Automation Bias in Mammography: The Impact of Artificial Intelligence BI-RADS Suggestions on Reader Performance

Thomas Dratsch, MD* • Xue Chen, MD* • Mohammad Rezaatude Mehrizi, PhD • Roman Kloeckner, MD • Aline Mähringer-Kunz, MD • Michael Pülsken, MD • Bettina Baefsser, MD • Stephanie Sauer, MD • David Maintz, MD • Daniel Pinto dos Santos, MD

From the Institute of Diagnostic and Interventional Radiology, University of Cologne, Faculty of Medicine and University Hospital Cologne, Kerpener Str 62, 50937 Cologne, Germany (T.D., X.C., M.P., D.M., D.P.d.S.); School of Business and Economics, Knowledge, Information and Innovation, Vrije Universiteit Amsterdam, Amsterdam, the Netherlands (M.R.M.); Institute of Interventional Radiology, University Clinic Schleswig-Holstein, Kiel, Germany (R.K.); Department of Diagnostic and Interventional Radiology, University Medical Centre of the Johannes Gutenberg-University Mainz, Mainz, Germany (A.M.K.); and Institute of Diagnostic and Interventional Radiology, University Clinic Würzburg, Würzburg, Germany (B.B., S.S.). Received September 9, 2022; revision requested November 2; revision received January 25, 2023; accepted March 13. Address correspondence to T.D. (email: thomas.dratsch@uk-koeln.de).

Supported by a grant from the German Ministry of Health (BMG) as part of the project EVA-KI (ZMVI1-2520DAT03B).

To determine how automation bias can affect inexperienced, moderately experienced, and very experienced radiologists when reading mammograms with the aid of an artificial intelligence (AI) system.

In this prospective experiment, 27 radiologists read 50 mammograms and provided their Breast Imaging Reporting and Data System (BI-RADS) assessment assisted by a purported AI system. Mammograms were obtained between January 2017 and December 2019 and were presented in two randomized sets. The first was a training set of 10 mammograms, with the correct BI-RADS category suggested by the AI system. The second was a set of 40 mammograms in which an incorrect BI-RADS category was suggested for 12 mammograms. Reader performance, degree of bias in BI-RADS scoring, perceived accuracy of the AI system, and reader confidence in their own BI-RADS ratings were assessed using analysis of variance (ANOVA) and repeated-measures ANOVA followed by post hoc tests and Kruskal-Wallis tests followed by the Dunn post hoc test.

The percentage of correctly rated mammograms by inexperienced (mean, 79.7% ± 11.7 [SD] vs 19.8% ± 14.0; P < .001; r = 0.93), moderately experienced (mean, 81.3% ± 10.1 vs 24.8% ± 11.6; P < .001; r = 0.96), and very experienced (mean, 82.3% ± 4.2 vs 45.5% ± 9.1; P = .003; r = 0.97) radiologists was significantly impacted by the correctness of the AI prediction of BI-RADS category. Inexperienced radiologists were significantly more likely to follow the suggestions of the purported AI when it incorrectly suggested a higher BI-RADS category than the actual ground truth compared with both moderately (mean degree of bias, 4.0 ± 1.8 vs 2.4 ± 1.5; P = .044; r = 0.46) and very (mean degree of bias, 4.0 ± 1.8 vs 1.2 ± 0.8; P = .009; r = 0.65) experienced readers.

The results show that inexperienced, moderately experienced, and very experienced radiologists reading mammograms are prone to automation bias when being supported by an AI-based system. This and other effects of human and machine interaction must be considered to ensure safe deployment and accurate diagnostic performance when combining human readers and AI.

© RSNA, 2023

Supplemental material is available for this article.

Artificial intelligence (AI) and machine learning–based algorithms are being increasingly used by radiologists to enhance diagnostic accuracy (1). When using these AI tools, radiologists independently examine an image but also consider the AI suggestion, assess it, and potentially modify their initial judgment accordingly. In an ideal scenario, radiologists correctly integrate the information provided by AI, benefiting from cases in which the AI provides a better suggestion and ignoring cases in which AI makes an incorrect prediction, and produce better diagnostic reports than they would without AI.

Several studies have shown a synergistic combination of radiologists and AI is possible (2,3). However, there is also the danger that radiologists may stop critically engaging with the AI results and start mindlessly following them. This overreliance on a decision support system, known as automation bias, has been observed in a wide range of fields, such as aviation, engineering, and medicine, and a long line of research has focused on the conditions that cause automation bias to arise (4). Several AI systems are commercially available (5); however, it remains a question whether a successful integration of AI into the radiologic workflow is possible. Additionally, the repetitive nature and highly standardized workflow of mammography screening paired with decision support systems may overtake some of the radiologist’s interpretative work. In this setting, automation bias may be more likely. Several studies have shown that introduction of computer-aided detection into the...
mammography workflow impairs the performance of radiologists (6,7). However, there are no studies showing the influence of AI-based systems on the performance of accurate mammogram readings by radiologists (8). We hypothesize that incorrect predictions by AI may have similar negative effects on reader performance.

In the case of mammography, two types of errors are relevant to automation bias: errors of commission (incorrectly diagnosing a normal mammogram as containing a malignancy) and errors of omission (incorrectly diagnosing a mammogram with malignancy as normal). Both types of errors can be the result of automation bias when radiologists using a decision support system do not critically engage with the decisions of the system and follow the suggestions even when they are incorrect. Thus, our goal was to determine how automation bias can affect inexperienced, moderately experienced, and very experienced radiologists when reading mammograms aided by an AI system.

Materials and Methods

Study Design and Participants

This study was approved by the University Hospital of Cologne ethics committee and adheres to the criteria of the ethics committee of University Hospital of Cologne (no. 20–1322). Written informed consent was given by all participants before study inclusion.

Twenty-seven radiologists from University Hospital Cologne, University Clinic Schleswig-Holstein, and University Clinic Würzburg (Germany)—11 inexperienced, 11 moderately experienced, and five very experienced in reading mammograms—participated in this prospective study (Table 1). Inexperienced participants had never read mammograms before or were starting their first mammography rotation and had fewer than 2 months of experience reading mammograms. Moderately experienced participants had an average of 13.0 months (1.1 years) of experience reading mammograms. Very experienced participants had an average of 129.6 months (10.8 years) of experience reading mammograms. Participation was voluntary, and data were anonymized.

Fifty mammograms were randomly selected from our institution's picture archiving and communication system. The search time frame was restricted to January 2017 to December 2019. Stratified random selection was performed to achieve the desired distribution of cases, as depicted in Figure 1. All mammograms had been interpreted by experienced readers. The ground truth for all malignant cases was established through biopsies. Benign cases were stable over a period of at least 4 years. Each mammogram included a craniocaudal and mediolateral oblique view for each breast, for a total of four images per patient. A table showing quantitative information for all Breast Imaging Reporting and Data Systems (BI-RADS) 3–5 lesions in the study can be found in Table S1.

In 38 of the 50 mammograms (76%), the AI-based system suggested the correct diagnosis. In 12 of the 50 mammograms (24%), the purported AI system suggested an incorrect diagnosis for one of the two breasts—suggesting a higher or lower BI-RADS category than the actual diagnosis.

Purported AI-Based System

For each mammogram, an experienced reader used the actual reports to draw a heat map highlighting the lesions in question in red (Fig 2). For cases in which the purported AI falsely reported a pathologic finding, plausible areas in the image that could be mistaken for a suspicious lesion were selected. For the cases in which the purported AI falsely suggested a pathologic finding was benign, green heat maps were generated to suggest a benign finding. The original diagnoses of all 50 mammograms as well as the heat maps were then reviewed by another independent reader to ensure their accuracy and plausibility. All mammograms used in the study with or without heat maps are available in the supplement (Appendix S1).

We used C++ and Qt Creator (version 4.0; https://www.qt.io/) to create the interface of the purported AI-based diagnostic system. This system resembled an actual AI-based system but was in fact a program to display the mammograms alongside the predictions and heat maps we created in advance (Fig

---

**Table 1: Reader Demographic Information**

<table>
<thead>
<tr>
<th>Characteristic</th>
<th>Inexperienced Readers (n = 11)</th>
<th>Moderately Experienced Readers (n = 11)</th>
<th>Very Experienced Readers (n = 5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age (y)</td>
<td>29.7 ± 2.1</td>
<td>32.6 ± 1.5</td>
<td>42.0 ± 9.8</td>
</tr>
<tr>
<td>Sex (male)*</td>
<td>7 (64)</td>
<td>8 (73)</td>
<td>0 (0)</td>
</tr>
<tr>
<td>Experience reading mammograms (mo)</td>
<td>0.5 ± 0.7</td>
<td>13.0 ± 9.4</td>
<td>129.6 ± 70.7</td>
</tr>
<tr>
<td>Experience in radiology (y)</td>
<td>2.8 ± 1.4</td>
<td>5.4 ± 1.5</td>
<td>15.2 ± 8.5</td>
</tr>
</tbody>
</table>

Note.—Unless otherwise indicated, data are means ± SDs.

* Data are numbers of individuals, with percentages in parentheses.
2). Because it was only a program to display the images and heat maps, no actual training took place. The system was also designed to track the time to read each mammogram. The code for the purported AI-based system is available at https://github.com/DrXCHEN/Automation-Bias.

Image Evaluation Procedure
The experiment was conducted using a dedicated Digital Imaging and Communications in Medicine and picture archiving and communication system workstation with a monitor and software in compliance with requirements for reviewing mammograms. The purported AI-based system was displayed on a separate monitor. At the beginning of the study, all participants were informed that they would be evaluating a new AI-based system by reading 50 cases with the support of the system. The 50 cases included mammograms that ranged from BI-RADS 2 (benign finding) to BI-RADS 5 (highly suspicious finding). Inexperienced participants watched a short video presentation on the basics of mammography screening and BI-RADS classification. Then all participants proceeded to read the 50 images supported by the purported AI in one session.

To not bias participants’ decisions by the BI-RADS rating of the other breast, we only selected mammograms where the BI-RADS rating of the other breast was BI-RADS 2. The side of the BI-RADS lesion in question was randomized. The order of the 50 mammograms was also randomized for each participant. At the beginning of the study, participants were presented with a randomized block of 10 mammograms (Fig 1, stage 1) in which the purported AI was always correct to establish the credibility of the AI. This was immediately followed by another randomized block of 40 mammograms, consisting of the 12 mammograms in which the purported AI suggested an incorrect diagnosis for one of the two breasts and 28 mammograms in which the purported AI suggested the correct diagnosis (Fig 1, stage 2).

At the start of each new case, participants were shown the predicted BI-RADS categories for both breasts and the heat maps. Participants could toggle the heat maps on and off to inspect the images more closely. After participants had reviewed all 50 images, they completed a short questionnaire with demographic questions and rated the overall accuracy of the purported AI on a 10-point Likert scale ranging from 1 (not very accurate) to 10 (very accurate) and their overall confidence in their own BI-RADS ratings on a seven-point Likert scale ranging from 1 (not very confident) to 7 (very confident).

Outcome Measures
We evaluated the percentage of correctly rated mammograms for each participant and recorded the time to read each mammogram. Reader scores for perceived accuracy of the purported AI-based system and confidence in their own BI-RADS ratings were also evaluated.

In 12 of the 50 mammograms, the purported AI suggested an incorrect diagnosis for one of the two breasts, suggesting a higher ($n = 6$) or lower ($n = 6$) BI-RADS category than the actual diagnosis. For each case, degree of bias was measured as the difference between the actual BI-RADS categories and the
incorrect AI suggestions. For instance, a bias score of 2 would be awarded for a case if the participant followed the prediction of the AI and incorrectly selected BI-RADS 4 instead of correctly selecting BI-RADS 2. A bias score of 1 was awarded for that trial if a participant partially followed the AI and incorrectly selected BI-RADS 3 for that mammogram instead of selecting BI-RADS 2. Separate bias scores were calculated for the six cases in which the AI suggested a higher BI-RADS category (bias toward over-diagnosis) and the six cases in which the AI suggested a lower BI-RADS category (bias toward under-diagnosis) by adding up the scores for all individual cases in each category. For two of the six cases, AI suggested a BI-RADS category two steps higher or lower than the actual BI-RADS category (eg, BI-RADS 4 instead of BI-RADS 2), and for four of the six cases, AI suggested a BI-RADS category one step higher or lower than the actual BI-RADS category (eg, BI-RADS 3 instead of BI-RADS 2). Thus, total bias scores for the six trials could range between 0 (participants never followed the incorrect suggestions of the AI) and 8 (participants always followed the incorrect suggestions of the AI: $2 \times 2 + 4 \times 1 = 8$).

**Statistical Analyses**

All statistical analyses were performed by one author (T.D.) using IBM SPSS Statistics (version 26.0; IBM). Normality of the data was assessed using the Kolmogorov-Smirnov test. Homogeneity of variances was assessed with the Levene test. Normally distributed variables are reported as means ± SDs, whereas nonparametric data are reported as medians and interquartile ranges. To analyze the effect of experience and time course on the time to read the mammograms, we performed a 3 × 5 repeated-measures analysis of variance (ANOVA) with level of experience (inexperienced vs moderately experienced vs very experienced) and time course (mammograms 1–10 vs 11–20 vs 21–30 vs 31–40 vs 41–50) as independent variables and time to read each mammogram as the dependent variable. To analyze the effect of experience and the correctness of the AI predictions on participant performance, we performed a 2 × 2 repeated-measures ANOVA with level of experience (inexperienced vs experienced) and correctness of AI predictions (correct vs incorrect prediction) as independent variables and percentage of correctly rated mammograms as the dependent variable. Errors of omission, errors of commission, and overall time to read all mammograms were analyzed by using one-way ANOVA. Perceived accuracy of the purported AI-based system and reader confidence in their own BI-RADS ratings were analyzed using the Kruskal-Wallis test. ANOVAs were followed by post hoc tests, and Kruskal-Wallis tests were followed by Dunn post hoc tests. The Bonferroni-Holm correction was used to correct for multiple comparisons (9). Two-sided $P < .05$ was considered indicative of a significant difference. Pearson $r$ value was used as the effect size with $r$ values of 0.10, 0.30, and 0.50 constituting small, medium, and large effects, respectively (10). Because of the novel nature of the research and the lack of prior effect sizes for comparison, no power analysis was performed. One mammogram was excluded from the final analysis because it was revealed during analysis that the sides of the lesions in the original report had been switched, which resulted in a mismatch between the AI predictions presented along with the mammogram and the actual BI-RADS categories.
Performance and Automation Bias

There was a significant main effect of correctness of AI predictions \((P < .001, r = 0.92)\), indicating that participants rated more mammograms incorrectly when the AI predictions were also incorrect (inexperienced readers: mean, 79.7\% ± 11.7 vs 19.8\% ± 14.0; \(P < .001\); \(r = 0.93\); moderately experienced readers: mean, 81.3\% ± 10.1 vs 24.8\% ± 11.6; \(P < .001\); \(r = 0.96\); very experienced readers: mean, 82.3\% ± 4.2 vs 45.5\% ± 9.1; \(P = .003\); \(r = 0.97\) (Fig 4). There was also a significant main effect for experience \((P < .001, r = 0.72)\), indicating that the three groups of readers differed regarding the percentage of correctly rated mammograms. As Table 3 shows, very experienced readers rated significantly more mammograms correctly when the AI predictions were incorrect compared with moderately experienced and inexperienced readers. The three groups of readers did not differ significantly in the percentage of correctly rated mammograms when the AI predictions were correct (Table 3). The interaction effect between the level of experience and correctness of AI predictions was not significant \((P = .11, r = 0.16)\).

Automation Bias: Errors of Commission and Errors of Omission

For errors of commission, there was a significant main effect for experience \((P = .01, r = 0.34)\), indicating that the three groups of readers differed in their propensity to follow the purported AI when it falsely suggested a higher BI-RADS category. As Table 4 and Figure 5 show, inexperienced readers were more likely to follow the purported AI when it falsely suggested a higher BI-RADS category compared with moderately and very experienced readers. There was no significant difference between moderately and very experienced readers.

For errors of omission, there was no significant main effect for experience \((P = .16, r = 0.14)\), indicating that the three groups of readers did not differ in their propensity to follow the purported AI when it falsely suggested a lower BI-RADS category (Table 4).

Perceived Accuracy of the Purported AI-based System and Confidence in Own BI-RADS Ratings

For the perceived accuracy of the purported AI, there was no significant main effect of experience \((P = .22, r = 0.21)\), indicating that the three groups of readers did not differ in their perceived accuracy of the purported AI (Table 5).

For confidence in a reader’s own BI-RADS ratings, there was a significant main effect of experience \((P < .001, r = 0.82)\), indicating that the three groups of readers differed in confidence in their own BI-RADS ratings. As Table 5 shows, inexperienced readers were less confident in their own BI-RADS ratings compared with moderately and very experienced readers.

Discussion

There is concern that radiologists might become overly reliant on artificial intelligence (AI) systems and might start...
Automation Bias in Mammography

Radiology: Volume 000: Number 0—Month 2023

To mindlessly follow their output—a tendency known as automation bias. To ensure safe deployment of these AI systems, it is important to study their effect on radiologists’ performance. Thus, our goal was to explore the degree to which automation bias can affect radiologists when reading mammograms. Our results show that an AI-based system introduced automation bias and decreased reader performance. Inexperienced (mean, 79.7% ± 11.7 vs 19.8% ± 14.0; \(P < .001; r = 0.93\)), moderately experienced (mean, 81.3% ± 10.1 vs 24.8% ± 11.6; \(P < .001; r = 0.96\)), and very experienced (mean, 82.3% ± 4.2 vs 45.5% ± 9.1; \(P = .003; r = 0.97\)) radiologists were worse at assigning the correct Breast Imaging Reporting and Data System (BI-RADS) scores for cases in which the purported AI suggested an incorrect BI-RADS category. These results suggest that all radiologists, regardless of expertise, can be subject to automation bias. However, very experienced readers rated significantly more mammograms correctly when the AI predictions were incorrect compared with both moderately experienced and inexperienced readers.

Table 2: Average Reading Time for One Mammogram for Each Batch of 10 Mammograms for Inexperienced, Moderately Experienced, and Very Experienced Radiologists

<table>
<thead>
<tr>
<th>Trial No. and Comparison</th>
<th>Inexperienced Readers ((n = 11))</th>
<th>Moderately Experienced Readers ((n = 11))</th>
<th>Very Experienced Readers ((n = 5))</th>
<th>Mean Difference</th>
<th>(P) Value</th>
<th>(r) Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>1–10</td>
<td>68.6 ± 21.6</td>
<td>67.9 ± 21.4</td>
<td>NA</td>
<td>0.7</td>
<td>.949</td>
<td>0.02</td>
</tr>
<tr>
<td>Inexperienced vs moderately experienced readers</td>
<td>68.6 ± 21.6</td>
<td>NA</td>
<td>118.6 ± 29.5</td>
<td>50.0</td>
<td>.009</td>
<td>0.73</td>
</tr>
<tr>
<td>Moderately experienced vs very experienced readers</td>
<td>NA</td>
<td>67.9 ± 21.4</td>
<td>118.6 ± 29.5</td>
<td>50.7</td>
<td>.008</td>
<td>0.73</td>
</tr>
<tr>
<td>11–20</td>
<td>42.1 ± 17.3</td>
<td>56.1 ± 11.7</td>
<td>NA</td>
<td>14.0</td>
<td>.441</td>
<td>0.45</td>
</tr>
<tr>
<td>Inexperienced vs moderately experienced readers</td>
<td>42.1 ± 17.3</td>
<td>NA</td>
<td>86.1 ± 24.7</td>
<td>44.0</td>
<td>.002</td>
<td>0.75</td>
</tr>
<tr>
<td>Moderately experienced vs very experienced readers</td>
<td>NA</td>
<td>56.1 ± 11.7</td>
<td>86.1 ± 24.7</td>
<td>30.0</td>
<td>.33</td>
<td>0.67</td>
</tr>
<tr>
<td>21–30</td>
<td>38.8 ± 19.5</td>
<td>49.5 ± 15.4</td>
<td>NA</td>
<td>10.7</td>
<td>.461</td>
<td>0.31</td>
</tr>
<tr>
<td>Inexperienced vs moderately experienced readers</td>
<td>38.8 ± 19.5</td>
<td>NA</td>
<td>70.5 ± 14.0</td>
<td>31.7</td>
<td>.025</td>
<td>0.68</td>
</tr>
<tr>
<td>Moderately experienced vs very experienced readers</td>
<td>NA</td>
<td>49.5 ± 15.4</td>
<td>70.5 ± 14.0</td>
<td>21.0</td>
<td>.200</td>
<td>0.58</td>
</tr>
<tr>
<td>31–40</td>
<td>32.6 ± 10.1</td>
<td>43.4 ± 19.0</td>
<td>NA</td>
<td>10.8</td>
<td>.461</td>
<td>0.35</td>
</tr>
<tr>
<td>Inexperienced vs moderately experienced readers</td>
<td>32.6 ± 10.1</td>
<td>NA</td>
<td>77.7 ± 17.0</td>
<td>45.1</td>
<td>.001</td>
<td>0.88</td>
</tr>
<tr>
<td>Moderately experienced vs very experienced readers</td>
<td>NA</td>
<td>43.4 ± 19.0</td>
<td>77.7 ± 17.0</td>
<td>34.3</td>
<td>.008</td>
<td>0.69</td>
</tr>
<tr>
<td>41–50</td>
<td>30.3 ± 9.8</td>
<td>41.6 ± 17.2</td>
<td>NA</td>
<td>11.3</td>
<td>.461</td>
<td>0.39</td>
</tr>
<tr>
<td>Inexperienced vs moderately experienced readers</td>
<td>30.3 ± 9.8</td>
<td>NA</td>
<td>67.2 ± 17.6</td>
<td>36.9</td>
<td>.003</td>
<td>0.83</td>
</tr>
<tr>
<td>Moderately experienced vs very experienced readers</td>
<td>NA</td>
<td>41.6 ± 17.2</td>
<td>67.2 ± 17.6</td>
<td>25.6</td>
<td>.034</td>
<td>0.60</td>
</tr>
</tbody>
</table>

Note.—Unless otherwise indicated, data are mean reading times (in seconds) ± SDs. NA = not applicable.

Figure 4: Bar graph shows the mean percentage of mammogram readers assigned the correct Breast Imaging Reporting and Data System (BI-RADS) score stratified by correctness of artificial intelligence (AI) predictions and level of experience (inexperienced, \(n = 11\); moderately experienced, \(n = 11\); very experienced, \(n = 5\)). Error bars represent 95% CIs. Circles, triangles, and squares represent individual data points.
inexperienced readers, indicating that experience may make readers more resilient toward the negative effect of incorrect AI suggestions.

When further analyzing the direction of bias, all three groups were equally likely to follow the suggestions of the purported AI when it incorrectly suggested a lower BI-RADS category compared with the actual ground truth (errors of omission) (inexperienced vs moderately experienced readers: mean degree of bias, 6.3 ± 1.4 vs 6.3 ± 1.3; \( P > .999; r = 0.00 \)) (inexperienced vs very experienced readers: mean degree of bias, 6.3 ± 1.4 vs 5.0 ± 0.7; \( P = .243; r = 0.45 \)) (moderately vs very experienced readers: mean degree of bias, 6.3 ± 1.3 vs 5.0 ± 0.7; \( P = .243; r = 0.47 \)). However, inexperienced radiologists were more likely to follow the suggestions of the purported AI when it incorrectly suggested a higher BI-RADS category (errors of commission) than the actual ground truth compared with both moderately (mean degree of bias, 4.0 ± 1.8 vs 2.4 ± 1.5; \( P = .044; r = 0.46 \)) and very experienced (mean degree of bias, 4.0 ± 1.8 vs 1.2 ± 0.8; \( P = .009; r = 0.65 \)) readers.

As previous research has shown, inexperienced radiologists are more likely than experienced radiologists to produce false-positive findings (11). Additionally, errors of commission can be the result of overconfidence in the capabilities of decision support systems and a lack of confidence in one’s own abilities (12,13). We did not find a significant difference between the three groups of radiologists regarding the perceived accuracy of the purported AI. However, inexperienced radiologists were less confident in their own BI-RADS ratings compared with moderately and very experienced readers, which may potentially make them more vulnerable to following incorrect suggestions by AI.

Overall, there was no significant difference in performance between the three groups for the cases in which the purported AI suggested the correct BI-RADS category. This suggests that inexperienced radiologists could potentially use the predictions of AI-based systems to perform on a level similar to that of experienced radiologists. This is in line with earlier research showing that inexperienced users benefit the most from computer-based...
decision support systems (14). While an AI-based system may help novices perform on a level similar to that of more experienced radiologists, introducing AI-based support early during radiology training may pose the risk of deskilling such that certain skills are lost or not properly learned in the first place and cannot be executed without the help of AI (15).

We found that the interpretation time significantly decreased over the course of the study for all readers. This could be possible evidence of automation complacency; however, it is also equally possible that readers became more comfortable with the system and therefore increased their speed when reading the mammograms. Additionally, the very experienced readers took significantly longer to read the mammograms compared with the inexperienced (mean time, 78.8 minutes ± 24.6 vs 36.3 minutes ± 9.0; \( P = .001; r = 0.81 \)) and moderately experienced (mean time, 78.8 minutes ± 24.6 vs 43.9 minutes ± 11.7; \( P = .003; r = 0.73 \)) readers. One possible explanation for this effect is that the very experienced readers engaged more critically with the suggestions of the AI-based system by comparing their own ratings with the output of the AI, resulting in longer reading times per mammogram. However, this is only speculation and should be explored in future studies.

Overall, our results show that automation bias can affect the performance of radiologists regardless of their experience. Therefore, specific strategies should be considered to mitigate automation bias. Previous research has shown that presenting users with the confidence levels of the decision support system can help reduce automation bias by keeping users more critically engaged with the output of the system (16).

In the case of an AI-based system, this could be implemented by displaying the probability of each output. Another strategy to mitigate automation bias involves educating users about the reasoning process of the decision support system (17). For AI-based solutions, this means focusing on explainable AI solutions that allow inferences about how a certain judgment by the system is generated. Lastly, ensuring that the users of a decision support system feel accountable for their own decisions can also help decrease automation bias (18).

Our study had several limitations. First, we did not investigate the performance of radiologists without the use of the purported AI-based system, which should be considered in future studies to better judge the effect of AI-based systems on performance and their implementation. Additionally, we did not investigate the effect of other sources of information on readers’ performance. For instance, Gaube et al (19) compared the effect of decision support in chest radiographs and found that readers were negatively influenced by suggestions coming from purported AI or other human readers. Second, we only focused on mammograms that

---

**Table 5: Accuracy of AI and Confidence in BI-RADS Ratings**

<table>
<thead>
<tr>
<th>Statistic and Comparison</th>
<th>Inexperienced Readers (n = 11)</th>
<th>Moderately Experienced Readers (n = 11)</th>
<th>Very Experienced Readers (n = 5)</th>
<th>P Value</th>
<th>r Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accuracy</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inexperienced vs moderately experienced readers</td>
<td>9.0 (7.0–9.0)</td>
<td>8.0 (7.0–9.0)</td>
<td>NA</td>
<td>.432</td>
<td>0.31</td>
</tr>
<tr>
<td>Inexperienced vs very experienced readers</td>
<td>9.0 (7.0–9.0)</td>
<td>NA</td>
<td>8.0 (4.5–8.5)</td>
<td>.432</td>
<td>0.37</td>
</tr>
<tr>
<td>Moderately experienced vs very experienced readers</td>
<td>NA</td>
<td>8.0 (7.0–9.0)</td>
<td>8.0 (4.5–8.5)</td>
<td>.750</td>
<td>0.08</td>
</tr>
<tr>
<td>Confidence</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inexperienced vs moderately experienced readers</td>
<td>2.0 (1.0–2.0)</td>
<td>5.0 (5.0–6.0)</td>
<td>NA</td>
<td>.001</td>
<td>0.79</td>
</tr>
<tr>
<td>Inexperienced vs very experienced readers</td>
<td>2.0 (1.0–2.0)</td>
<td>NA</td>
<td>6.0 (5.0–6.0)</td>
<td>.001</td>
<td>0.85</td>
</tr>
<tr>
<td>Moderately experienced vs very experienced readers</td>
<td>NA</td>
<td>5.0 (5.0–6.0)</td>
<td>6.0 (5.0–6.0)</td>
<td>.646</td>
<td>0.12</td>
</tr>
</tbody>
</table>

Note.—Data are median overall accuracy of artificial intelligence (AI) on a 10-point Likert scale and median overall reader confidence in their own Breast Imaging Reporting and Data System (BI-RADS) ratings on a seven-point Likert scale; data in parentheses are interquartile ranges. NA = not applicable.
were categorized with a standardized BI-RADS system. Future studies should also investigate the effect of AI-based systems on the creation of complex narrative reports. Third, our sample size was small.

In conclusion, we show that all radiologists, regardless of experience reading mammograms, were prone to automation bias. This and other effects of human-machine interaction must be considered to ensure safe deployment and accurate diagnostic performance when combining human readers and AI.

**Author contributions:** Guarantors of integrity of entire study, M.P., D.P.d.S.; study concepts/study design or data acquisition or data analysis/interpretation, all authors; manuscript drafting or manuscript revision for important intellectual content, all authors; approval of final version of submitted manuscript, all authors; agree to ensure any questions related to the work are appropriately resolved, all authors; literature research, T.D., X.C., M.P., B.B., D.P.d.S.; clinical studies, X.C., M.P., S.S., D.M.; experimental studies, T.D., X.C., M.R.M., R.K., A.M.K., M.P., B.B., S.S., D.P.d.S.; statistical analysis, T.D., X.C., M.P., D.P.d.S.; and manuscript editing, T.D., X.C., R.K., A.M.K., M.P., B.B., D.M., D.P.d.S.

**Disclosures of conflicts of interest:** T.D. Grant from German Federal Ministry of Health. X.C. No relevant relationships. M.R.M. Patent held in association with Vrij University Amsterdam. R.K. Consulting fees from Boston Scientific, Bristol-Myers Squibb, Guerbet, Roche, and SIRTEX; payment or honoraria from BTG, EISAI, Honorarium from Bayer Vital; founder and CEO of Lernrad. B.B. Honorarium from Bayer Vital; founder and CEO of Lernrad. S.S. No relevant relationships. D.M. Honorarium from Phillips. D.P.d.S. Grant from the German Federal Ministry of Health; consulting fees from Cook Medical; honorarium from Bayer; unpaid roles as Chair of the IT subcommittee of the German Radiological Society and Vice President of the European Society for Medical Imaging Informatics.

**References**