets, please describe the dataset and source.*
Sampling strategy	*Describe the sampling procedure (e.g. random, snowball, stratified, convenience). Describe the statistical methods that were used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient. For qualitative data, please indicate whether data saturation was considered, and what criteria were used to decide that no further sampling was needed.*
Data collection	*Provide details about the data collection procedure, including the instruments or devices used to record the data (e.g. pen and paper, computer, eye tracker, video or audio equipment) whether anyone was present besides the participant(s) and the researcher, and whether the researcher was blind to experimental condition and/or the study hypothesis during data collection.*
Timing	*Indicate the start and stop dates of data collection. If there is a gap between collection periods, state the dates for each sample cohort.*
Data exclusions	*If no data were excluded from the analyses, state so OR if data were excluded, provide the exact number of exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Non-participation	*State how many participants dropped out/declined participation and the reason(s) given OR provide response rate OR state that no participants dropped out/declined participation.*
Randomization	*If participants were not allocated into experimental groups, state so OR describe how participants were allocated to groups, and if allocation was not random, describe how covariates were controlled.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging