Published in partnership with Seoul National University Bundang Hospital

6

https://doi.org/10.1038/s41746-024-01192-8

Expert gaze as a usability indicator of medical AI decision support systems: a preliminary study

Check for updates

Nora Castner **1**^{1,2,7} , Lubaina Arsiwala-Scheppach^{3,7}, Sarah Mertens³, Joachim Krois **3**, Enkeleda Thaqi⁴, Enkelejda Kasneci⁴, Siegfried Wahl **1**^{1,5} & Falk Schwendicke⁶

Given the current state of medical artificial intelligence (AI) and perceptions towards it, collaborative systems are becoming the preferred choice for clinical workflows. This work aims to address expert interaction with medical AI support systems to gain insight towards how these systems can be better designed with the user in mind. As eye tracking metrics have been shown to be robust indicators of usability, we employ them for evaluating the usability and user interaction with medical AI support systems. We use expert gaze to assess experts' interaction with an AI software for caries detection in bitewing x-ray images. We compared standard viewing of bitewing images without AI support versus viewing where AI support could be freely toggled on and off. We found that experts turned the AI on for roughly 25% of the total inspection task, and generally turned it on halfway through the course of the inspection. Gaze behavior showed that when supported by AI, more attention was dedicated to user interface elements related to the AI support, with more frequent transitions from the image itself to these elements. When considering that expert visual strategy is already optimized for fast and effective image inspection, such interruptions in attention can lead to increased time needed for the overall assessment. Gaze analysis provided valuable insights into an AI's usability for medical image inspection. Further analyses of these tools and how to delineate metrical measures of usability should be developed.

Even in an era where artificial intelligence (AI) is becoming more pervasive in the workforce, human experts are still sought-after for the final decision. In a wide range of fields, such "collaborative" AI-human systems are the preferred constellation: In education^{1,2}, marketing^{3,4}, semi-autonomous vehicles⁵, and ever-increasing in healthcare^{6–8}. AI serves as support or assistant systems, while any final decision is taken by a human, fulfilling the demand for human oversight and autonomy⁹.

AI decision support systems for medical image interpretation – e.g., inspecting x-rays or volumetric scans – have been shown to improve diagnostic accuracy¹⁰⁻¹². Collaborative AI that supports radiologist inspection has led to faster and more accurate diagnoses compared with experts or AI alone^{11,13-16}. However, other studies show mixed results regarding performance and allude to concerns that AI can hinder expert diagnoses or offer little support¹⁷⁻²¹. These concerns can transfer to patients, who feel more assured when they know the expert has made the final diagnosis after

seeking AI support^{22,23}. Medical experts have also expressed apprehension using AI, regarding concerns of liability, trust and understanding, and reliability^{24–27}. These concerns solidify AI as an assistant tool and not autonomously making decisions.

Though a wealth of peer-reviewed publications highlight the potential for medical AI support systems, very few systems are successfully adopted into daily clinical workflows (See overviews in refs. 6,28–30). There are a number of factors that are holding back its integration into clinical environments, like regulatory hurdles³¹, unclear diagnostics, efficacy, and usefulness^{6,26}. Others suggest that critical limitations affecting integration of these systems are unsatisfactory experience and interoperability difficulties^{32–34}. Given interaction problems between the human user and the AI system, trust in AI systems can be greatly impacted^{35–37}.

For collaborative AI systems, ease of interaction is essential. So far, little effort has been directed towards improving the interaction between human

¹Carl Zeiss Vision International GmbH, Tübingen, Germany. ²University of Tübingen, Tübingen, Germany. ³Charité – Univesitätsmedizin, Oral Diagnostics, Digital Health and Services Research, Berlin, Germany. ⁴Technical University of Munich, Human–Centered Technologies for Learning, Munich, Germany. ⁵Institute for Ophthalmic Research, University of Tübingen, Tübingen, Germany. ⁶Ludwig Maximilian University, Operative, Preventative and Pediatric Dentistry and Period-ontology, Munich, Germany. ⁷These authors contributed equally: Nora Castner, Lubaina Arsiwala-Scheppach. 🖂 e-mail: nora.castner@zeiss.com

expert and system. When AI interaction is uncomfortable for the user, trust in the system is diminished regarding the system's performance and practicality³⁸. Currently, the majority of research on AI medical support has focused on performance metrics, while ignoring practicality and interaction between the user and the system. Better interaction may improve usability and, subsequently, integration in professional workflows. If AI interactions are not deemed *useful*, these devices could be pushed aside in favor of human only workflows³⁴. One approach to assessing interaction is expert gaze behavior.

Efficient and thorough inspection of medical images leads to faster feature recognition and better clinical reasoning^{39–41}. The visual strategies of medical professionals are an interplay of heightened sensitivity to certain features or structures and prior knowledge, i.e., experience and case based. This context dependent gaze behavior is known as the scanpath, consisting of fixations (attentional information) and saccades (transitions between attentional areas)^{42,43}.

Much of the previous literature in medical expert gaze analysis has focused on comparing experts and novices, highlighting the faster and more accurate perceptual ability of experts^{44,45}: With shorter time to first fixation on relevant areas (e.g., an anomaly) than novices^{40,46–52}. Experts also have more fixations and fixations of longer duration on relevant instead of irrelevant areas, which can be attributed to reducing extraneous attentional processing^{44,53–56}. However, image content affects expert eye movements^{40,44,57,58}. Obvious and easy to spot anomalies do not require as many fixations than harder to detect anomalies^{40,58–63}. For further descriptions of gaze in the context of medical image inspection, see refs. 39,64,65.

Expert gaze and scanpath behavior have been investigated in the context of dental images, and depend on image type⁶⁶. For panoramic radiographs, image coverage is key; outer and inner structural areas are assessed quickly, then thoroughly in a global to local search strategy^{67,68}. For periapical radiographs, tooth-by-tooth and circular search strategies are preferred, depending on the nature of anomalies present⁶³. Recently, this systematic tooth-by-tooth scanning strategy was also found when experts inspected bitewings for caries, which promoted faster recognition⁶⁹. When anomalies are harder to detect, experts' pupillary response indicates that their cognitive load adjusts to accommodate the information level⁷⁰. This adaptability highlights experts' effective information processing abilities. Whether this behavior can also accommodate information presented by a decision support system has yet to be explored.

Eye tracking has been used to evaluate the usability of systems in research fields such as marketing, software testing, and product design^{71,72}. From an interaction perspective, eye tracking can address not only the *how* (e.g., how do they navigate the interface), but also the *why* (e.g., why is the image inspected in this way)^{73,74}. Metrics such as fixation behavior and scanpath transitions and length related to interface elements can represent a user's attention or understanding of taskflows^{75,76}. These metrics have also been shown to correlate with usability reports^{77,78}. Pupil diameter changes as an indicator of cognitive load can also indirectly assess usability⁷⁹. Eye movement patterns can also indicate specific usability concerns, such as inconsistencies in design, architecture, and formatting⁷⁴. This information can improve accessibility⁸⁰, content highlighting⁸¹, and even realtime attention guiding⁶².

Eye tracking metrics have also been used as an indicator for usability of medical technologies^{83,84}, for example, to assess the usability of intensive care ventilators⁸⁵ or prosthetic arms⁸⁶. Other studies have leveraged eye tracking findings to better design interfaces based on patient or clinician needs^{86–88}. See Asan et al.⁸⁹ for more literature on medical interface design from gaze research. Concerning AI decision support systems, eye tracking is a powerful tool in addressing system improvements. An array of research has used expert gaze to assist AI models with region segmentation and labeling^{90–92}. These improvements can transform information visualization, which has shown to improve diagnostic performance^{93,94}. In summary, eye tracking measures offer insight into system usability and can be collected during the task. However, these metrics have not yet been used to evaluate how experts integrate AI support into their own clinical decision-making strategies or how AI support could potentially interrupt their workflows.

The aim of the present study was to evaluate how experts interact with an AI-based decision support tool to investigate dental bitewings, i.e. radiographs used for detecting caries. We use gaze behavior analysis via eye tracking as a non-invasive, naturalistic, and objective measure of interaction. Eye tracking measures have been shown to be robust indicators of usability, which research on medical AI systems has yet to fully utilize. Also, expert gaze behavior during medical image examination is well understood, offering a link between gaze features and cognitive processes. We employed gaze behavior as an indicator of the visual strategies related to clinical decision-making when using AI support versus not using AI support. We hypothesized that gaze behavior when using AI support will be different from gaze behavior without AI support. We also investigated how experts interact with the AI support system in the context of how they incorporate components of the system as well as control the system, and how their interaction changed over the experiment.

Results

Statistical analysis

As we are concerned about factors related to usability, we offer no analysis related to how experts look at bitewings in the realm of clinical decision-making. Instead, we analyzed visual strategies in the context of how experts employ an AI-support software that presents a bitewing plus informational content as part of an interface. We group the informational content under the category of user interface (UI) elements. These parts of the interface are depicted in Fig. 3a. As scanpath behavior specifically linked to medical image content was out of the scope of this current research, we restricted our analysis to transitional eye movements around the software, which was presented in a web browser.

We report only fixational metrics, as the eye tracker sampling rate is too low to fully understand saccadic behavior. In addition to the gaze behavior between dentists with and without AI support, we provide descriptions of how they interact with the AI system: i.e., their mouse interactions and gaze behavior differences when the AI overlay is turned off and on.

All variables exhibited non-normal distributions and thus were summarized using median and interquartile range (IQR) and were analyzed using non-parametric tests. Differences in each gaze metric between relevant groups were tested using the Wilcoxon rank sum test, where level of significance was set to p < 0.05. Missing data was not imputed. To account for any possible spatial offsets in the gaze data, defined areas of interest (AOIs) were given an extra pixel padding based on their relative pixel area: A pixel padding of 3 degrees of visual angle. For fixation behavior analysis, we counted fixations that land in overlapping AOI as a hit in both AOI. All statistical analyses and data management were performed using Python (version 3.8 and above). Table 1 reports the results of the statistics.

Time on task

We found that dentists took longer on the task when using AI support (109.19 s [78.91, 140.78]) than when not using AI support (87.1 s [73.37, 99.86]), while this difference was not significant given the wide spread of time taken (p = 0.0742). Figure 1 shows the distributions of task times for both conditions. We also observed a slight effect of stimulus viewing order with the plotted task time for each image (Fig. 2). For the first few images, there was a large difference in the inspection time between AI support and No AI support. This trend remains over the course of the experiment, but with slight inconsistencies.

Regarding the distribution of attention during the task, we normalized each participant's time viewing the bitewing and viewing UI elements over their total task time to get a percent proportion. The proportion of time spent looking at the UI vs. looking at the bitewing showed that with AI support, total time viewing UI elements was doubled from 4.94 % to 8.4 %.

Fixation behavior

Significantly more time was spent looking at the user interface when AI was present (5481.76 ms [3820.35, 6966.61]) than without (2348.39 ms [1071.496, 4963.75]: p = 0.0003). More time was spent viewing the bitewing with AI (54,393.53 ms [43,855.15, 84,542.02]) than without (49,139.55 ms

Table 1 | Summary statistics of fixation metrics based relative to the inspection task, attention to the bitewing and attention to user interface (UI) elements

	With AI support		Without AI support		
Variable	Med	IQR	Med	IQR	<i>p</i> -val
Task time [s]	109.19	[78.91, 140.78]	87.1	[73.37, 99.86]	0.0742
Total fixation duration UI [ms]	5481.76	[3820.35, 6966.61]	2348.39	[1071.496, 4963.75]	0.0003*
Total fixation duration bitewing [ms]	54393.53	[43855.15, 84542.02]	49139.55	[41812.34, 70003.19]	0.3168
Task average fixation duration [ms]	313.27	[243.597, 367.13]	317.29	[239.797, 349.86]	0.7660
Task fixation count	208.22	[182.6, 260.6]	182.70	[164.05, 212.39]	0.0548
UI fixation count	20.86	[18.83, 26.1]	10.63	[4.93, 15.65]	<0.001*
Bitewing fixation count	191.90	[161.8, 230.75]	176.55	[153.58, 204.25]	0.19
Task fixation frequency [fix/s]	2.21	[2.03, 2.33]	2.16	[1.98, 2.43]	0.8287

* indicates statistical significance at p < 0.05.



Fig. 1 | **Total time on task, in seconds, for viewing bitewings with AI support and without.** There are differing distributions between the two conditions, though their median task time is not significantly different.

[41,812.34, 70,003.19]), again without statistical significance, though (p = 0.3168). However, this behavior resulted in a higher number of total fixation counts with AI support (208.22 [182.6, 260.6]) compared to without AI support (182.70 [164.05, 212.39]: p = 0.05) Although, fixation count in the bitewing was not significantly different between AI (191.90 [161.8, 230.75]) and No AI (176.55 [153.58, 204.25] : p = 0.19), fixation count in the user interface (AI: 20.86 [18.83, 26.1], No AI: 10.63 [4.93, 15.65]) was significantly higher for the AI support condition (p < 0.001).

Fixation frequency remained unaffected by AI support, fixations per second with AI support (2.21 [2.03, 2.33]) and without AI support (2.16 [1.98, 2.43]: p = 0.8287). Average fixation duration was also not significantly different between conditions (p = 0.7660), with median fixation durations at 313.27 ms [243.597, 367.13] (AI) and 317.29 ms [239.797, 349.86] (without AI). Figure 3 shows visual attention to relevant regions using fixation count to AOIs defined based on the user interface (AOIs illustrated in Fig. 3a). There were more fixations in the map and AI list of findings area when experts had the option for AI than without AI. Moreover, slight increases in the fixation counts on other AOIs related to the interface were apparent when AI was available (see Fig. 3b).

Scanpath transitional behavior

We also looked at experts' gaze transitions between AOIs when they have the option for AI support. Figure 4 reports the total number of transitions from each AOI to another AOI shown in Fig. 3a. Higher transitions are represented by warmer colors from one AOI to another AOI read in the



Fig. 2 | Task time for each image with and without AI support. We observe a slight effect of stimulus viewing order with the plotted task time for each image (means, with bars representing the standard deviation) with AI support (diamonds) and without AI support (circles).

manner of from left to bottom. With the AI support option, dentists had the highest gaze transitions from the AI list of findings to the bitewing (912 transitions) and the second-highest transitions being from bitewing back to the AI list of findings (896 transitions). The map of the teeth and bitewing also had a high number of transitions to and from. In general, fewer transitions were apparent when experts had no option for AI support, though interesting enough, the highest transitions were similar to the AI support condition, but at a much lower magnitude: 131 transitions from AI findings to bitewing and 127 transitions from bitewing to the AI findings. Without the AI support option, transitions to and from the tooth map were much less frequent, as the map was empty.

User interface interactions

Even though AI support was optional, all dentists made use of the AI on the images where they were allowed to according to the randomization schedule. AI was turned on an average of 8.4 times (*range*: 1.1–19.5) and AI stayed on an average of 21012.05 ms (*range*: 9103.11–48506.51 ms).

Dentist on average turned the AI on after an average of 46793.94 ms (sd = 42731.65 ms), i.e. after the first 43.57 %(sd = 32.35 %) of task time. There was a higher number of fixations when they had the AI turned off (23.41 [15.98, 74.43]) compared to when they had the AI turned on (11.56 [4.94, 20.45]: p = 0.001), but this behavior can be attributed to them having the AI turned on for an average of 24.31 % of their total viewing time (Fig. 5). Average fixation duration slightly decreases when the user had AI



a. Defined AOIs

Fig. 3 | **Visual attention to relevant regions (AOIs) of the AI support interface. a** is a depiction of the stimuli used in the experiment, with the bitewing being in the center surrounded by user interface elements with the right-side elements related to

b. Fixation Count for AOIs

the AI support, which were not visible in the no AI condition. **b** shows the average fixation count for dentists viewing with (light blue) and without AI (dark blue). The black error bars represent a confidence interval at 95%.



With AI Support

Without AI Support



Fig. 4 | **Gaze transitional behavior between AOIs with and without AI support.** As we are interested in gaze transitions from one AOI to another, we exclude gaze transitions within an AOI as visualized by missing values along the diagonal. The

warmer color tiles represent a higher number of gaze transition from one AOI to another AOI, read in the manner of from left to bottom.

support on (346.80 ms [264.57, 530.38]) compared to when they had it off (379.96 ms [282.37, 475.28]), though not significantly (p = 0.6781). Nor was fixation frequency significantly different when the user had AI support on (2.30 [1.92, 2.43]) compared to when they had it off (2.02 [1.76, 2.28]: p = 0.1784).

Discussion

The aim of the present study was to use gaze analysis to observe how experts interact with an AI-based decision support tool to investigate dental bitewings. Although dentists' time to inspect bitewings increased when they have the option to use AI support compared to normal visual inspection, the difference was not significant. However, their gaze behavior suggests they can accommodate to the increase in content from AI. These findings are decisive when considering that AI should set out to improve workflows⁹⁵. Even the slightest increase in inspection time can evolve to large delays and fatigue in clinical environments.

Fixation metrics such as the fixation duration and frequency suggested that expert gaze behavior does not change when they have the option for AI support. Even in the context of when the AI was used, these metrics show no significant changes between when the AI is toggled on and off. From previous research, it is known that both these fixation metrics are affected by information content and extraction^{96–100}. It would be expected that average fixation durations would increase to incorporate the additional visual overlays and interface information (tooth map and anomaly labels) the AI support offers. However, this was not the case for the current research. It has been found that fixation frequencies decreased to increasing uncertainty¹⁰¹

and increasing content density¹⁰⁰. Thus, we expected lower fixation frequencies with AI support if the information the system provided would highlight areas dentists were unsure of. It seems that dental experts incorporated AI decision support information into their visual search strategies with little to no change from their usual (i.e, no AI support) visual inspection strategies.

However, there were significantly higher fixation counts when AI was available, especially attending to user interface elements containing AI support information. This finding aligns with usability research, where a higher number of fixations generally indicate more effort during the task^{78,102} or more components to investigate¹⁰³. Dentists had nearly doubled their time spent attending to UI elements, which also is reflected in the higher fixation count on UI elements related to the AI support. Yet, fixation counts viewing the bitewing with AI support were only slightly higher than when there was no option for AI support. Similar behavior was found when using interactive AI systems for fact-checking¹⁰⁴. Overall, dentists spent more time on task when AI support was available, which contributes to more fixations, but it does not affect the rate at which they visually processed the information, evident from the fixation duration and frequency metrics.

Regarding visual attention to specific elements of the system, aspects of the UI that were available for both conditions show almost equal attention







One subject viewing an image

(e.g., tools, info, and browser). However, there is a large increase in attention to the right side of the UI - where elements related to AI were - when there was the option for AI support. Regarding attentional transitions between the AOIs, there was an overall a higher number of transitions with AI support, mainly between the AI list of findings and the bitewing and between the tooth map and the bitewing. This behavior has implications that can be important for future designs of medical AI support systems, as these systems should not draw too much attention away from the medical image, which can increase inspection time and interrupt the already rapid and effective viewing patterns that experts have developed over their years of experience. Such long interruptions during expert visual inspection of CT scans have been found to increase inspection time, but not necessarily affect the diagnostic accuracy¹⁰⁵. We can anticipate such behaviors, even if they are only slight disruptions, can build up over time and can contribute to fatigue.

From our analysis, it seems that dentists employed the AI support as a second reviewer. Experts first investigated the image independently half of the task time, then used the AI, likely confirm their already made findings. Generally, they had the AI support turned on for almost 25% of the task time, but this behavior was varying, ranging from 12.72% to 50.04% of the time. Toggling the AI overlay on and off also varied between being turned on only once in one dataset up to 19 times in another dataset. Figure 6 shows two example interaction behaviors from two subjects for two different bitewings: One expert turned the AI on early and had it on longer than the average behavior (left subfigure), the other turned it on late and more frequently toggled it on and off (right subfigure). Fixation durations during these intervals also show quite the range, where overall durations were shorter when AI support is turned on (apparent in the right subfigure). It could be that experts adopt their own interaction styles, but further research is needed to confirm this behavior. Research that depicts AI as a secondreviewer system has promoted performance and human talents (e.g., creativity and heuristics) and has brought focus to system explainability¹⁰⁶⁻¹⁰⁹. Thus, finding the ideal harmony between user individuality and system information presentation - possibly when to offer suggestions and the reasoning behind it - can enhance interaction and even trust.

A limitation of the current work is that it does not address one of the most popular eye metrics for usability, the pupillary response as an indicator of task workload. This metric would have provided another level of understanding of mental effort with respect to AI support. The omission of pupillary response was needed, as we did not control for additional factors that can naturally reflect the pupillary reflex¹¹⁰. Pupillary response in experts has shown how experts accommodate increasingly difficult image information in medical image inspection⁷⁰. Whether this behavior can also



Another subject viewing another image

Fig. 6 | How experts interact with the AI support interface. When they have the AI turned off/ turned on is indicated in by the dark/light blue bars, respectively, at the bottom of the graph. The orange/navy blue bars are the median fixation durations during the respective on/off interval.

accommodate information presented by a decision support system has yet to be explored. Also, corroborating the visual attention with self reports towards the usability of the system would also provide a more direct link to professional opinion and potential improvements of the system. Since previous research has already established this link between gaze and usability⁷⁵⁻⁷⁸, we chose to focus on how gaze can allude to better interaction suggestions for future systems.Follow-up studies should consider standard usability questionnaires such as the System Usability Scale and User Experience Questionnaire after the experiment^{111,112}. Performance was also not discussed in the current analysis, but in a detailed previous report¹²; dentists with AI support increased their sensitivity (without a decrease in specificity) for detecting caries compared to those without AI support. Lastly, acquiring data from more experts on more images in different environments would help promote the generalizability of our results.

Preliminary design suggestions from these findings could be offering the toggle option closer to the bitewing image, if not slightly overlaying it so as not to occlude relevant features. As experts employed the toggle option often, large transitions away from the main content to a small button on the side may not be favorable. Also, informational content, such as maps or reports, could be positioned closer to the relevant content (the bitewing), favoring shorter saccade lengths. Both suggestions need more research, though operate on a known design principle¹¹³. Finally, the ideal next steps of this research should include the factor of system explainability on how professionals visually interact with AI support systems. How error is communicated to users has shown varying effects on the interaction as well as trust in the system^{114–116}. Eye movements can offer indicators for better ways to promote explainability without overloading the professional. We have shown that expert gaze behavior can help create better, more usable systems that are designed to promote the best abilities of both expert and AI, which ultimately benefits patient care.

Methods

Participants

Twenty-two dental experts (6 women, 16 men) volunteered to participate in the study. They were either employees at the dental hospital of Charité -Universitätsmedizin Berlin or worked in private practices in Berlin, Germany. Criteria for participation was having more than two years of clinical experience (i.e., had finished postgraduate education according to German insurance law), clinically active, and regularly detecting caries in their workflows (orthodontists and oral surgeons were excluded). All participants had normal to corrected vision.

Materials

This research was part of a larger study encompassing dental professional performance with an AI decision-support system¹² and their attention to specific lesions through the support of AI^{117} . More specific details related to the task, the system and specific attention to dental features can be found in the previous studies^{12,69}. There were two conditions that each participant

experienced, bitewing inspection with AI support and without AI support, with all bitewings presented in a web browser that runs the AI software. As interacting with the software is the current research focus, we only briefly detail the bitewings, but¹² provides further details on the bitewing content. From a database of 140 bitewings, 20 were randomly selected and presented in random order to each participant. Of these 20, ten bitewings were randomly selected to have the AI support available, which meant AI could be toggled on/off in the software. Due to the randomization process, bitewings were seen by multiple participants or in different conditions. All bitewings were of the permanent dentition, with at least the crowns of one jaw being visible. Each bitewing was checked and annotated for caries and restorations by four experts, with a fifth expert for crosschecking. A more detailed explanation of this labeling process can be found in ref. 69.

The AI decision support system used was dentalXrai Pro 1.0.4, (dentalXrai Ltd, Berlin, Germany). This system is capable of fully automated AI X-ray reporting in everyday dental clinical practice. It detects pathologies and restorations, highlighting the findings in color, and automatically generates the written documentation. The software allowed the participant to view the native radiograph and its augmented version, where AI software detections are shown as pixel overlays on the bitewing. The participant could also add, remove, or change findings and generate an automated report. Participants in the AI condition can toggle the AI overlay on/off. Figure 7b shows an example of the software interface with the AI overlayed on the bitewing and the respective tooth mappings on the right panel. This example is what participants see when the AI is toggled on. The backbone of the software uses cloud-based machine learning to detect teeth, proximal carious lesions, and restorations visible on the bitewing images. For marking the teeth, a detection model based on U-Net¹¹⁸, whose findings had been validated by an experienced dentist for each bitewing image, was employed. The software version used for this study had a reported accuracy of 0.80, specificity of 0.75, and sensitivity of 0.83¹¹⁹. Further details of the algorithm and its performance can be found in ref. 119.

For consistency in presentation, the non-AI condition viewed images within the same web browser platform, but were not given the option to toggle the AI on and off. Figure 7a shows an example of the software for the non-AI condition. The stimulus remains almost the same for both conditions, but there is less information presented on the right side of the browser for the non-AI condition, as this region provided the AI content. Additionally, there are no pixel overlays on the bitewing images in this condition.

Procedure

At least one week before the study, participants received a handbook about the AI software and were advised to try it out on a minimum of four independent bitewing radiographs. This was encouraged so that they had better understanding of the system and its capabilities. The study was run at either the dental hospital of Charité - Universitätsmedizin Berlin or the participant's private practice.



Fig. 7 | Example of web browser interface with and without AI overlays. Example of the web browser interface with presentation of the bitewing and left panel for image manipulations and right panel presented the AI information for the AI condition (b) or was blank for the non-AI condition (a).

The task evaluated for this current research was interacting with the software with AI support and without AI support, and how experts use the software to visually inspect images. 20 bitewing images were randomly chosen from the bitewing pool and presented in random order in the web browser that runs the dentalXrai software. This generation of images was unique to each participant. Prior to uploading an image into the software, each participant drew a slip of paper from a pool of 20 slips contained in a sealed opaque envelope (ten indicating to use the AI software and ten not) to determine which image would have the AI software (intervention) or not (control).

Participants performed the task in one session, operating at their own pace. They viewed the images in the web browser and depending on which condition the image was, they could use the AI software or could not. In the AI condition, dentists could then enable or disable the AI augmentations as needed. For both conditions, they then verbally reported any proximal caries detections and their corresponding treatment decisions to the study assistant, though this was not evaluated in the current analysis. The participants concluded the examination of the image, and the next one could be uploaded, following the protocol for drawing a slip of paper.

We chose this design method to be convenient for our participants. Asking medical professionals to set aside long or multiple windows of time can become harder for them to fit into their busy schedules. To avoid dropout rate or inconsistent lengths between two sessions, we chose one session. This choice can also control for errors in replicability of the setup, as we traveled to them. Additionally, one session, with highly randomized stimuli, better controls for any fatigue or learning effects participants may exhibit.

This evaluation is nested within a randomized, controlled, nonblinded, clustered cross-over, superiority trial with an allocation ratio of 1:1¹², assessing the impact of an artificial intelligence (AI) software for detection of carious lesions. The trial was not conducted during clinical care and on actual patients, but on retrospectively sampled imagery material, which was randomly assessed with and without assistance from the AI software. The trial was registered at Deutsches Register Klinischer Studien (DRKS00022357). Ethical approval was provided by the Charité -Universitätsmedizin Berlin (EA/144/20). During the study, we had recorded dentists' gaze patterns, and here we present the gaze pattern behavior of the control group (i.e., dentists not using AI). Written informed consent was obtained from all participating dentists.

Eye tracking

To record gaze data, we used the SmartEye Aurora remote eye tracker running at 60 Hz, positioned under a monitor with a resolution of 1920×1080 pixels. Data collection took place in dimly lit rooms at either Charité or in private clinics for participant convenience; the study investigator brought the monitor to their clinic. Participants were unconstrained and sat approximately 70 cm from the tracker. An initial 9-point calibration and validation were performed. Recalibration was done if the software indicated that the calibration quality was poor. Gaze data was collected for the whole duration of the study using the *iMotions* software (version 8.2.22899.4). Event detection was performed using the iMotions implementation of the I-VT algorithm, with a minimum fixation duration of 60 milliseconds (ms) and a velocity threshold of 30deg/s. The current analysis used the fixations reported from the software, which are interpolated between the left and the right eye. We interpret fixations as the areas of attentional focus related to the stimuli presented on the screen.

Data preparation

Data collection resulted in 445 datasets from the participants viewing bitewing radiographs. As five participants unintentionally examined one image twice, we excluded the first time they viewed the image, as it was too short for proper investigation (440 Datasets). To ensure gaze pattern data quality, we removed datasets with an average reported gaze signal quality lower than 0.60 (valid signal over total signal, using a scale of 0.0 being the

lowest and 1.0 being the highest quality). 80 datasets were excluded from this criterion. A stimulus presentation error resulted in the removal of 11 further datasets. These exclusion criteria adhere to standard guidelines used in eye tracking research on data quality control^{120,121} Overall, 349 datasets (170 without AI and 179 with AI) were included in the current analysis.

Reporting summary

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

Data availability

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Code availability

The code created and used for analyzing the preprocessed data for the current study is available from the corresponding author on reasonable request.

Received: 10 November 2023; Accepted: 12 July 2024; Published online: 27 July 2024

References

- Kasneci, E. et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learn. Individ. Differences* 103, 102274 (2023).
- Ji, H., Han, I. & Ko, Y. A systematic review of conversational ai in language education: focusing on the collaboration with human teachers. J. Res. Technol. Educ. 55, 48–63 (2022).
- Huang, M.-H. & Rust, R. T. A strategic framework for artificial intelligence in marketing. J. Acad. Mark. Sci. 49, 30–50 (2021).
- Ma, L. & Sun, B. Machine learning and ai in marketing–connecting computing power to human insights. *Int. J. Res. Mark.* 37, 481–504 (2020).
- Xing, Y., Lv, C., Cao, D. & Hang, P. Toward human-vehicle collaboration: Review and perspectives on human-centered collaborative automated driving. *Transportation Res. part C: Emerg. Technol.* **128**, 103199 (2021).
- Topol, E. J. High-performance medicine: the convergence of human and artificial intelligence. *Nat. Med.* 25, 44–56 (2019).
- Ostheimer, J., Chowdhury, S. & Iqbal, S. An alliance of humans and machines for machine learning: Hybrid intelligent systems and their design principles. *Technol. Soc.* 66, 101647 (2021).
- Martín-Noguerol, T. et al. Artificial intelligence in radiology: relevance of collaborative work between radiologists and engineers for building a multidisciplinary team. *Clin. Radiol.* **76**, 317–324 (2020).
- 9. World Health Organization. *Ethics and governance of artificial intelligence for health: Who guidance*. Tech. Rep. (World Health Organization, 2021).
- Rodríguez-Ruiz, A. et al. Detection of breast cancer with mammography: effect of an artificial intelligence support system. *Radiology* 290, 305–314 (2019).
- Wu, N. et al. Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Trans. Med. Imaging* 39, 1184–1194 (2019).
- Mertens, S., Krois, J., Cantu, A. G., Arsiwala, L. T. & Schwendicke, F. Artificial intelligence for caries detection: Randomized trial. *J. Dent.* 115, 103849 (2021).
- Waite, S. et al. A review of perceptual expertise in radiology-how it develops, how we can test it, and why humans still matter in the era of artificial intelligence. *Academic Radiol.* 27, 26–38 (2020).
- Houssami, N., Kirkpatrick-Jones, G., Noguchi, N. & Lee, C. I. Artificial intelligence (ai) for the early detection of breast cancer: a scoping review to assess ai's potential in breast screening practice. *Expert Rev. Med. Devices* 16, 351–362 (2019).

- Benzekry, S. Artificial intelligence and mechanistic modeling for clinical decision making in oncology. *Clin. Pharmacol. Therapeutics* 108, 471–486 (2020).
- Rubin, D. L. Artificial intelligence in imaging: the radiologist's role. J. Am. Coll. Radiol. 16, 1309–1317 (2019).
- 17. Oakden-Rayner, L. The rebirth of CAD: how is modern Al different from the CAD we know? *Radiol Artif Intell* **1**, e180089 (2019).
- 18. Tchou, P. M. et al. Interpretation time of computer-aided detection at screening mammography. *Radiology* **257**, 40–46 (2010).
- Levy, A., Agrawal, M., Satyanarayan, A. & Sontag, D. Assessing the impact of automated suggestions on decision making: Domain experts mediate model errors but take less initiative. In *Proceedings* of the 2021 CHI Conference on Human Factors in Computing Systems, 1–13 (ACM, 2021).
- Dreiseitl, S. & Binder, M. Do physicians value decision support? a look at the effect of decision support systems on physician opinion. *Artif. Intell. Med.* 33, 25–30 (2005).
- Fenton, J. J. et al. Influence of computer-aided detection on performance of screening mammography. *N. Engl. J. Med.* 356, 1399–1409 (2007).
- Ongena, Y. P., Yakar, D., Haan, M. & Kwee, T. C. Artificial intelligence in screening mammography: a population survey of women's preferences. J. Am. Coll. Radiol. 18, 79–86 (2021).
- Lennox-Chhugani, N., Chen, Y., Pearson, V., Trzcinski, B. & James, J. Women's attitudes to the use of ai image readers: a case study from a national breast screening programme. *BMJ Health Care Informatics* 28, e100293 (2021).
- 24. Price, W. N., Gerke, S. & Cohen, I. G. Potential liability for physicians using artificial intelligence. *JAMA* **322**, 1765–1766 (2019).
- Eschert, T. et al. A survey on the use of artificial intelligence by clinicians in dentistry and oral and maxillofacial surgery. *Medicina* 58, 1059 (2022).
- Cabitza, F. & Zeitoun, J.-D. The proof of the pudding: in praise of a culture of real-world validation for medical artificial intelligence. *Ann. Transl. Med.* 7, 161 (2019).
- 27. Rainey, C. et al. Uk reporting radiographers' perceptions of ai in radiographic image interpretation–current perspectives and future developments. *Radiography* **28**, 881–888 (2022).
- Bohr, A. & Memarzadeh, K. The rise of artificial intelligence in healthcare applications. In *Artificial Intelligence in healthcare*, 25–60 (Elsevier, 2020).
- Jiang, F. et al. Artificial intelligence in healthcare: past, present and future. Stroke Vascular Neurol. 2, 230–243 (2017).
- Sendak, M. P. et al. A path for translation of machine learning products into healthcare delivery. *EMJ Innov.* 10, 19–00172 (2020).
- Davenport, T. & Kalakota, R. The potential for artificial intelligence in healthcare. *Future Healthc. J.* 6, 94 (2019).
- Guo, J. & Li, B. The application of medical artificial intelligence technology in rural areas of developing countries. *Health equity* 2, 174–181 (2018).
- Laï, M.-C., Brian, M. & Mamzer, M.-F. Perceptions of artificial intelligence in healthcare: findings from a qualitative survey study among actors in france. *J. Transl. Med.* 18, 1–13 (2020).
- Esmaeilzadeh, P. Use of ai-based tools for healthcare purposes: a survey study from consumers' perspectives. *BMC Med. Inform. Decis. Mak.* 20, 1–19 (2020).
- Caspers, J. Translation of predictive modeling and ai into clinics: a question of trust. *Eur. Radiol.* **31**, 4947–4948 (2021).
- Vössing, M., Kühl, N., Lind, M. & Satzger, G. Designing transparency for effective human-ai collaboration. *Inf. Syst. Front.* 24, 877–895 (2022).
- 37. Ma, J. et al. Towards trustworthy ai in dentistry. *J. Dent. Res.* **101**, 1263–1268 (2022).
- Amershi, S. et al. Guidelines for human-ai interaction. In Proceedings of the 2019 chi conference on human factors in computing systems, 1–13 (ACM, 2019).

- Ganesan, A., Alakhras, M., Brennan, P. C. & Mello-Thoms, C. A review of factors influencing radiologists' visual search behaviour. J. Med. Imaging Radiat. Oncol. 62, 747–757 (2018).
- 40. Van der Gijp, A. et al. How visual search relates to visual diagnostic performance: a narrative systematic review of eye-tracking research in radiology. *Adv. Health Sci. Educ.* **22**, 765–787 (2017).
- Ericsson, K. A. & Lehmann, A. C. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annu. Rev. Psychol.* 47, 273–305 (1996).
- 42. Noton, D. & Stark, L. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vis. Res.* **11**, 929–IN8 (1971).
- 43. Holmqvist, K. et al. *Eye Tracking: A Comprehensive Guide to Methods and Measures* (OUP Oxford, 2011).
- Gegenfurtner, A., Lehtinen, E. & Säljö, R. Expertise differences in the comprehension of visualizations: A meta-analysis of eye-tracking research in professional domains. *Educ. Psychol. Rev.* 23, 523–552 (2011).
- 45. Nakashima, R. et al. The effect of expert knowledge on medical search: medical experts have specialized abilities for detecting serious lesions. *Psychological Res.* **79**, 729–738 (2015).
- Brams, S. et al. The relationship between gaze behavior, expertise, and performance: A systematic review. *Psychological Bull.* 145, 980 (2019).
- Nodine, C. F., Kundel, H. L., Lauver, S. C. & Toto, L. C. Nature of expertise in searching mammograms for breast masses. *Academic Radiol.* 3, 1000–1006 (1996).
- Kundel, H. L., Nodine, C. F., Conant, E. F. & Weinstein, S. P. Holistic component of image perception in mammogram interpretation: gaze-tracking study. *Radiology* 242, 396–402 (2007).
- Krupinski, E. A. Visual scanning patterns of radiologists searching mammograms. *Academic Radiol.* 3, 137–144 (1996).
- Cooper, L. et al. The assessment of stroke multidimensional ct and mr imaging using eye movement analysis: does modality preference enhance observer performance? *Med. Imaging 2010 Image Percept. Observer Perform. Technol. Assess.* **7627**, 76270B (2010).
- 51. Mallett, S. et al. Tracking eye gaze during interpretation of endoluminal three-dimensional ct colonography: visual perception of experienced and inexperienced readers. *Radiology* **273**, 783–792 (2014).
- Warren, A. L., Donnon, T. L., Wagg, C. R., Priest, H. & Fernandez, N. J. Quantifying novice and expert differences in visual diagnostic reasoning in veterinary pathology using eye-tracking technology. *J. Vet. Med. Educ.* 45, 295–306 (2018).
- Haider, H. & Frensch, P. A. Eye movement during skill acquisition: More evidence for the information-reduction hypothesis. *J. Exp. Psychol.* 25, 172 (1999).
- 54. Topczewski, J. J., Topczewski, A. M., Tang, H., Kendhammer, L. K. & Pienta, N. J. Nmr spectra through the eyes of a student: eye tracking applied to nmr items. *J. Chem. Educ.* **94**, 29–37 (2017).
- Just, M. A. & Carpenter, P. A. A capacity theory of comprehension: individual differences in working memory. *Psychological Rev.* 99, 122 (1992).
- 56. Marshall, S. P. Identifying cognitive state from eye metrics. *Aviat. Space Environ. Med.* **78**, B165–B175(11) (2007).
- Kok, E. M., De Bruin, A. B., Robben, S. G. & Van Merrienboer, J. J. Looking in the same manner but seeing it differently: Bottom-up and expertise effects in radiology. *Appl. Cogn. Psychol.* 26, 854–862 (2012).
- Turgeon, D. P. & Lam, E. W. Influence of experience and training on dental students' examination performance regarding panoramic images. *J. Dent. Educ.* 80, 156–164 (2016).
- 59. Wood, G. et al. Visual expertise in detecting and diagnosing skeletal fractures. *Skelet. Radiol.* **42**, 165–172 (2013).
- Donovan, T. & Litchfield, D. Looking for cancer: Expertise related differences in searching and decision making. *Appl. Cogn. Psychol.* 27, 43–49 (2013).

- Suwa, K., Furukawa, A., Matsumoto, T. & Yosue, T. Analyzing the eye movement of dentists during their reading of ct images. *Odontology* 89, 0054–0061 (2001).
- Alamudun, F. et al. Fractal analysis of visual search activity for mass detection during mammographic screening. *Med. Phys.* 44, 832–846 (2017).
- 63. Hermanson, B. P., Burgdorf, G. C., Hatton, J. F., Speegle, D. M. & Woodmansey, K. F. Visual fixation and scan patterns of dentists viewing dental periapical radiographs: an eye tracking pilot study. *J. Endod.* **44**, 722–727 (2018).
- 64. Gandomkar, Z. & Mello-Thoms, C. Visual search in breast imaging. Br. J. Radiol. **92**, 20190057 (2019).
- Fox, S. E. & Faulkner-Jones, B. E. Eye-tracking in the study of visual expertise: Methodology and approaches in medicine. *Frontline Learn. Res.* 5, 29–40 (2017).
- Castner, N. et al. Deep semantic gaze embedding and scanpath comparison for expertise classification during opt viewing. In ACM Symposium on Eye Tracking Research and Applications, 1–10 (ACM, 2020).
- Grünheid, T., Hollevoet, D. A., Miller, J. R. & Larson, B. E. Visual scan behavior of new and experienced clinicians assessing panoramic radiographs. *J. World Federation Orthod.* 2, e3–e7 (2013).
- Castner, N., Frankemölle, J., Keutel, C., Huettig, F. & Kasneci, E. LSTMS can distinguish dental expert saccade behavior with high" plaque-urracy". In 2022 Symposium on Eye Tracking Research and Applications, 1–7 (ACM, 2022).
- 69. Arsiwala-Scheppach, L. T. et al. Gaze patterns of dentists while evaluating bitewing radiographs. *J. Dentistry* **135**, 104585 (2023).
- 70. Castner, N. et al. Pupil diameter differentiates expertise in dental radiography visual search. *PloS one* **15**, e0223941 (2020).
- Novák, J. Š., Masner, J., Benda, P., Šimek, P. & Merunka, V. Eye tracking, usability, and user experience: A systematic review. *Int. J. Hum. Comput. Interaction* **39**, 1–17 (2023).
- Wang, J., Xu, Z., Wang, X. & Lu, J. A comparative research on usability and user experience of user interface design software. *Int. J. Adv. Comput. Sci. Appl.* **13**, 21–29 (2022).
- Jacob, R. J. & Karn, K. S. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye*, 573–605 (Elsevier, 2003).
- Majrashi, K., Hamilton, M. & Uitdenbogerd, A. L. Correlating crossplatform usability problems with eye tracking patterns. In *Proceedings of the 30th International BCS Human Computer Interaction Conference 30*, 1–11 (ACM, 2016).
- Goldberg, J. H., Stimson, M. J., Lewenstein, M., Scott, N. & Wichansky, A. M. Eye tracking in web search tasks: design implications. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, 51–58 (ACM, 2002).
- Kotval, X. P. & Goldberg, J. H. Eye movements and interface component grouping: An evaluation method. In *Proceedings of the human factors and ergonomics society annual meeting*, vol. 42, 486–490 (SAGE Publications Sage CA, 1998).
- 77. Bruneau, D., Sasse, M. A. & McCarthy, J. D. *The eyes never lie: the use of eyetracking data in hci research* (ACM, 2002).
- Goldberg, J. H. & Kotval, X. P. Computer interface evaluation using eye movements: methods and constructs. *Int. J. Ind. Ergonomics* 24, 631–645 (1999).
- Pomplun, M. & Sunkara, S. Pupil dilation as an indicator of cognitive workload in human-computer interaction. In *Human-Centered Computing*, 542–546 (CRC Press, 2019).
- Harley, L. et al. Designing usable voting systems for voters with hidden barriers: A pilot study. *Int. J. Hum. Comput. Interact.* 32, 103–118 (2016).
- Dong, W., Liao, H., Xu, F., Liu, Z. & Zhang, S. Using eye tracking to evaluate the usability of animated maps. *Sci. China Earth Sci.* 57, 512–522 (2014).

- Mariakakis, A., Goel, M., Aumi, M. T. I., Patel, S. N. & Wobbrock, J. O. Switchback: Using focus and saccade tracking to guide users' attention for mobile task resumption. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2953–2962 (ACM, 2015).
- Zhou, M., Wang, Z. & Zheng, Y. Study on usability evaluation of medical software interface based on eye movement tracking technology. In *Twelfth International Conference on Machine Vision* (*ICMV 2019*), vol. 11433, 148–156 (SPIE, 2020).
- Koester, T., Brøsted, J. E., Jakobsen, J. J., Malmros, H. P. & Andreasen, N. K. The use of eye-tracking in usability testing of medical devices. In *Proceedings of the International Symposium on Human Factors and Ergonomics in Health Care*, vol. 6, 192–199 (SAGE Publications Sage CA, 2017).
- 85. Jiang, M., Liu, S., Feng, Q., Gao, J. & Zhang, Q. Usability study of the user-interface of intensive care ventilators based on user test and eye-tracking signals. *Med. Sci. Monit. Int. Med. J. Exp. Clin. Res.* **24**, 6617 (2018).
- Eghdam, A., Forsman, J., Falkenhav, M., Lind, M. & Koch, S. Combining usability testing with eye-tracking technology: evaluation of a visualization support for antibiotic use in intensive care. In User Centred Networked Health Care, 945–949 (IOS Press, 2011).
- Forsman, J., Anani, N., Eghdam, A., Falkenhav, M. & Koch, S. Integrated information visualization to support decision making for use of antibiotics in intensive care: design and usability evaluation. *Inform. Health Soc. Care* 38, 330–353 (2013).
- 88. Heijsters, F. et al. A usability evaluation of the perceived user friendliness, accessibility, and inclusiveness of a personalized digital care pathway tool. *Int. J. Med. Inform.* **175**, 105070 (2023).
- Asan, O. & Yang, Y. et al. Using eye trackers for usability evaluation of health information technology: a systematic literature review. *JMIR Hum. factors* 2, e4062 (2015).
- Theissler, A., Kraft, A.-L., Rudeck, M. & Erlenbusch, F. Vial-ad: Visual interactive labelling for anomaly detection-an approach and open research questions. In *IAL@ PKDD/ECML*, 84–89 (Springer, 2020) https://www.springer.com/gp/computer-science/Incs/societiesand-Incs/ecml-pkdd.
- Krak, I., Barmak, O. & Manziuk, E. Using visual analytics to develop human and machine-centric models: A review of approaches and proposed information technology. *Comput. Intell.* 38, 921–946 (2020).
- Khosravan, N. et al. Gaze2segment: a pilot study for integrating eyetracking technology into medical image segmentation. In *Medical Computer Vision and Bayesian and Graphical Models for Biomedical Imaging*, 94–104 (Springer, 2016).
- Pesce, E. et al. Learning to detect chest radiographs containing pulmonary lesions using visual attention networks. *Med. Image Anal.* 53, 26–38 (2019).
- Schlemper, J. et al. Attention gated networks: Learning to leverage salient regions in medical images. *Med. Image Anal.* 53, 197–207 (2019).
- Kapoor, N., Lacson, R. & Khorasani, R. Workflow applications of artificial intelligence in radiology and an overview of available tools. *J. Am. Coll. Radiol.* **17**, 1363–1370 (2020).
- Bylinskii, Z., Borkin, M. A., Kim, N. W., Pfister, H. & Oliva, A. Eye fixation metrics for large scale evaluation and comparison of information visualizations. In *Eye Tracking and Visualization: Foundations, Techniques, and Applications. ETVIS 2015 1*, 235–255 (Springer, 2017).
- 97. Just, M. A. & Carpenter, P. A. The role of eye-fixation research in cognitive psychology. *Behav. Res. Methods Instrum.* **8**, 139–143 (1976).
- Altin Gumussoy, C., Pekpazar, A., Esengun, M., Bayraktaroglu, A. E. & Ince, G. Usability evaluation of tv interfaces: Subjective evaluation vs. objective evaluation. *Int. J. Hum. Comput. Interact.* 38, 661–679 (2022).

- Ooms, K., De Maeyer, P., Fack, V., Van Assche, E. & Witlox, F. Interpreting maps through the eyes of expert and novice users. *Int. J. Geographical Inf. Sci.* 26, 1773–1788 (2012).
- Liao, H., Wang, X., Dong, W. & Meng, L. Measuring the influence of map label density on perceived complexity: A user study using eye tracking. *Cartogr. Geographic Inf. Sci.* 46, 210–227 (2019).
- Brunyé, T. T. & Gardony, A. L. Eye tracking measures of uncertainty during perceptual decision making. *Int. J. Psychophysiol.* **120**, 60–68 (2017).
- 102. Sharif, B., Falcone, M. & Maletic, J. I. An eye-tracking study on the role of scan time in finding source code defects. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, 381–384 (ACM, 2012).
- Dong, W., Wang, S., Chen, Y. & Meng, L. Using eye tracking to evaluate the usability of flow maps. *ISPRS Int. J. Geo Inf.* 7, 281 (2018).
- 104. Shi, L., Bhattacharya, N., Das, A., Lease, M. & Gwizdka, J. The effects of interactive ai design on user behavior: An eye-tracking study of fact-checking covid-19 claims. In *Proceedings of the 2022 Conference on Human Information Interaction and Retrieval*, 315–320 (ACM, 2022).
- Williams, L. H. & Drew, T. Distraction in diagnostic radiology: how is search through volumetric medical images affected by interruptions? *Cogn. Res.: Princ. Implic.* 2, 1–11 (2017).
- 106. Yang, Q., Steinfeld, A. & Zimmerman, J. Unremarkable ai: Fitting intelligent decision support into critical, clinical decision-making processes. In *Proceedings of the 2019 CHI conference on human factors in computing systems*, 1–11 (ACM, 2019).
- 107. Petrone, J. Fda approves stroke-detecting ai software. *Nat. Biotechnol.* **36**, 290–290 (2018).
- Meskó, B. & Görög, M. A short guide for medical professionals in the era of artificial intelligence. NPJ Digital Med. 3, 126 (2020).
- Cai, C. J., Winter, S., Steiner, D., Wilcox, L. & Terry, M. "hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making. *Proc. ACM Hum. Comput. Interact.* 3, 1–24 (2019).
- Pfleging, B., Fekety, D. K., Schmidt, A. & Kun, A. L. A model relating pupil diameter to mental workload and lighting conditions. In Proceedings of the 2016 CHI conference on human factors in computing systems, 5776–5788 (ACM, 2016).
- Bangor, A., Kortum, P. T. & Miller, J. T. An empirical evaluation of the system usability scale. *Int. J. Hum. Comput. Interact.* 24, 574–594 (2008).
- 112. Laugwitz, B., Held, T. & Schrepp, M. Construction and evaluation of a user experience questionnaire. In HCI and Usability for Education and Work: 4th Symposium of the Workgroup Human-Computer Interaction and Usability Engineering of the Austrian Computer Society, USAB 2008, Graz, Austria, November 20-21, 2008. Proceedings 4, 63–76 (Springer, 2008).
- MacKenzie, I. S. Fitts' law as a research and design tool in humancomputer interaction. *Hum. Comput. Interact.* 7, 91–139 (1992).
- 114. Schaekermann, M. et al. Ambiguity-aware ai assistants for medical data analysis. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, 1–14 (ACM, 2020).
- 115. Lukashova-Sanz, O., Dechant, M. & Wahl, S. The influence of disclosing the ai potential error to the user on the efficiency of user–ai collaboration. *Appl. Sci.* **13**, 3572 (2023).
- Alam, L. & Mueller, S. Examining the effect of explanation on satisfaction and trust in ai diagnostic systems. *BMC Med. Inform. Decis. Mak.* 21, 178 (2021).
- Arsiwala-Scheppach, L. T. et al. Impact of artificial intelligence on dentists' gaze during caries detection: A randomized controlled trial. *J. Dent.* 140, 104793 (2024).
- 118. Ronneberger, O., Fischer, P. & Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *International*

Conference on Medical image computing and computer-assisted intervention (Springer, 2015).

- Cantu, A. G. et al. Detecting caries lesions of different radiographic extension on bitewings using deep learning. *J. Dent.* **100**, 103425 (2020).
- Fiedler, S., Schulte-Mecklenbeck, M., Renkewitz, F. & Orquin, J. L. Increasing reproducibility of eye-tracking studies: The eyeguidelines. In *A handbook of process tracing methods*, 65–75 (Routledge, 2019).
- Cho, V. Y., Loh, X. H., Abbott, L., Mohd-Isa, N. A. & Anthonappa, R. P. Reporting eye-tracking studies in dentistry (reside) checklist. *J. Dent.* **129**, 104359 (2023).

Acknowledgements

This research is supported by European Union's Horizon 2020 research and innovation program under grant agreement No.951910 (acronym: MAIA). We would also like to extend our gratitude to Prof. Dr. Fabian Hüttig for his support in this project.

Author contributions

N.C. and L.A-S. performed the data measurements, analysis, and interpretation and drafted and critically revised the manuscript. S.M. contributed to conception, design, data acquisition. E.T. contributed to the initial data preprocessing and statistical investigation. J.K. and F.S. contributed to conception. E.K., S.W. and F.S. contributed to results discussion and manuscript review and corrections.

Funding

Open Access funding enabled and organized by Projekt DEAL.

Competing interests

F.S. and J.K. are co-founders of the startup dentalXrai GmbH. dentalXrai GmbH had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results. The remaining authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41746-024-01192-8.

Correspondence and requests for materials should be addressed to Nora Castner.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2024