Human Cognitive Biases in Explanation-based Interaction: The Case of Within and Between Session Order Effect

Dario Pesenti[*1], **Alessandro Bogani**[*2], **Katya Tentori**[1] **Stefano Teso**[1,2]
[1]CIMeC, University of Trento
[2]DISI, University of Trento
name.surname@unitn.it

Abstract

Explanatory Interactive Learning (XIL) is a powerful interactive learning framework designed to enable users to customize and correct AI models by interacting with their explanations. In a nutshell, XIL algorithms select a number of items on which an AI model made a decision (*e.g.* images and their tags) and present them to users, together with corresponding explanations (*e.g.* image regions that drive the model's decision). Then, users supply corrective feedback for the explanations, which the algorithm uses to improve the model. Despite showing promise in debugging tasks, recent studies have raised concerns that explanatory interaction may trigger *order effects*, a well-known cognitive bias in which the sequence of presented items influences users' trust and, critically, the quality of their feedback. We argue that these studies are not entirely conclusive, as the experimental designs and tasks employed differ substantially from common XIL use cases, complicating interpretation. To clarify the interplay between order effects and explanatory interaction, we ran two larger-scale user studies ($n = 713$ total) designed to mimic common XIL tasks. Specifically, we assessed order effects both *within* and *between* debugging sessions by manipulating the order in which correct and wrong explanations are presented to participants. Order effects had a limited, though significant, impact on users' agreement with the model (*i.e.*, a behavioural measure of their trust), and only when examined within debugging sessions, not between them. The quality of users' feedback was generally satisfactory, with order effects exerting only a small and inconsistent influence both within and between sessions. Overall, our findings suggest that order effects do not pose a significant issue for the successful employment of XIL approaches. More broadly, our work contributes to the ongoing efforts for understanding human factors in AI.

Code — https://gitlab.pavlovia.org/DarioP/ordereffects

1 Introduction

Explainable AI (XAI) tools help stakeholders inspect, understand, and evaluate the behavior of – otherwise opaque – AI models by generating *explanations* of their decisions (Molnar 2020; Schwalbe and Finzel 2024). These can often reveal defects in the model's reasoning that would be difficult to detect through standard evaluation alone, such as sub-optimal feature usage (Kulesza et al. 2015) and reliance on confounded features (Geirhos et al. 2020; Lapuschkin et al. 2019; Ye et al. 2024). *Explanatory interactive learning* (XIL) is a powerful framework that builds on this observation to improve user control (Kulesza et al. 2015; Teso and Kersting 2019). XIL algorithms repeatedly select a number of items on which an AI model made a decision (*e.g.* images and their tags) and present them to users together with corresponding local explanations (*e.g.* highlighting what regions of the image drove the model's decision). Then, they ask users to evaluate whether the provided explanations are satisfactory and, if not, to supply corrective feedback, which the algorithms use to improve the AI model. By steering directly the model's explanations, XIL enables users to quickly debug and customize AI models (Lertvittayakumjorn and Toni 2021; Teso et al. 2023); see Section 3.2 for a broader overview.

Recent studies (Nourani et al. 2021) have shown that the order in which AI predictions are presented to users may affect how users interact with them. *Order effects* refer to a class of cognitive biases (Hogarth and Einhorn 1992) in which human judgments are systematically influenced by the sequence in which information is presented. More specifically, a *primacy* effect occurs when earlier information is given disproportionately more weight than more recent information, whereas a *recency* effect arises when the opposite happens, with more recent information exerting greater influence. In the context of intelligent systems, Nourani et al. (2021) have shown that, in an non-XIL interactive setting, users witnessing correct AI behaviour early tend to build a more accurate mental model of the AI but also to *over-rely on its suggestions* and to *make more mistakes* when prompted to decide themselves, while those that encounter errors early tend to underestimate the AI's competencies. This suggests that, in XIL, *order effects may compromise the quality of human feedback and thus its reliability in applications*.

We contend that the question is not yet settled, as Nourani et al.'s study differs substantially from typical XIL use cases (cf. Section 2 for a discussion). Their findings also partially conflict with those of Honeycutt, Nourani, and Ragan (2020), who, in a more interactive design, found a detrimental impact of letting users interact with the model on their

[*]These authors contributed equally.

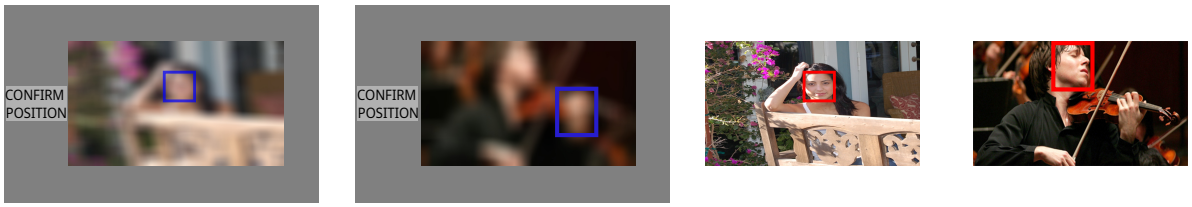

Figure 1: Interface of the user studies (left): participants were asked to evaluate the model's *explanation* (represented by a bounding box, in blue) of a fictional image classifier. Ideally, this should entirely enclose the face of the person pictured in the image. Participants were instructed to press the "Confirm" button if they deemed the box enclosed the face, or to move it onto the face otherwise, all within a 6 seconds time limit. Right: original images and ground-truth bounding boxes (in red), for reference.

perception of it, but no significant order effects.

To understand the interplay between explanatory interaction and order effects, we carry out two controlled, larger scale user studies ($n = 713$ total) simulating a XIL task. During the experiments, participants had to interact with explanations from an image classifier that was fictitious, unbeknownst to them. More specifically, participants were instructed to either accept the model's explanation, or to correct it (cf. Fig. 1).

We considered two settings representative of actual XIL usage: in the **within-session** setting, we focused on order effects occurring within a single debugging session, before the AI model was updated; whereas in the **between-session** setup, participants carried out two consecutive debugging sessions, believing that the AI model had been updated in between. Our results indicate that the order of presentation of AI outputs exerted only small effects on the quality of users' feedback, which presented overall high levels, in both within- and between-session settings, as well as on their implicit agreement with the model. Only a small order effect was found within-session. No effect was found on users' perceptions of accuracy and trustworthiness of the model. In summary, *order effects do not appear to pose a significant issue for the successful employment of XIL approaches*. Our work contributes to ongoing efforts for understanding human factors in AI and explanation-based interaction.

2 Preliminaries

Explanations and interaction. *Explanatory interactive learning* (XIL) operationalizes the observation that if a (sufficiently expert) user understands how a model works, they can – and, typically, proactively want to (Kulesza et al. 2015) – supply corrective feedback useful for improving the model itself (Teso and Kersting 2019; Schramowski et al. 2020). For instance, in a medical diagnosis task using X-ray scans, machine learning classifiers can achieve high accuracy by exploiting confounding factors, *e.g.* background cues that correlate with the decision but are not causally related. This compromises out-of-sample performance while being invisible to standard metrics like accuracy. By highlighting what features the model relies on, local explanations such as saliency maps (Miller 2019) help identify such issues (Geirhos et al. 2020; Lapuschkin et al. 2019) and enable users to formulate corrective feedback (*e.g.* "don't use this part of the scan") (Ross, Hughes, and Doshi-Velez 2017).

XIL algorithms loop through two steps. During the **debugging session**, the machine iteratively selects a number of items (*e.g.* images) from a pool of options and computes predictions (tags) and explanations (saliency maps) for them. The items and the corresponding predictions and explanations are then presented to a user, who, for each of them, can indicate what features the machine is using improperly. In the **update step**, the collected feedback is used to update the model. XIL approaches were shown to, *e.g.* help laypeople to quickly tailor spam filters to their needs (Kulesza et al. 2015) and domain experts to rectify confounding in scientific studies (Schramowski et al. 2020). However, it has been mostly neglected whether human biases – such as order effects, which are naturally triggered by the XIL loop – can impair the adoption of XIL algorithms.

Order effects and XIL. Research in psychology has shown that order effects are pervasive and impactful, influencing item recall (Furnham and Boo 2011; Baddeley and Hitch 1993), belief updating (Hogarth and Einhorn 1992), response accuracy (Eisenberg and Barry 1988), and even preferences in political elections (Matsusaka 2016). They have received increasing attention also in the XAI literature (Nourani et al. 2021; Nourani, King, and Ragan 2020; Nourani et al. 2022). Indeed, both within- and between-session order effects can influence the quality of debuggers' performance, and each requires ad hoc solutions. For example, the former can be addressed by randomizing the order of item presentation, whereas the latter would necessitate recruiting different groups of debuggers across sessions. It is therefore essential to ascertain the extent to which such effects affect performance; nevertheless, only a limited number of studies have investigated them within interactive learning contexts. Honeycutt, Nourani, and Ragan (2020) showed that allowing the users to correct and interact with a model reduced their trust in it and perception of its accuracy overall, but they found no order effects on trust over multiple debugging sessions. However, they examined only between-session order effects, and focused exclusively on self-report measures, which are less reliable than behavioral indices of accuracy (Warren, Byrne, and Keane 2024) and are often poor predictors of individuals' actual behaviour (Sheeran and Webb 2016). In another work, Nourani et al. (2021) reported an increased mistrust in the model if they

were exposed to more incorrect decisions first; conversely, they found an automation effect (*i.e.* over-reliance on model predictions (Cummings 2012; Rastogi et al. 2022)) when the users were presented with more correct AI decisions first, which was also reflected in a worse task performance by the participants. In this case, however, they studied only within-session order effects where participants, unlike typical XIL use cases, self-selected the stimuli to inspect and could not correct the model's outputs.

3 User Studies

Our experiments aim to provide a more reliable exploration of the potential negative impact of order effects on XIL algorithms. Specifically, we investigate the influence of order effects both *within-session* (Experiment 1) and *between-sessions* (Experiment 2), evaluating not only users' perceptions, but also *their actual behaviour* in a more ecological debugging task. Next, we introduce the task, variables and data processing common to both experiments.

Task. Based on the work of (Honeycutt, Nourani, and Ragan 2020), we implemented a debugging task of a binary classifier trained to determine the presence of human faces in noisy images. In this task, that comprised either a single (Experiment 1) or two debugging sessions (Experiment 2), participants were shown a series of blurred images containing a face, each accompanied by a bounding box indicating the face location according to the model (*i.e.* the explanation for what part of the image led the model to categorize the image as presenting a face).[1] The box was placed either correctly or incorrectly (*i.e.* perfectly around the human face or not; see below for further details) and participants were asked to provide feedback about it: if they agreed with it, they had to click a button to confirm its position; conversely, if they deemed the box misplaced, they had to move the bounding box to the position they considered correct with the mouse (the interface is shown in Fig. 1.) Importantly, participants' could change only the position of the box, but not its dimensions. For each image, participants had 6 seconds to provide their feedback: if no response was given within this time frame, the trial was recorded as a "missed" response.

Images and correct bounding boxes. All images were sourced from the Open Images Dataset V7 (Kuznetsova et al. 2020) a well-known repository of natural images. We selected images containing a single unoccluded, front facing human face, and recorded them along with the face's correct "ground-truth" bounding boxes provided by OpenImages. The selected images were applied a substantial amount of Gaussian blur (85×85 kernel, $\sigma = 40$). This was done to make the task less trivial compared to the one employed in Honeycutt, Nourani, and Ragan (2020), as a task in which the model's outputs are clearly correct or incorrect might reduce the chances of observing possible order effects (as partially supported by our findings, see Section 3.1).

[1]We chose face recognition as a use case because it is accessible, requiring little domain knowledge, yet sufficiently complex to vary stimulus difficulty and avoid floor or ceiling effects.

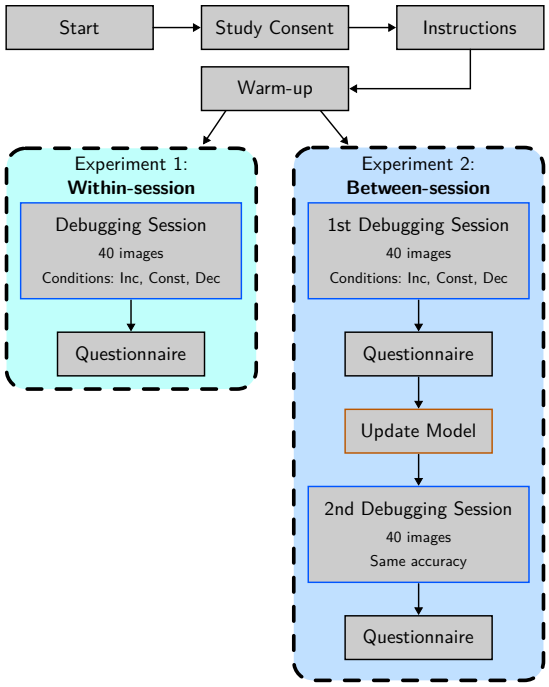

Figure 2: Schematic illustration of our user studies.

Variables and Analyses. In both experiments, we manipulated the following three independent variables.

1. *Order*: It refers to how the accuracy of the model (represented by the proportion of images featuring a correct box) evolved throughout the debugging task (*i.e.* between the first and second half of the single session in Experiment 1 and between the two sessions in Experiment 2; see Section 3.1). It presented 3 levels (between subjects): increasing (*i.e.* accuracy improved throughout the task), constant (*i.e.* accuracy remained constant), and decreasing (*i.e.* accuracy worsened), abbreviated as *Inc*, *Const*, and *Dec*, respectively.
2. *Placement*: It refers to the accuracy of the model's placement of the bounding box, defined as the percentage of overlap between model-placed boxes and the ground truth. It presented 2 levels (within-subject): *correct* vs. *incorrect*. Correct placements consist in bounding boxes coinciding perfectly with the ground truth. Among the incorrect placements, we distinguish between two categories: 'partially wrong' (with a 25% overlap between the model's placement and the ground truth) and 'wrong' (with no overlap at all). We introduced this distinction to represent the variety of errors that a model might make, but due to the low number of stimuli belonging to each of this two sub-classes (see Section 3.1) they were considered as a single "incorrect" class in the statistical analyses (but see Appendix B for plots of the results split by the three placement levels).[2]

[2]False positives, where the model detects a non-existent face, were excluded, as XIL methods chiefly target models that make correct predictions for the wrong reasons (Teso et al. 2023).

3. *Difficulty*: It refers to how challenging it was to locate the human face in the image, and it presented 2 levels (within-subjects): *easy* vs. *difficult*. This variable was included to assess whether order effects, if present, influenced responses across all stimuli or specifically in those where the correct answer was more ambiguous. Image difficulty was determined through pilot studies in which candidate images (together with their respective bounding boxes) were presented to participants who had to confirm or correct the position of the box. Easy and difficult images were selected, respectively, among the ones presenting the highest and lowest levels of accuracy (calculated as the percentage of overlap between participant-placed boxes and the ground truth).

In both experiments, participants across all conditions viewed the same set of images. Thus, any differences between the Inc and Dec groups could be attributed solely to presentation order, with each group effectively serving as a control for the other. The inclusion of the Const group, in which no order effects were present, provided a baseline condition that enabled an even more precise interpretation of potential differences between the other two groups.

We then measured three dependent variables:

1. *Accuracy of participants' feedback*, quantified as the overlap between the area corresponding to the ground truth and that of participants' placement, divided by the former;
2. *Participants' agreement with the model*, quantified as the overlap between the areas of the model's box and the participants' box, divided by the former;
3. *Participants' perception of the model's accuracy and their trust in it*, quantified through a questionnaire comprising four 7-point Likert scale items adapted from Honeycutt, Nourani, and Ragan (2020) and Hoffman et al. (2019).

The exact wording of the task instructions and the questionnaire items is reported in Appendix A.4.

The accuracy and agreement of the participants with the model were analyzed using mixed linear models that present the order condition, the placement of the model, the difficulty of the image, and their interactions as fixed effects. Random intercepts were included for both participants and images. Significant fixed effects involving more than one contrast between groups were further investigated through Bonferroni-corrected post-hoc comparisons. Participants' answers to the four questionnaire items were averaged to compute an index of perceived model accuracy and trustworthiness and analyzed by means of a Kruskal-Wallis rank sum test to investigate potential differences between the three order conditions.

Participants and data exclusions. An a priori power analysis, conducted by means of a simulation approach implemented in R (Green and MacLeod 2016; Kumle, Võ, and Draschkow 2021), indicated that, for both experiments, a total sample of at least 330 participants evaluating 40 images each would provide 82% power to detect a small-to-medium effect of the interaction among the three independent variables, as well as their main effects. Participants were recruited on Prolific[3] among those with a Prolific approval rate of at least 98%, and compensated in accordance with the hourly payment suggested by Prolific (£1 for Experiment 1 and £1.30 for Experiment 2, plus a possible bonus payment of £10; see Section 3.1).

Before running the analyses, we excluded all observations in which participants did not confirm nor changed the position of the box, as well as those in which the last registered input was made at 5.95 sec or later from the onset of the image out of the total 6 seconds available, as these likely reflected cases in which participants had not reached a conclusive answer by the end of the trial.[4]

3.1 Experiment 1: Within-session Order Effect

Aim. In Experiment 1, we investigated *whether the distribution of the model's errors within a single debugging session influenced how participants interacted with and perceived the model*. More specifically, we examined whether a prevalence of the model's errors either at the beginning or at the end of the session, while keeping the model's overall accuracy exactly the same, affected participants' performance in the debugging task and their trust in the model.

Procedure. All participants provided informed consent and received identical instructions on how to perform the debugging task, without any mention of the different experimental conditions. To encourage attentiveness, we told participants that at the end of data collection, five debugging trials would be randomly selected. Three participants, chosen at random from those who provided correct responses in all five trials, would receive a £10 bonus. To ensure fairness, both random selections were carried out, and the bonus was awarded accordingly. Participants completed six warm-up trials to familiarize themselves with the task before starting the actual debugging session, which involved evaluating 40 images. All conditions presented a model with an overall accuracy equal to 60%, but they differed in how the model's performance evolved throughout the debugging session. Between the first and the second half of the session, model's accuracy increased (from 40% to 80%) in the *Inc* condition, remained constant (60%) in the *Const* condition, and decreased (from 80% to 40%) in the *Dec* condition (see Section 3.1). The order of presentation of the images was randomized once and kept fixed afterward for all participants within each condition. After the debugging task, participants filled out the questionnaire on perceived model's accuracy and trustworthiness. Finally, they were asked to provide information about their experience with intelligent systems and debugging procedures.

Sample composition and data quality. A total of 359 participants ($M_{Age} = 34.58 \pm 9.72$; 51% female) were recruited and evenly distributed across the three order conditions ($N_{Inc} = 119$; $N_{Const} = 121$; $N_{Dec} = 119$). A chi-

[3] www.prolific.com

[4] Still, running the analyses without excluding this second class of stimuli did not change the results.

Within-session	First half of the session						Second half of the session					
	Easy			Hard			Easy			Hard		
Condition	Corr.	P. Corr.	Wrong	Corr.	P. Corr.	Wrong	Corr.	P. Corr.	Wrong	Corr.	P. Corr.	Wrong
Increasing	4	3	3	4	3	3	8	1	1	8	1	1
Constant	6	2	2	6	2	2	6	2	2	6	2	2
Decreasing	8	1	1	8	1	1	4	3	3	4	3	3

Between-session	First session						Second session					
	Easy			Hard			Easy			Hard		
Condition	Corr.	P. Corr.	Wrong	Corr.	P. Corr.	Wrong	Corr.	P. Corr.	Wrong	Corr.	P. Corr.	Wrong
Increasing	8	9	3	8	9	3						
Constant	12	6	2	12	6	2	12	6	2	12	6	2
Decreasing	16	3	1	16	3	1						

Table 1: Number of images presented in Experiment 1 (within-session, top) and 2 (between-sessions, bottom), split by Condition (Increasing, Constant, and Decreasing model performance over time), by Image Difficulty (Easy vs. Hard), and by Correctness of the model's explanation (Correct, Partially Wrong, and Wrong).

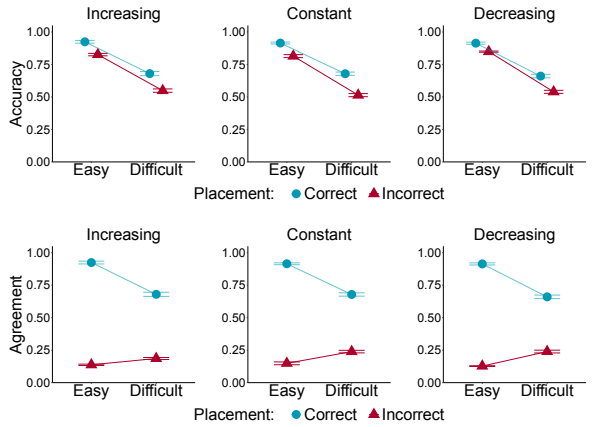

Figure 3: Average accuracy (top) and agreement (bottom) in Experiment 1 divided by order condition, correctness of model's placement (correct and incorrect), and image difficulty. Error bars represent standard errors.

square test of independence indicated that participants' experience with programming and debugging machine learning algorithms did not differ significantly between the three conditions ($p = .188$), ensuring balanced samples in this regard. Also, the proportion of excluded observations was limited and roughly equal in the three order conditions (*Inc* $= 8\%$; *Const* $= 7\%$; *Dec* $= 8\%$), suggesting that participants actively engaged with the task during the majority of trials.

Results

Accuracy. Participants' overall accuracy in the debugging task was fairly high in all three order conditions (*Inc*: 0.76 ± 0.10; *Const*: 0.75 ± 0.08; *Dec*: 0.76 ± 0.08), albeit it varied depending on stimuli features: participants were more accurate when evaluating easy (0.88 ± 0.08) than difficult images (0.62 ± 0.12; $F(1, 36) = 63.33$, $p < .001$) and when evaluating correct (0.80 ± 0.11) than incorrect images (0.69 ± 0.10; $F(1, 36) = 10.65$, $p = .002$, see Figure 3).[5] As for the effect of order condition, only its two-way interaction with model's placement was significant, $F(2, 12875) = 4.56$, $p = .011$. Post-hoc tests revealed that the interaction was driven by participants' accuracy being more similar between incorrect and correct images in the *Dec* condition (Correct: 0.79 ± 0.11; Incorrect: 0.70 ± 0.08) compared to the *Const* one (Correct: 0.80 ± 0.10; Incorrect: 0.67 ± 0.11), $p = .008$. A possible interpretation of this result is that participants in the *Dec* condition encountered most of the incorrect images in the latter half of the experiment, by which point they may have gained more confidence with the task. However, given that the difference was small and limited to the contrast between *Dec* and *Const*, it may also reflect random noise.

Agreement. As expected, there was a main effect of the model's placement, with agreement being higher when the box was placed correctly (0.80 ± 0.11) than when it was not (0.18 ± 0.08), $F(1, 36) = 174.38, p < .001$. The interaction between model's placement and image difficulty was significant as well, $F(1, 36) = 12.80$, $p = .001$, with agreement rates for difficult images being significantly lower than that for easy ones when the box was correctly placed (Easy: 0.92 ± 0.10; Difficult: 0.67 ± 0.16, $p < .001$) but not when it was placed incorrectly (Easy: 0.14 ± 0.08; Difficult: 0.22 ± 0.11, $p = .458$). Crucially, the three-way interaction was significant, $F(2, 12862) = 7.99$, $p < .001$: when the box placement was incorrect, in all conditions, the agreement between participants and the model tended to be higher for difficult compared to easy images; however this tendency was reduced in the *Inc* condition (Easy: 0.14 ± 0.06; Difficult: 0.19 ± 0.08) compared to the *Const* (Easy: 0.15 ± 0.12; Difficult: 0.24 ± 0.11; $p = .027$) and the *Dec* ones (Easy: 0.13 ± 0.03; Difficult: 0.24 ± 0.12; $p < .001$). This finding may be interpreted as evidence of a small primacy effect: *early exposure to the model's inaccuracy, as in the Inc condition, led participants to rely less on the model, especially when the images were ambiguous*. In line with this interpre-

[5] For significant effects of the mixed models we report the value of the test statistic (F) and relative degrees of freedom (in parentheses).

tation, the effect described above was primarily driven (see Figure 3) by difficult, partially wrong images, for which participants were likely most uncertain about the correctness of the model's placement.

Questionnaire. The Kruskal-Wallis test on the perceived accuracy and trust index indicated that the three order conditions did not differ in how they evaluated the model after the debugging session ($p = .909$). Indeed, the average values of the index were almost identical for participants in the *Inc* (3.06 ± 1.03), *Const* (3.11 ± 1.11), and *Dec* (3.13 ± 1.00) conditions.

Discussion. The results of Experiment 1 suggest that the order in which participants are exposed to correct and incorrect explanations from a model within a single debugging session have, at best, a limited effect on the quality of their feedback and on their tendency to agree with the model. The only finding that is clearly interpretable as an order effect is that experiencing a consistent number of model failures early in a debugging session may reduce participants' reliance on the model's outputs, particularly when dealing with difficult stimuli. Interestingly, participants appeared to be unaware of this effect, as their responses to the questionnaire on perceived accuracy and trust showed no differences across order conditions.

3.2 Experiment 2: Between-session Order Effect

Aim. In Experiment 2, we investigated possible order effects *between two distinct sessions of a debugging task*. We presented participants with a simulated model whose accuracy either increased, remained constant, or decreased after a fictitious update between the two sessions, with the second session being identical for all participants.

Procedure. The procedure was similar to that of Experiment 1. However, in this case, after the six warm-up trials, participants completed two debugging sessions, each consisting in the evaluation of 40 images. In the first session, the model's accuracy varied across the three order conditions (*Inc*: 40%; *Const*: 60%; *Dec*: 80%). The images were selected, based on the number required by each condition, from a larger set consisting of 32 correct images (16 easy and 16 difficult), 18 partially wrong images (9 easy and 9 difficult), and 6 wrong images (3 easy and 3 difficult). (See Section 3.1.) This selection was carried out through 6 random draws, one for each image type. For example, the 8 correct easy images of the Inc condition were a subset of the 12 correct easy images of the Const condition, which, in turn, were a subset of the 16 correct easy images of the Dec condition. These subsets were randomly drawn once, and the order of images was shuffled and kept constant across participants within each condition.

Following the first session, participants completed the same questionnaire used in Experiment 1. They then waited for five seconds, during which they were told the model was being updated based on feedback from all users. Participants then proceeded to the second session, which was identical across all order conditions in terms of both the images shown (model's accuracy was 60%) and presentation order. Importantly, within both sessions, image presentation was controlled to distribute errors evenly in order to minimize potential within-session order effects. After completing the second session, participants, once again, filled in the same questionnaire and reported on their prior experiences with intelligent systems and debugging procedures.

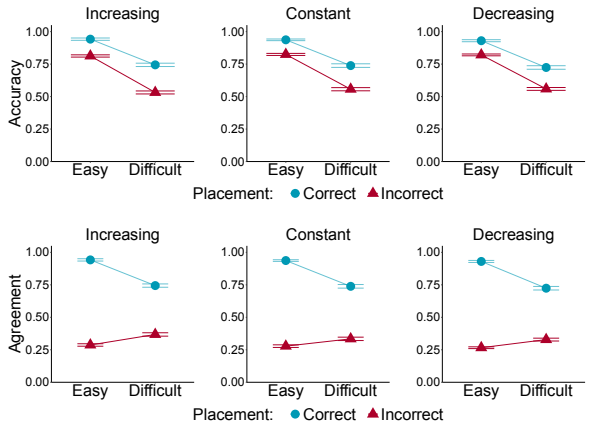

Figure 4: Average accuracy (top) and agreement (bottom) in Experiment 2 divided by order condition, correctness of model's placement (correct and incorrect), and image difficulty. Error bars represent standard errors.

Sample composition and data quality. A total of 354 participants ($M_{Age} = 36.33 \pm 10.64$, 47.5% female) were recruited and randomly distributed across the three order conditions ($N_{Inc} = 121$; $N_{Const} = 117$; $N_{Dec} = 116$). Yet again, experience with programming and debugging machine learning algorithms did not differ significantly between the three order conditions, $p = .278$, and engagement with the task was satisfactory, as indicated by the limited proportion of excluded observations *Inc* $= 6\%$; *Const* $= 6\%$; *Dec* $= 7\%$).

Results

Manipulation check. To ensure that the manipulation of the model performance in the first session was effective, we assessed the agreement with the model through a mixed model including order condition as a fixed effect and random intercepts for participants. Order condition predicted agreement, $F(2, 305.92) = 483.39$, $p < .001$, with post-hoc comparisons indicating that agreement was significantly lower in the *Inc* (0.42 ± 0.36) than in the *Const* condition (0.57 ± 0.38, $p < .001$) and in the *Const* than in the *Dec* condition (0.70 ± 0.35, $p < .001$).

Accuracy. In the second session, participants in all three order conditions showed a good average level of accuracy (*Inc*: 0.78 ± 0.08; *Const*: 0.78 ± 0.07; *Dec*: 0.78 ± 0.08). As expected, this was significantly higher for easy (0.89 ± 0.06) than for difficult images (0.66 ± 0.11; $F(1, 36) = 47.81$, $p < .001$) and for correct (0.84 ± 0.11) than for incorrect images (0.69 ± 0.09; $F(1, 36) = 19.55$, $p < .001$, see Fig. 4, top). As in Experiment 1, the effect of order condition was significant only in interaction with model's

placement, $F(2, 12915.6) = 8.61$, $p < .001$. Post-hoc contrasts indicated that, in this case, the effect was driven by a slightly more pronounced reduction in the accuracy for incorrect than correct images in the *Inc* condition (Correct: 0.84 ± 0.11; Incorrect: 0.68 ± 0.09) compared to both the *Const* (Correct: 0.84 ± 0.10; Incorrect: 0.69 ± 0.09; $p = .002$) and *Dec* conditions (Correct: 0.83 ± 0.11; Incorrect: 0.70 ± 0.09; $p = .001$). One possible interpretation of this result is that participants in the Inc condition became more reliant on the model during the second session, following a perceived improvement in its accuracy. However, the small magnitude of the effect warrants caution in interpreting this finding, as it may simply reflect noise.

Agreement. In the second session, as in Experiment 1, we observed a significant main effect of model's placement on participants' agreement with the model, which was higher when the box was placed correctly (0.84 ± 0.11) than when it was not (0.31 ± 0.10; $F(1, 36) = 112.04$, $p < .001$). The effect was further qualified by a significant interaction with image difficulty ($F(1, 36) = 7.06$, $p = .012$): difficult images resulted in significantly lower agreement than easy ones when the box was correct (Easy: 0.94 ± 0.08; Difficult: 0.74 ± 0.15; $p = .006$), but not when it was incorrect (Easy: 0.28 ± 0.10; Difficult: 0.34 ± 0.14; $p = .816$). Crucially, no significant effect of order condition was observed, and indeed participants' overall agreement in the *Inc* (0.63 ± 0.10), *Const* (0.62 ± 0.10), and *Dec* conditions (0.61 ± 0.09) were similar and all converging toward intermediate values (see Fig. 4, bottom).

Questionnaire. As in Experiment 1, the Kruskal-Wallis test on the perceived accuracy and trust index suggested that participants in the *Inc* (3.38 ± 1.21), *Const* (3.43 ± 1.17), and *Dec* (3.34 ± 1.23) conditions did not differ in their perceptions of the model ($p = .821$).

Discussion. The results of Experiment 2 indicate that between-session order effects have a limited impact on users during XIL debugging sessions. Specifically, while there may be a minimal influence of order on users' accuracy, the order of debugging sessions clearly did not affect participants' agreement with the model or their perception of it.

4 Conclusion

XIL can substantially help to enhance and steer the behavior of AI models. A number of XIL approaches have been proposed, see (Lertvittayakumjorn and Toni 2021; Teso et al. 2023) for an overview. While some methods are model-agnostic (Plumb, Ribeiro, and Talwalkar 2021; Slany et al. 2022; Michiels, De Vos, and Suykens 2023), others are tailored for specific architectures, including neural networks (Teso 2019; Mitsuhara et al. 2019; Schramowski et al. 2020; Shao et al. 2021) and concept-based models (Lertvittayakumjorn, Specia, and Toni 2020; Stammer, Schramowski, and Kersting 2021). However, *most works focus on algorithmic, rather than human, factors*, such as how to best integrate feedback into the model (Schramowski et al. 2020; Shao et al. 2021; Michiels, De Vos, and Suykens 2023), and how to leverage alternatives to saliency maps, like examples (Teso et al. 2021; Zylberajch, Lertvittayakumjorn, and Toni 2021) and high-level concepts (Stammer, Schramowski, and Kersting 2021). In XIL, items are choosen by the machine, usually according to their predictive uncertainty (Settles 2011), and presented sequentially, potentially triggering order effects. Popordanoska, Kumar, and Teso (2020) argue that this setup prioritizes easy-to-learn items, and as such it may fool users into overtrusting the model, and propose human-initiated interaction as a possible solution (Attenberg and Provost 2010). Their observation that the order in which items are presented – which depends entirely on the machine's query selection strategy – can induce overreliance partially motivates our work.

Comparison with Related Work. Previous results suggested that letting users interact with a model during debugging can unduly decrease their trust in it (Honeycutt, Nourani, and Ragan 2020) and expose them to the influence of biases, worsening the reliability of their feedback (Nourani et al. (2021)). We examined these potential issues by engaging users in a realistic XIL debugging task and assessing whether participants' performance and perceptions were influenced by within- and between-session order effects, which may be both naturally triggered by the XIL loop. Our results showed that participants provided high-quality feedback and appropriately adjusted their agreement with the model based on its performance, and that these variables were only minimally affected by order effects. These results were confirmed by participants' self-reported perceptions, which did not show any difference between conditions In particular, we observed that participants' agreement with the model was influenced by presentation order only within (but not between) sessions, and only with certain kind of stimuli (i.e. difficult images), suggesting that participants may reset their expectations when the model is updated. This possibility is reassuring, since it indicates that participants can adapt to model changes without being biased by prior exposure, making XIL algorithms overall robust to the influence of order effects, provided users are informed about model updates. This finding warrants further investigation to determine the optimal procedure for implementing debugging sessions.

Limitations and future directions. Our work focuses exclusively on order effects arising from differences in the distribution of model errors, which are naturally triggered by the XIL loop. Future work should consider other types of prediction tasks, beyond image classification, and other widely used families of explanations, such as concept-level explanations (Stammer, Schramowski, and Kersting 2021) and concrete examples (Zylberajch, Lertvittayakumjorn, and Toni 2021). It could also be worth exploring different types of order effects, *e.g.* those stemming from the sequence in which more or less challenging items or explanations are presented, as these may influence participants' accuracy by modulating their confidence in their ability to perform the task. More broadly, future work could examine whether the implementation of XIL algorithms is effective across different types of tasks, and whether other forms of bias might more substantially hinder their adoption.

Ethical Statement

Our study has received approval from the Ethics board of our university, document identifier code 2025-001ESA.

Acknowledgments

Funded by the European Union. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Health and Digital Executive Agency (HaDEA). Neither the European Union nor the granting authority can be held responsible for them. Grant Agreement no. 101120763 - TANGO.

References

Attenberg, J.; and Provost, F. 2010. Why label when you can search? Alternatives to active learning for applying human resources to build classification models under extreme class imbalance. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 423–432.

Baddeley, A. D.; and Hitch, G. 1993. The recency effect: Implicit learning with explicit retrieval? *Memory & Cognition*, 21: 146–155.

Cummings, M. 2012. Automation Bias in Intelligent Time Critical Decision Support Systems. In *Collection of Technical Papers - AIAA 1st Intelligent Systems Technical Conference*.

Eisenberg, M.; and Barry, C. 1988. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science*, 39(5): 293–300.

Furnham, A.; and Boo, H. C. 2011. A literature review of the anchoring effect. *The Journal of Socio-Economics*, 40(1): 35–42.

Geirhos, R.; Jacobsen, J.-H.; Michaelis, C.; Zemel, R.; Brendel, W.; Bethge, M.; and Wichmann, F. A. 2020. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11): 665–673.

Green, P.; and MacLeod, C. J. 2016. SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4): 493–498.

Hoffman, R. R.; Mueller, S. T.; Klein, G.; and Litman, J. 2019. Metrics for Explainable AI: Challenges and Prospects. arXiv:1812.04608.

Hogarth, R. M.; and Einhorn, H. J. 1992. Order effects in belief updating: The belief-adjustment model. *Cognitive Psychology*, 24(1): 1–55.

Honeycutt, D.; Nourani, M.; and Ragan, E. 2020. Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 63–72.

Kulesza, T.; Burnett, M.; Wong, W.-K.; and Stumpf, S. 2015. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th international conference on intelligent user interfaces*, 126–137.

Kumle, L.; Võ, M. L.-H.; and Draschkow, D. 2021. Estimating power in (generalized) linear mixed models: An open introduction and tutorial in R. *Behavior research methods*, 53(6): 2528–2543.

Kuznetsova, A.; Rom, H.; Alldrin, N.; Uijlings, J.; Krasin, I.; Pont-Tuset, J.; Kamali, S.; Popov, S.; Malloci, M.; Kolesnikov, A.; Duerig, T.; and Ferrari, V. 2020. The Open Images Dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*.

Lapuschkin, S.; Wäldchen, S.; Binder, A.; Montavon, G.; Samek, W.; and Müller, K.-R. 2019. Unmasking Clever Hans predictors and assessing what machines really learn. *Nature communications*, 10(1): 1–8.

Lertvittayakumjorn, P.; Specia, L.; and Toni, F. 2020. FIND: human-in-the-loop debugging deep text classifiers. In *Conference on Empirical Methods in Natural Language Processing*, 332–348.

Lertvittayakumjorn, P.; and Toni, F. 2021. Explanation-Based Human Debugging of NLP Models: A Survey. *arXiv preprint arXiv:2104.15135*.

Matsusaka, J. G. 2016. Ballot order effects in direct democracy elections. *Public choice*, 167: 257–276.

Michiels, J.; De Vos, M.; and Suykens, J. 2023. Increasing Performance And Sample Efficiency With Model-agnostic Interactive Feature Attributions. *arXiv preprint arXiv:2306.16431*.

Miller, T. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267: 1–38.

Mitsuhara, M.; Fukui, H.; Sakashita, Y.; Ogata, T.; Hirakawa, T.; Yamashita, T.; and Fujiyoshi, H. 2019. Embedding Human Knowledge into Deep Neural Network via Attention Map. *arXiv preprint arXiv:1905.03540*.

Molnar, C. 2020. *Interpretable machine learning*. Lulu. com.

Nourani, M.; King, J.; and Ragan, E. 2020. The role of domain expertise in user trust and the impact of first impressions with intelligent systems. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing*, volume 8, 112–121.

Nourani, M.; Roy, C.; Block, J. E.; Honeycutt, D. R.; Rahman, T.; Ragan, E.; and Gogate, V. 2021. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *26th International Conference on Intelligent User Interfaces*, 340–350.

Nourani, M.; Roy, C.; Block, J. E.; Honeycutt, D. R.; Rahman, T.; Ragan, E. D.; and Gogate, V. 2022. On the importance of user backgrounds and impressions: Lessons learned from interactive AI applications. *ACM Transactions on Interactive Intelligent Systems*, 12(4): 1–29.

Plumb, G.; Ribeiro, M. T.; and Talwalkar, A. 2021. Finding and Fixing Spurious Patterns with Explanations. *arXiv preprint arXiv:2106.02112*.

Popordanoska, T.; Kumar, M.; and Teso, S. 2020. Machine Guides, Human Supervises: Interactive Learning with Global Explanations. *arXiv preprint arXiv:2009.09723*.

Rastogi, C.; et al. 2022. Deciding Fast and Slow: The Role of Cognitive Biases in AI-Assisted Decision-Making. *Proc. ACM Hum.-Comput. Interact.*

Ross, A. S.; Hughes, M. C.; and Doshi-Velez, F. 2017. Right for the right reasons: training differentiable models by constraining their explanations. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, 2662–2670.

Schramowski, P.; Stammer, W.; Teso, S.; Brugger, A.; Herbert, F.; Shao, X.; Luigs, H.-G.; Mahlein, A.-K.; and Kersting, K. 2020. Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nature Machine Intelligence*, 2(8): 476–486.

Schwalbe, G.; and Finzel, B. 2024. A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts. *Data Mining and Knowledge Discovery*, 38(5): 3043–3101.

Settles, B. 2011. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, 1467–1478.

Shao, X.; Skryagin, A.; Schramowski, P.; Stammer, W.; and Kersting, K. 2021. Right for Better Reasons: Training Differentiable Models by Constraining their Influence Function. In *Proceedings of Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI)*.

Sheeran, P.; and Webb, T. L. 2016. The intention–behavior gap. *Social and personality psychology compass*, 10(9): 503–518.

Slany, E.; Ott, Y.; Scheele, S.; Paulus, J.; and Schmid, U. 2022. CAIPI in Practice: Towards Explainable Interactive Medical Image Classification. In *IFIP International Conference on Artificial Intelligence Applications and Innovations*, 389–400. Springer.

Stammer, W.; Schramowski, P.; and Kersting, K. 2021. Right for the Right Concept: Revising Neuro-Symbolic Concepts by Interacting with their Explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3619–3629.

Teso, S. 2019. Toward Faithful Explanatory Active Learning with Self-explainable Neural Nets. In *Proceedings of the Workshop on Interactive Adaptive Learning (IAL 2019)*, 4–16.

Teso, S.; Alkan, Ö.; Stammer, W.; and Daly, E. 2023. Leveraging Explanations in Interactive Machine Learning: An Overview. *Frontiers in Artificial Intelligence*.

Teso, S.; Bontempelli, A.; Giunchiglia, F.; and Passerini, A. 2021. Interactive Label Cleaning with Example-based Explanations. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*.

Teso, S.; and Kersting, K. 2019. Explanatory interactive machine learning. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, 239–245.

Warren, G.; Byrne, R. M.; and Keane, M. T. 2024. Categorical and continuous features in counterfactual explanations of AI systems. *ACM Transactions on Interactive Intelligent Systems*, 14(4): 1–37.

Ye, W.; Zheng, G.; Cao, X.; Ma, Y.; and Zhang, A. 2024. Spurious correlations in machine learning: A survey. *arXiv preprint arXiv:2402.12715*.

Zylberajch, H.; Lertvittayakumjorn, P.; and Toni, F. 2021. HILDIF: Interactive Debugging of NLI Models Using Influence Functions. *Workshop on Interactive Learning for Natural Language Processing*, 1.