Initial Responses to False Positives in AI-supported Continuous Interactions – A Colonoscopy Case Study

NIELS VAN BERKEL, Aalborg University, Denmark and University College London, United Kingdom
JEREMY OPIE, University College London, United Kingdom
OMER F. AHMAD, University College London, United Kingdom
LAURENCE LOVAT, University College London Hospitals, United Kingdom
DANAIL STOYANOV, University College London, United Kingdom
ANN BLANDFORD, University College London, United Kingdom

The use of Artificial Intelligence in clinical support systems is increasing. In this paper we focus on AI support for continuous interaction scenarios. A thorough understanding of end-user behaviour during these continuous Human-AI interactions, in which user input is sustained over time and during which AI suggestions can appear at any time, is still missing. We present a controlled lab-study involving 21 endoscopists and an AI colonoscopy support system. Using a custom-developed application and an off-the-shelf videogame controller, we record participants' navigation behaviour and clinical assessment across 14 endoscopic videos. Each video is manually annotated to mimic an AI recommendation, being either true positive or false positive in nature. We find that time between AI recommendation and clinical assessment is significantly longer for incorrect assessments. Further, the type of medical content displayed significantly affects decision time. Finally, we discover that the participant's clinical role plays a large part in the perception of clinical AI support systems. Our study presents a realistic assessment of the effects of imperfect and continuous AI support in a clinical scenario.

CCS Concepts: • **Human-centered computing** → **Human computer interaction (HCI)**; *HCI design and evaluation methods*; *Empirical studies in HCI*; • **Applied computing** → *Life and medical sciences.*

Additional Key Words and Phrases: Human-AI interaction, artificial intelligence, colonoscopy, support system, false positives, continuous interaction, clinical decision support

ACM Reference Format:
Niels van Berkel, Jeremy Opie, Omer F. Ahmad, Laurence Lovat, Danail Stoyanov, and Ann Blandford. 2021. Initial Responses to False Positives in AI-supported Continuous Interactions – A Colonoscopy Case Study. *ACM Trans. Interact. Intell. Syst.* 1, 1, Article 1 (January 2021), 19 pages. https://doi.org/10.1145/3480247

1 INTRODUCTION

Artificial Intelligence (AI) based support systems are increasingly common across a variety of industries, with the healthcare sector often identified as one of the areas that can be positively transformed by AI technology [9, 17]. Although the idea of AI support is not new (see *e.g.* the 1993 December issue of the Communications of the ACM [13]), only recently has the widespread and real-world integration of AI in end-user facing software really commenced [9]. Within the medical

Authors' addresses: Niels van Berkel, nielsvanberkel@cs.aau.dk, Aalborg University, Aalborg, Denmark, University College London, UCL Interaction Centre, London, United Kingdom; Jeremy Opie, j.opie@ucl.ac.uk, University College London, UCL Interaction Centre, London, United Kingdom; Omer F. Ahmad, o.ahmad@ucl.ac.uk, University College London, Wellcome/EPSRC Centre for Interventional & Surgical Sciences, London, United Kingdom; Laurence Lovat, l.lovat@ucl.ac.uk, University College London Hospitals, London, United Kingdom; Danail Stoyanov, danail.stoyanov@ucl.ac.uk, University College London, Medical Physics and Bioengineering, London, United Kingdom; Ann Blandford, a.blandford@ucl.ac.uk, University College London, UCL Interaction Centre, London, United Kingdom.

domain, AI has been identified as a beneficial technology for a wide range of application areas such as image recognition to support diagnosis in radiology and pathology [19, 32], robot-supported surgery [17], and home monitoring technology for fall detection [31].

Despite the far-reaching and beneficial possibilities of AI support technology, recent work in Human-Computer Interaction (HCI) and beyond has raised growing concerns regarding numerous downsides related to the use of AI [2]. This includes questions on how to deal with incorrect classifications by AI systems [24], concerns regarding the fairness of AI systems [42, 49], and the presentation of irrelevant information to the user [7]. In this paper, we specifically focus on the effect of imperfect AI support systems during continuous user interaction scenarios. The majority of HCI literature on AI support has focused on intermittent scenarios [4, 7, 12], *e.g.*, pathology classification on a stationary image, in which the interaction between user and AI can be defined as a turn-taking process. Continuous interaction scenarios, in which user input is sustained over a period of time and may receive AI input at any time, remain under-explored despite the unique challenges faced in terms of user interaction [40]. Clinical decision support during colonoscopy, the case presented in this paper, is indicative of this different type of human-AI interaction, as AI-powered suggestions may repeatedly appear and disappear on a screen while a clinician navigates through a patient's colon.

In a setting of continuous AI-support, AI recommendations (regardless of their correctness) may overlap with the user's visual interest area, may interrupt the user's flow, or even distract the user from relevant information elsewhere on the display. Capturing the effect of AI classifications, in particular true positives (a correct classification of an entity, *e.g.* a polyp) and false positives (incorrectly classifying an entity that should not be classified), on user interaction and perceptions towards AI support is critical to understand how AI support systems can be successfully integrated in daily clinical practice. The effect of false positives in particular is a critical question for the medical domain, where unnecessary interventions or interruptions in a procedure can result in negative medical outcomes, distressing patient experiences, and increased financial costs [16, 22]. A recent research priority setting study for AI in colonoscopy furthermore highlighted the effect of false positives as a key research concern [1]. While the rates of false negatives (failing to classify an element that should be classified) and true negatives (correctly not classifying an element that should not be classified) are also critical to the effectiveness of AI support systems, this is out of scope for the current paper which focuses on true and false positives during user interaction.

In order to systematically study the effect of (imperfect) AI support in a continuous support scenario, we conducted a controlled lab-based experiment in which we studied the behaviour of endoscopists (N = 21) when navigating through AI-overlaid patient footage. We annotated a total of 14 video recordings of real-world endoscopic procedures across six distinct categories, each highlighting an entity that can be seen during a colonoscopy (see Figure 1). Using a videogame controller, participants navigated through these videos and were presented with either a false positive or true positive AI recommendation. Participants were asked to provide their clinical assessment of the highlighted object (non-polyp or polyp) while navigating through the video. We continuously and unobtrusively captured the participants' viewing behaviours and collected their perspectives on both the perceived usefulness and hindrance of the AI for each video.

Our results identify significant differences in both the effect of AI and perceptions towards AI support across both the different professional roles included in our sample and their respective endoscopic experience, as quantified through their number of completed colonoscopies. Furthermore, we find substantial differences in both the participants' decision time and navigation behaviour between the different video categories. Interestingly, the impact of false positives and true positives on participant browsing behaviour is highly similar. Our work contributes towards a better understanding of the impact of (in)correct AI support in continuous support scenarios. Furthermore, we

provide a methodological contribution to enhance the ecological validity of studying participants' behaviour and perceptions towards AI in the context of medical studies.

2 RELATED WORK

Colonoscopies are performed primarily to detect pre-cancerous polyps. Adenomas are the most common type of pre-cancerous polyp. For cancer surveillance, the adenoma detection rate (ADR) is considered the optimal quality indicator for colonoscopy examinations [26]. It has been highlighted that ADR is an independent predictor of the risk of interval cancer [20], and that every 1% increase in ADR leads to a reduction in the risk of interval cancer by 3% [8]. However, Ahn et al. discovered that even with controlled bowel preparation there is a 17% miss rate of adenomas [3]. Lee et al. revealed that with the assistance of an experienced gastrointestinal (GI) endoscopy nurse the ADR can be improved. They also discovered that this was particularly beneficial when coupled with inexperienced endoscopists [23]. Buchner et al. also found that having a second pair of eyes assisting with inspection improved ADR, and identified that with assistance there was an 8% increase in the detection rate of small adenomas [6]. It is important that adenomas are detected during colonoscopies so that they can be removed, as unlike radiology procedures, there is no opportunity to review and remove them afterwards [18]. As stated by Hassan et al., "*Differently from radiology procedures, colonoscopy and endoscopy, in general, are real-time procedures requiring complex analysis of millions of frames without the opportunity to review them afterwards.*" [18].

2.1 AI Support Systems in Colonoscopy

Computer-aided detection (CADe) systems have been developed in an attempt to improve ADR, by highlighting polyps during the colonoscopy procedure that might otherwise be overlooked by the human observer. Several pre-clinical studies have been published in the field of CADe for colonoscopy, where algorithmic performance is often evaluated on a per-frame basis for video colonoscopy data. For example, Misawa et al. described a CADe system that was able to detect 94% of test polyps, but with a false positive detection rate of 60% [28].

More recently, prospective, randomised, clinical-trials (RCT) of CADe software have been published. Wang et al. conducted a double-blind RCT involving 1010 patients, where 508 were randomised to CADe colonoscopy and 502 to colonoscopy with a sham CADe system (trained to present alert boxes on polyp-like non-polyp structures, *e.g.*, bubbles, wrinkled colon wall) [47]. There was a statistically significant increase in ADR for CADe colonoscopy compared to the sham system. The authors reported 48 'consistent' false positive alarms by the CADe system. It is important to note that the definition of false positives in the study was highly subjective since these were not defined on a per video frame basis. False positives were defined as an area that was continuously traced by the system but deemed by the endoscopist not to be a polyp. In practice, many false positives appear briefly, and therefore the actual false positive rate is likely to be higher in real-world clinical practice. Furthermore, evaluation of the CADe system was avoided in situations where the colon was not fully inflated, since the wrinkled appearance of the colonic wall can easily be incorrectly classified as polyps. This trial highlights the difficulties faced in handling false positives in clinical evaluations.

Another prospective RCT by Repici et al. demonstrated a significant increase in ADR with CADe assistance [36]. However, the study did not record the number of false positives, and instead reported the non-neoplastic resection rate (*i.e.*, surgical removal of non-polyp tissue). There was no statistically significant difference in non-neoplastic resection rates between the CADe and control arm using standard colonoscopy. In addition, withdrawal time, the time taken to inspect the colon while withdrawing the colonoscope following insertion, was also not significantly different. However, the clinical trial only involved expert endoscopists, and therefore the user interaction

with false positives and associated clinical implications warrants further investigation using a wider range of operators. In the current study, we recruited a wide range of expertise in order to assess potential differences in perceptions towards AI systems.

2.2 Designing for Error in AI Support

Despite the extensive developments of their capabilities, AI applications produce – and will continue to produce – incorrect assessments of some situations. Even though the body of work on AI applications has grown substantially, Dove et al. highlight a lack of discussion and research around the specific challenges of *interacting* with machine learning-based systems [10], specifically referring to the impact of false positives and false negatives on end-users. Furthermore, Dove et al. state that the majority of work in this area has focused on intelligent agents that have a physical presence or manifest themselves virtually (*e.g.*, voice assistants, robots).

A number of HCI papers explore the role of AI in the healthcare domain. In a 2019 CSCW workshop, Park et al. stress that any understanding of the application of AI in healthcare needs to "*extend beyond its technical capabilities, to consider normative, regulatory, and ethical challenges*" [33]. Park et al. state that the potential negative effects of AI on healthcare can have direct negative consequences for both patients and staff, urging the time-sensitive need for the HCI community to investigate the human role in the integration of AI-based systems.

These concerns are reflected in a number of studies, presenting insights into the ways healthcare professionals integrate AI applications into their daily work. Molin et al. are the first to systematically explore HCI considerations in the field of digital pathology [29]. Through a thematic analysis of clinicians' communication and their tools, Molin et al. propose four design considerations for digital image analysis; verification and correction, algorithmic transparency, verification on different levels of detail, and communication with clinicians [29]. Also focusing on pathology, Cai et al.'s inspirational work explores how to support clinical decision making by designing interfaces to overcome imperfect algorithmic suggestions [7]. Following the identification of pathologist needs, the authors present an interactive 'refinement' tool which allows users to identify similar images based on a variety of parameters (*e.g.*, by region, by concept) [7]. Wang et al. explore the tensions between an AI-based clinical decision support system and the rural clinical context [45]. In both studies [7, 48], interaction in the studied scenarios is intermittent rather than continuous, raising different interaction needs and difficulties to those studied in our paper. More generally, Dudley & Kristensson have reviewed user interfaces for interactive machine learning applications, highlighting the back-and-forth nature of user interaction with these systems [12]. In contrast to these works, all of which focus on intermittent interaction, our paper aims to explore AI-user interaction in a *continuous* scenario. The constraints of this setting, such as the direct control over a live video feed, the need to keep focus on the endoscopic image at all times, and, therefore, the limited amount of information that can be overlaid on the image, impose a different way of working on *e.g.* endoscopists [40]. Prior work on continuous interaction stresses that the HCI and digital health community "*need to ensure that guidelines on the design of AI systems accurately reflect user needs when the user is not necessarily the starting nor the end point of an interaction, but instead operates along a continuum*" [40].

In the medical domain, false positives can pose a range of negative consequences, including extended procedure times, unnecessary and potentially harmful interventions, and increased medical costs. We therefore set out to study the effect of false positives on doctors in a continuous setting.

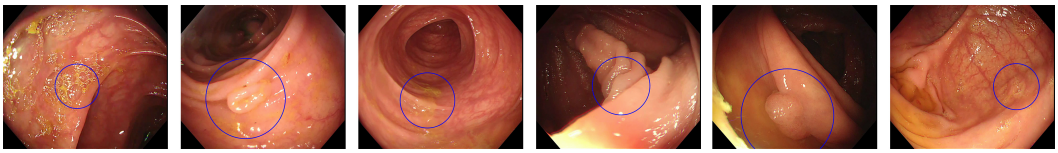

Fig. 1. Video categories from left to right: bubbles, diverticulum, mucus, wrinkled, obvious polyp, subtle polyp (best viewed in colour).

3 METHOD

To systematically study the effect of AI suggestions on medical practitioners we conducted a lab-study in which participants were presented with a total of fourteen unique videos of real-world colonoscopy footage. Through manual annotation of these videos, we mimicked an AI recommendation system which consistently overlaid selected elements of the videos with an 'AI recommendation'. We asked participants to navigate through these videos as if they were inspecting a patient, allowing our participants to navigate both forward and backward at their own chosen speed. Furthermore, we instructed participants to assess and indicate whether the object highlighted in the video is either a polyp or not a polyp. Subsequent to each interactive video, we asked participants to reflect on the role of the AI recommendation in the completed scenario. Following the completion of all scenarios, participants answered a number of open-text and multiple-choice questions to collect insights on their perceptions towards the use of AI support systems in their clinical practice.

3.1 Video materials

All videos used in the study were obtained from real-life patient footage. The videos were manually annotated for this study by a domain expert with extensive colonoscopy experience, with the annotations subsequently assessed by a second independent endoscopists. In selecting example videos and providing these annotations, our expert drew on his experience of errors encountered during initial trials of an AI-support system. Through a bespoke Python script, we rendered the AI support visuals 'on top' of the annotated areas. The design of the AI indicator (*i.e.*, a circle encompassing the annotated area, its size adjusting in alignment with the annotation) is based on earlier recommendations on continuous AI support [40]. Out of the total of 14 videos, six videos contain an actual polyp (*i.e.*, true positives). Three of these videos contain an easy to spot polyp, with the other three true positive videos containing a subtle polyp – making it more challenging to spot the polyp. As the video materials were obtained from real patients, in all six cases the polyp was removed and pathologically confirmed in the lab. The remaining eight videos contain false positives across a range of categories previously identified as being flagged by AI systems by our endoscopist collaborators and prior work [46]. These categories are: bubbles, diverticulum, mucus, and wrinkles. We summarise and briefly describe our selected videos in Table 1, and include an example frame from each category in Figure 1.

The duration of the videos is standardised to 10 seconds and they were presented to participants in randomised order. We obtained ethical approval for the use of these anonymous videos for research purposes prior to colonoscopy. For this study we did not consider the assessment of false negatives (*i.e.*, the AI system failing to highlight a polyp) as it is challenging to evaluate whether or not a participant's assessment is correct. Merely recording the frame where the participant indicates to observe a polyp is insufficient, as the participant can believe to spot the polyp in a number of different locations in the frame – requiring further manual and time-consuming annotation by the

participant. In addition, a participant indicating to have found a polyp in an unexpected location cannot be categorically refuted without a pathological assessment of the annotated area. Finally, true negatives (*i.e.*, the AI system correctly not highlighting a non-polyp) were also considered as out of scope for this paper.

3.2 Hardware and software

Participants were given a videogame controller (Xbox One) to manipulate the video feed. Using the controller's joystick, participants could control the *playback direction* of the videos. By moving the joystick to the right, the video plays forward, moving the joystick to the left plays the video backwards, returning the joystick to the neutral position pauses the video. Through the position of the joystick, participants could also control the *playback speed* of the video. Moving the joystick all the way to the right plays the video at 150% of the original playback speed (*vice versa* when moving the joystick all the way to the left). Participants were asked to classify the highlighted object in each video as either a polyp or non-polyp by pressing respectively the green (A) button or the red (B) button as soon as they arrived at their decision.

The choice for a joystick was inspired by daily clinical colonoscopy practice. By enabling participants to control both the direction and speed of the video playback, as opposed to watching a video linearly and without playback controls, we simulated the physical process of controlling the video feed during the colonoscopy procedure through withdrawal and manipulation of the endoscope. Although traditional input methods, such as keyboard or mouse, would allow us to manipulate video direction (*e.g.*, left and right arrow keys) or navigation speed (*e.g.*, duration of mouse-press on a forward button), we did not identify a traditional input method that would satisfy both these requirements simultaneously.

The controller was connected via Bluetooth to a laptop, where a bespoke *Node.js* application continuously read the current joystick and button input over a serial connection. The readings were appended to a .CSV file together with a timestamp, a randomly generated participant id, and an identifier of the current video. A web-application ran locally on the computer and presented participants with the study content. Following a pilot study with two members of the target population, we augmented the application with a progress bar, an indicator of the current joystick direction, and a visual feedback mechanism which activates when a participant presses one of the two aforementioned buttons. We used *FFmpeg* to export our AI-overlaid video files to individual frames saved as images. Based on the joystick's current value, the application incremented or

AI overlay	Video category	Nr. videos	Description
False positive	Bubbles	2	Cluster of transparent bubbles
False positive	Diverticulum	2	Small sac of tissue pushing inward or outward from the colon wall
False positive	Mucus	2	Sticky and slimy substance that can be found on the colon wall
False positive	Wrinkled	2	Creased and/or folded colon wall
True positive	Polyp - Obvious	3	Clearly visible protruding growth from the polyp wall
True positive	Polyp - Subtle	3	A flat growth from the polyp wall, difficult to identify

Table 1. Overview and description of the 14 different videos used in our study.

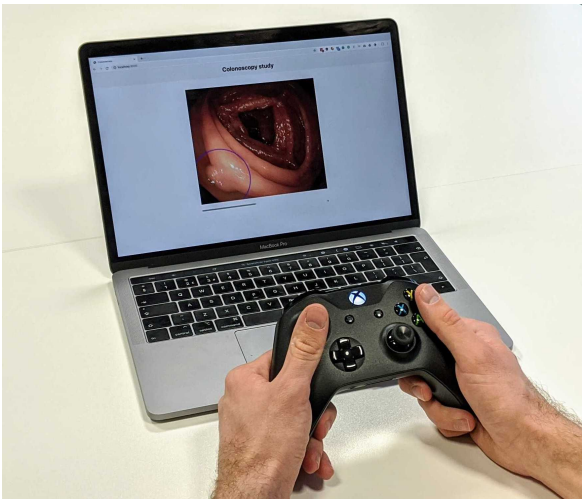

Fig. 2. Participant using a controller to navigate through patient footage.

decremented the image on display. We publicly release the source code of our application in order to support future research in this domain[1].

3.3 Recruitment and procedure

We recruited a total of 21 participants over the course of four months using a combination of mailing lists and snowball recruitment among endoscopists at our local hospital. The University College London Hospital is a major academic hospital and has one of the largest endoscopy units in the United Kingdom. We heavily relied on the connections of our endoscopist co-authors to engage this difficult to reach target group [44]. Our participant sample consisted of 7 gastroenterology consultants, 11 specialist registrars, and 3 nurse endoscopists – we summarise the sample's professional roles and their number of completed colonoscopies in Table 2. Based on their number of completed colonoscopies, we classify participants as either high experienced endoscopists or less experienced endoscopists according to the widely used threshold of 500 completed colonoscopies [35, 39]. All of our participants currently perform colonoscopies on a regular basis, and have an average age of 38.0 (SD = 7.61). Participants were not compensated for their participation in this study.

We invited participants to a designated room for individual evaluation sessions. We explained the research goal and obtained participants' informed consent prior to data collection. Participants were positioned in front of a laptop (13") with the controller placed in both hands, as is shown in Figure 2. Participants first completed a demographic survey and answered a number of questions with regards to their professional role and level of experience. Subsequently, participants were presented with instructions on how to operate the controller for navigation and were given the opportunity to interact with a 'tutorial' video until they felt comfortable to proceed. Following this, participants were presented with the aforementioned videos.

To ensure a fair and equal presentation of all video files, our application only allowed participants to proceed when the last frame was on display. We did, however, not verify whether participants pressed either the green or red green button before proceeding, as we did not want to interfere in

[1]Please see the supplemental materials.

Colonoscopies	Gastroenterology consultants	Specialist registrars	Nurse endoscopists
0-100	-	1	-
100-200	-	2	-
200-500	-	5	1
500-1000	2	2	2
1000-2500	1	1	-
>2500	4	-	-

Table 2. Overview of participant roles and their total number of performed colonoscopies. Dashed line separates experienced and less experienced endoscopists according to a threshold of 500 colonoscopies [35, 39].

the participant's browsing behaviour and reduce the ecological validity of the captured navigation data.

4 RESULTS

We now present the results of our 21 participants. First, we analyse the correctness of the participants' classifications. For each video shown to a participant, we consider only the *final* classification input (*i.e.*, polyp – green (A) button, non-polyp – red (B) button) and dismiss any preceding classification input. Figure 3 shows the overall correctness of assessment for each video category as described in Table 1. We note that for 10.5% of videos (*i.e.*, 31 out of 294 total viewed videos) participants failed to press either the green (A) or red (B) button. These missing data can be traced back primarily to one participant with missing classifications for 12 out of 14 videos, with the remaining missing data points more equally distributed between participants. For the majority of 79.8% of the videos (210 videos) one button was pressed, in 16.3% of cases (43 videos) the participants pressed the confirmation buttons twice, and for the remaining 3.8% of cases (10 videos) participants pressed more than twice. Of the 53 videos in which multiple buttons were pressed, only six videos contained a change in input (*i.e.*, classification from polyp to non-polyp or *vice versa*).

To identify the effects of video category and participant profession on participants' assessment results, we constructed a binomial (correct or incorrect) generalised linear mixed model using participant id as the random factor. The model is constructed using the *glmer* function in R package 'lme4' [5]. A likelihood ratio test as compared to the null model showed that our logistic regression model is not statistically significant ($\chi^2(17) = 25.294$, $p = 0.09$). We subsequently ran separate models

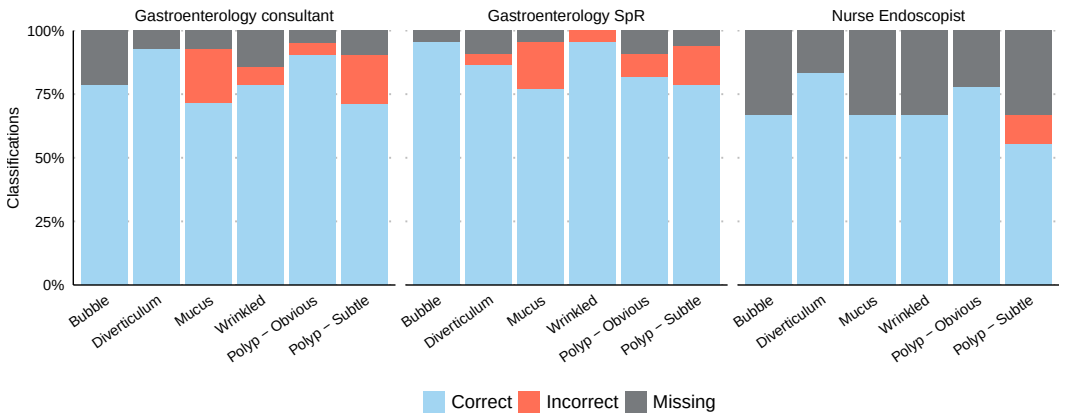

Fig. 3. Participant's polyp classifications across video categories.

containing only the video category ($\chi^2(5) = 19.121$, $p < 0.01$) and only the participant's profession ($\chi^2(2) = 2.467$, $p = 0.291$) as fixed effects. To ensure the validity of the models, we checked for the existence of multicollinearity among the models' parameters. We found that the variance inflation factor (VIF) was below the often-used threshold of ten for all our parameters [15], indicating the validity of the models. These results highlight a significant association between video category and the correct/incorrect classification of participants' assessments, but no association between their professional role and their assessment result.

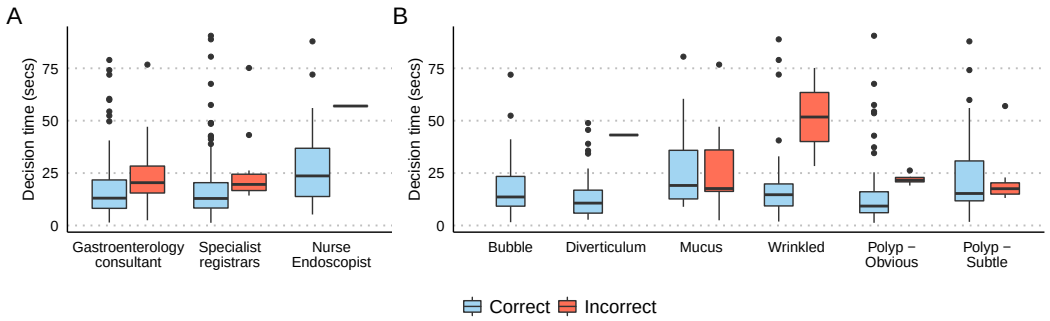

Fig. 4. Decision time for correct and incorrectly assessed videos, as split by Profession and Video category.

4.1 Viewing behaviour

We calculate participants' decision time for each video, defined as the time between the initial display of the AI recommendation and the time of the participants' first decision. For correctly classified videos, we find an average decision time of 19.4 seconds (SD = 17.4) and for incorrectly classified videos an average decision time of 27.1 seconds (SD = 19.0). To account for the non-parametric nature of the data, we conduct a Kruskal-Wallis test and identify a significant association between decision time and correctness of participants' classification, $\chi^2(1) = 9.956$, $p = 0.002$). Subsequently we assess the difference in decision time between the three professions included in our sample. Average decision time was respectively 20.1, 18.3, and 29.3 seconds for consultants, specialist registrars, and nurse endoscopists. A Kruskal-Wallis test confirms a significant association between decision time and profession ($\chi^2(2) = 9.564$, $p = 0.008$). We show the distribution of participant decision time as split between profession in Figure 4-A. An analysis of the difference in decision time between less experienced (M = 20.4 seconds, SD = 17.4) and highly experienced (M = 19.9, SD = 18.0) endoscopists reveals no significant difference between these two groups ($\chi^2(1) = 0.810$, $p = 0.368$).

Finally, we assess the difference in decision time between video categories and find a significant relation between decision time and video category ($\chi^2(5) = 24.713$, $p < 0.001$). Participants' average decision time was longest for the 'mucus' videos and shortest for the 'diverticulum' videos. Figure 4-B shows the distribution of participant decision time across video categories.

Next, we explore how participants navigated through the videos. Figure 5 shows the average *playback speed* 2.5 seconds prior to and 5 seconds following the first frame containing an AI overlay as split by true positive and false positive videos. We used Pettitt's test to inspect for a shift in the central tendency (*i.e.*, change-point detection) of the respective time series [34]. For the true positive videos we identified a significant shift in the central tendency at 0.5 seconds following the onset of the AI recommendation ($U^* = 1373$, $p < 0.001$). For the false positive videos we found a similar

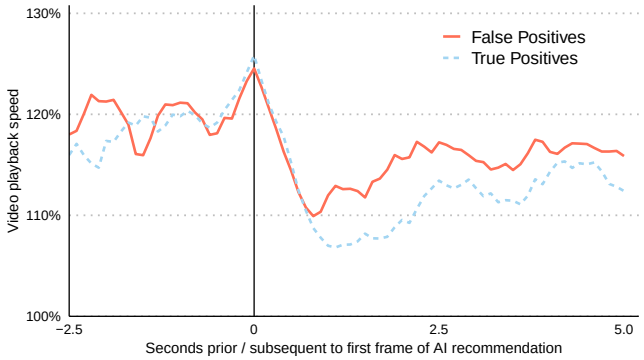

Fig. 5. Participant navigation speed, change in viewing pace before and after the first AI presentation.

shift in the central tendency at 0.3 seconds following onset of the AI recommendation ($U^* = 1275$, $p < 0.001$). These results highlight the near-simultaneous and abrupt change in participant navigation behaviour as shown in Figure 5 across both true and false positive videos. Playback speed remains lower for a longer period of time for the true positive videos.

Following each video, participants were asked to select any subsequent actions they would have taken in real life. Table 3 shows the options presented to participants and the frequency with which they were selected. Participants could select anywhere from none to all (*i.e.*, six) actions per video. Closer inspection of the area and washing of the area are the most commonly selected actions, whereas dye based chromoendoscopy (colouring of the area of interest) and deflating of the colon are most rare.

Action	**Bubble**	**Diver-ticulum**	**Mucus**	**Wrinkled**	**Polyp - Obvious**	**Polyp - Subtle**
Inspect the area more closely	73.8%	54.8%	83.3%	88.1%	76.2%	90.5%
Enhanced imaging	35.7%	35.7%	47.6%	50.0%	92.1%	85.7%
Dye based chromoendoscopy	0.0%	0.0%	7.1%	0.0%	6.3%	4.8%
Washing	88.1%	31.0%	88.1%	54.8%	58.7%	76.2%
Deflating	9.5%	9.5%	7.1%	4.8%	9.5%	6.3%
Inflating	16.7%	28.6%	19.0%	59.5%	22.2%	41.3%

Table 3. Frequency with which participants indicated to undertake various subsequent actions in a real colonoscopy, as grouped by video category.

4.2 Perceptions on AI support

Following each video, participants were asked to assess whether the AI was beneficial and whether the AI presented a hindrance to their workflow. Participants answered on a 7-point Likert scale. As seen in Figure 6, participants find the AI support most beneficial for the true-positive videos. We use the non-parametric aligned ranks transformation ANOVA to analyse our Likert responses, using the R-package *ARTool* [50]. We find a significant difference between the perceived benefit of the AI support and the video category ($F(5,288) = 16.764$, $p < 0.001$). Pairwise post hoc comparison (Tukey multiple comparisons of means) shows a significant difference between both polyp categories (*i.e.*,

'Polyp - Obvious' and 'Polyp - Subtle') and all false positive video categories ($p < 0.001$) – with the AI being perceived as significantly more beneficial in videos containing a polyp.

We repeat this analysis of Likert responses for the participants' perceived hindrance of the AI system. Our results again indicate a significant difference between self-reported hindrance of the AI and the video category ($F(5,288) = 3.605$, $p = 0.004$). Subsequent pairwise post hoc comparison shows a significant difference between the 'Polyp - Obvious' and the 'Wrinkled' video categories ($p = 0.003$) and the 'Polyp - Obvious' and the 'Bubble' video categories ($p = 0.045$), with the 'Polyp - Obvious' category reporting significantly lower hindrance levels.

Next, we analyse the participants' perceptions on AI support as grouped by their profession and their level of experience. Our sample consisted of 8 gastroenterology consultants, 11 specialist registrars, and 3 nurse endoscopists who all watched the same videos. We visualise the Likert responses of these three participant groups in Figure 7-A. We find a significant difference between the perceived AI benefit and the participant's profession ($F(2,291) = 24.948$, $p < 0.001$). Pairwise posthoc comparison (Tukey multiple comparison) reveals a significant difference between the Gastroenterology consultants and the Specialist registrars ($p < 0.001$) as well as the Gastroenterology consultants and the Nurse endoscopists ($p < 0.001$). In both cases, the Gastroenterology consults perceive the AI as less beneficial than the other professions. The Likert responses of both highly experienced ($N = 12$) and less experienced ($N = 9$) endoscopists is visualised in Figure 7-B. We find a significant difference between the perceived AI benefit and participant's experience level ($F(1,292) = 40.589$, $p < 0.001$), with participants with less experience valuing the AI as more beneficial.

Repeating the analysis for the perceived hindrance of the AI, we find a significant difference between professions ($F(2,291) = 5.180$, $p = 0.006$). A pairwise posthoc comparison shows a significant difference between Gastroenterology consultants and Specialist registrars ($p = 0.005$), with the Gastroenterology consultants reporting higher levels of perceived hindrance. Similarly, we find that endoscopists with more experience find the AI to be more of a hindrance to their workflow as compared to endoscopists with less experience ($F(2,292) = 5.689$, $p = 0.018$).

In addition to reporting the perceived benefit and hindrance of the AI, participants were also asked to provide their clinical assessment of the object highlighted in the video on a 7 point Likert scale (1 = Definitely not a polyp, 7 = Definitely a polyp). This allows us to assess the relationship between participants' clinical assessment and their perception of the AI. Given the ordinal nature of the responses we investigate this relationship using polychoric correlations, suitable for evaluating the relationship between two Likert items [25]. Similar to Pearson correlations, a polychoric

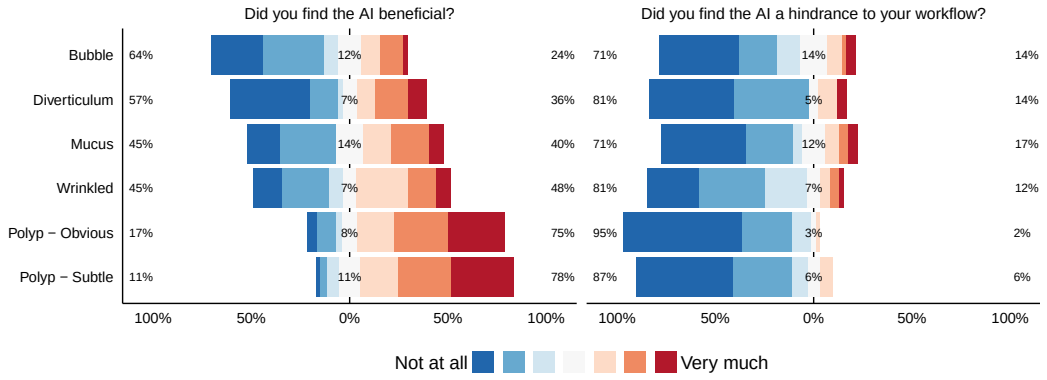

Fig. 6. Participant perceptions of the benefit and hindrance of the AI support as split by video category.

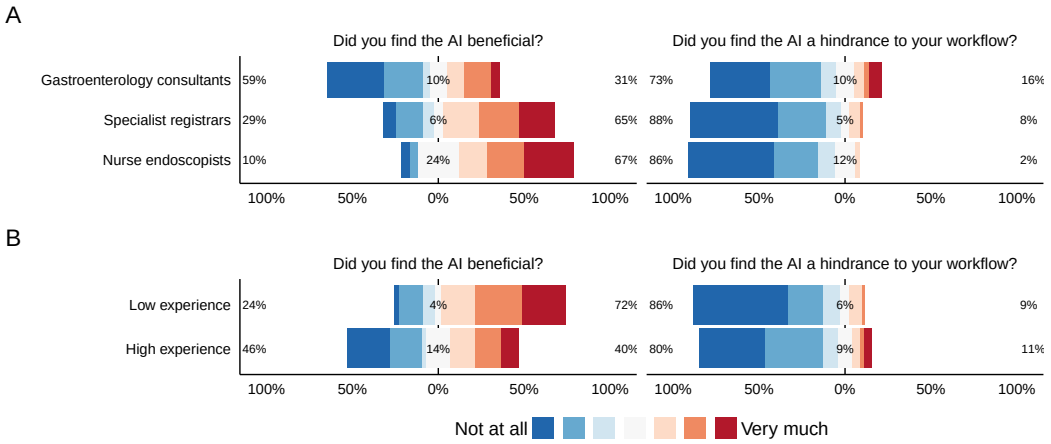

Fig. 7. Perceived benefit and hindrance of the AI support as split by Profession and Experience.

correlation value ranges between -1 and 1, with the extreme values indicating a perfect linear relationship and 0 indicating no linear relationship between the two variables. We find a positive correlation coefficient of 0.52 for the relationship between the participants clinical assessment and the perceived benefit of the AI, indicating that the perceived benefit of the AI increases the more certain participants are that a polyp is shown in the video. For the relationship between perceived hindrance of the AI and the participants' clinical assessment of the highlighted object we find a weak negative correlation of -0.20. This indicates that the perceived hindrance slightly decreases in line with the assessment of participants that the object is a polyp.

4.2.1 Participant responses. When asked to reflect on the AI support in a completion survey, participants highlighted both positive and negative aspects of this technology. From a positive perspective, participants highlighted how AI support could ensure that operators take a second look during colonoscopy at areas they might have otherwise skipped. "*The AI makes me question my initial diagnosis of an area, and this may be beneficial. It definitely acts as a deterrent against lazy endoscopy and makes you double check areas which might otherwise be wrongly considered normal.*" (P14, Nurse Endoscopist). In line with our results on perceived benefit in Figure 7, participants noted that the support might be most beneficial to those with limited experience; "*Definitely would help novice user. Experts is more questionable.*" (P04, Gastroenterology SpR).

Participants raised concerns around the number of false positives highlighted by an AI system, which are perceived as annoying and interrupting, and could lead to users disabling AI support altogether; "*It is important not to have too many false positives otherwise the endoscopist will stop using the system*" (P10, Gastroenterology consultant). Similarly, participants note that a high number of false positives would increase procedure time – although not all participants believe this is necessarily a bad thing; "*Areas falsely flagged as polyps are still areas for washing and closer inspection which achieves objective of thorough endoscopy.*" (P09, Gastroenterology SpR). Highlighting the delicate balance between the positive effect AI may have in identifying more polyps and the potential negative consequences on medical staff, one of the participants points to the increased (mental) effort required by incorrect AI suggestions; "*Knowing that optical diagnosis is imperfect, I would worry when disregarding an area the AI has flagged as potentially abnormal in case I am wrong. I am a naturally cautious practitioner and worry the extra inspection of potentially normal*

areas would add a lot of time as I try to really satisfy myself it is the machine and not myself making the incorrect call." (P14, Nurse Endoscopist).

Participants reflected positively on the use of a gamepad controller, with 17 participants specifically commenting on the intuitiveness and ease of use of the controller without being specifically asked to do so. Furthermore, two participants positively contrast the controller with the use of a keyboard for colon video navigation, as used in *e.g.* training material; "*Gamepad effective as reduces the distraction of a keyboard in front of the screen.*" (P09, Gastroenterology SpR).

5 DISCUSSION

Future deployments of AI technologies, including those used during clinical examinations and surgeries, will inevitably present users with incorrect classifications and recommendations. This raises questions regarding the end-user acceptance and trust towards these systems [21, 37], (legal) accountability in case of errors [30, 38], and the impact of AI support systems on user interaction. This work presents colonoscopy as a case study for understanding the effects of imperfect AI support on end-users during continuous interactions.

Our results reflect that the identification of adenomas – the primary goal of a colonoscopy – is a challenging task, with participants incorrectly classifying videos containing a polyp in 14.1% of cases (obvious and subtle videos combined). Recent meta-analyses on adenoma identification rates highlight substantial miss rates during patient inspection, with averages ranging between 22 and 27% [43, 52]. As our videos were only 10 seconds in length and clearly indicated the object of interest, a lower error rate than observed in real patient examinations is expected. These results highlight that, even if a polyp is identified by an AI support system, endoscopists may still incorrectly classify the object as a non-polyp.

5.1 False positives: perceptions and behaviour

Our results show that the participants' perceived benefit of AI significantly increases substantially for true positive videos, whereas the increase in the perceived hindrance among false positive videos is much lower (Section 4.2). As seen in Figure 6, false positives are not necessarily perceived as a bad thing. Participants noted that even if the AI points them in the direction of an object of interest that turns out not to be a polyp it would still be worth the extra inspection time. The AI support is, therefore, not necessarily perceived as a system used solely for the identification of polyps, but as a more generic system that can point out objects requiring further inspection – *i.e.*, highlighting 'red flags' during inspection. Our results align with earlier considerations on AI support, with Zachariah et al. stating that "*we should embrace CADe [computer-aided detection] systems only as a 'second observer,' one that questions us: 'what is this; is it important?'*" [51]. We note that these perceptions may change were a system to be integrated long-term in the participants' workflow.

Although participants indicated that they do not perceive false positives as a major hindrance, our results highlight that their presence may significantly affect the viewing behaviour of clinicians. As seen in Figure 5, participants' navigation speed is significantly reduced upon seeing a false positive. This could lead to prolonged examination times and subsequently increase procedure cost. While this depends on the frequency with which false positives are presented, recent work shows that current iterations of AI frequently flag false positives [27]. Figure 4 indicates that the required time for participants to dismiss a false positive can be well over a minute for some participants, with significant differences in decision time between video categories. Whereas the colonoscopy literature has identified ambitious and challenging goals for future AI systems, *e.g.* "*real time (<10 ms latency), easy to implement, reliable, provide near 100% sensitivity, and a nondistracting low false positive rate*" [51], we argue that, based on our results, metrics such as a system's overall

false positive rate may not be sufficiently informative to predict real-world world performance and interaction. Instead, developers of AI support systems should prioritise the reduction of false positives in areas that are most responsible for error and delays during actual use.

5.2 Studying continuous AI support applications in healthcare

A critical concern during the design of our study was to uphold the study's ecological validity while ensuring a balanced and controlled evaluation protocol, avoiding patient harm, and ensuring a sufficient sample size. Dove et al. described the difficulties of working with machine learning as a 'design material' [10], citing the unpredictability of AI prototypes as a barrier to systematic evaluation. To overcome these challenges, we augmented real-world colonoscopy footage with realistic AI support based on manual annotations. Furthermore, we considered it critical for the ecological validity of our study that participants were able to control both the direction and speed of the video playback – similar to the navigation of the endoscope during colonoscopy. As shown in Figure 5, this allowed participants to navigate freely – similar to how an endoscopist would inspect a colon in real-life while being provided continuous AI support. By preparing our study material in advance, we are able to overcome some of the challenges highlighted by Dove et al. in studying and designing for Human-AI interaction [10].

Our approach stands in stark contrast with evaluations in which participants are shown a video and asked to press a button when they believe a polyp appears in the video (see *e.g.* [18]), which are unable to capture the users' real-world navigation behaviour. At the same time, we highlight that our approach only approximates reality, as participants do not have complete control over the endoscope - being unable to inflate the colon or wash away any debris. This aligns with a recent call in the literature, which urges researchers to increase their studies' ecological validity while simultaneously acknowledging the necessary compromises and their effects on the presented results [41]. For this study, an evaluation with a live AI support system that runs during a routine hospital operation would have provided the most valid observation data. However, such a study would raise major ethical concerns with regards to patient safety (given the large number of false positives and the unknown effects on medical outcomes), as well as removing our ability to systematically compare different video categories and AI overlays (true positive, false positive). As shown in Table 3, participants frequently indicated that they would have taken further actions to inspect the area in more detail in a real life scenario. Whether or not they would do so in clinical practice, especially when confronted with a high false positive rate, cannot be assessed through the present study.

The recruitment of medical professionals as study participants is challenging, *inter alia* due to their busy schedules [44]. This is especially true for non-survey research, as we had to ensure participants' continuous availability for upwards of 20 minutes. Based on our experiences, we identify three aspects that were indispensable in the recruitment of participants. First, a close collaboration with gastroenterologists enabled access to professional circles which would normally be out of our reach. Second, the presence of a physical artefact (*i.e.*, the game controller) sparked interest among potential participants. Subsequent participant responses highlight that they considered this as both a suitable and engaging instrument for data collection. Third, we repeatedly visited the hospital and set up our study equipment in order to ensure that we could immediately commence data collection whenever a participant became available. We were less successful in participant recruitment at a national colonoscopy conference, in which attendees were more interested to utilise breaks between presentations for networking and other activities.

5.3 Professional role and AI perception

The role of AI support systems for endoscopy is a topic of active discussion within the colonoscopy literature [27, 30, 51]. While the literature discusses important questions regarding the integration of AI systems [2], such as a potential decrease in polyp recognition skills [51] and legal concerns when not following AI recommendations [30], the role of the human operator and their interaction with the AI system remains both under-discussed and under-explored. Our results highlight that participants with less endoscopic experience generally perceived the AI as more beneficial (Figure 6). The sentiment that those with less experiences have the most to gain from AI support is repeated both in participant responses and the wider literature [11, 14].

The analyses across different professional roles highlight that no unanimous response to the introduction of AI systems can be expected among end-users, with differences emerging in both the interactions (Figure 4) and perspectives (Figure 6) between different endoscopic roles. As such, we urge for AI support systems to be evaluated with a diverse range of participants prior to their deployment.

5.4 Limitations and future work

We recognise and discuss a number of limitations to the presented work. First, we solely considered false positives and true positives whereas real-world deployments will inevitably demonstrate other type of flaws, most critically failing to identify polyps (*i.e.*, false negatives). The consequences of this can be substantial if attention levels of endoscopists were to drop due to continuous AI support [51]. However, the effect of false negatives is outside of the scope of our study's focus on the participants' interaction with AI recommendations. Collecting participant input on false negatives requires not only a binary 'polyp/non-polyp' decision, but will also need subsequent manual annotation in order to assess whether the participant has indeed correctly identified a polyp (see Section 3.1).

The videos included in our experiment were manually annotated and are not the result of a 'live' AI system. This ensured consistency in the material presented to participants, thereby enabling us to compare participant results. As our video material contained a wide range of both false positives and true positive annotations, participants' perception of the AI system may have been affected by the contrast between surprisingly poor and superior AI 'detection' performance. While this was in line with our study's focus on initial responses to AI-support, future work on long-term AI perception must integrate actual AI systems.

We believe this to be a sensible consideration given previously raised concerns on maintaining experimental control when working with AI technology [10]. Informed by prior work [46] and through extensive collaboration with gastroenterologists, we were able to identify the most commonly occurring false positives in current AI systems (see Table 1). However, we stress that not all possible types of false positives were included in our study (*e.g.*, undigested debris).

Future work should aim to explore the (long-term) impact of embedding AI technology in clinical practice – including the potential over-reliance on AI support and a decrease in user trust and usage when faced with repeated false positives. We have made the source code of our application publicly available in order to support future researchers in the systematic evaluation of continuous AI support scenarios.

6 CONCLUSION

In this paper we report on a controlled study in which we investigated the behaviour and experiences of endoscopists through a continuous support application. Through the use of expert annotated colonoscopy videos and a videogame controller, we were able to capture the navigation

and decision behaviour of 21 expert end-users. Our work highlights that AI recommendations significantly slow down participant navigation, regardless of the content of the object highlighted. Yet, time for participants to make a decision on the nature of the highlighted object did differ significantly between video categories. We therefore argue that a single metric of AI performance is insufficient to assess real-world impact on user-interaction, as the effect of false positives on end-users differs between clinical content presented. Furthermore, our results highlight that the participant's professional role and experience significantly affected viewing behaviour and perceptions towards AI systems. Development and evaluation of AI applications should therefore carefully consider the full breadth of end-users who will interact with the technology. Finally, we highlight and discuss the challenges faced by researchers aiming to study AI support in continuous support scenarios. Maintaining sufficient levels of ecological validity should be a key consideration in Human-AI interaction studies going forward.

ACKNOWLEDGMENTS

We are grateful to the participants for their time in participating in our study, and acknowledge the financial support of the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [NS/A000050/1] at UCL.

REFERENCES

[1] Omer F Ahmad, Yuichi Mori, Masashi Misawa, Shin-Ei Kudo, John T Anderson, Jorge Bernal, Tyler M Berzin, Raf Bisschops, Michael F Byrne, Peng-Jen Chen, James E East, Tom Eelbode, Daniel S Elson, Suryakanth R Gurudu, Aymeric Histace, William E Karnes, Alessandro Repici, Rajvinder Singh, Pietro Valdastri, Michael B Wallace, Pu Wang, Danail Stoyanov, and Laurence B Lovat. 2020. Establishing key research questions for the implementation of artificial intelligence in colonoscopy: a modified Delphi method. *Endoscopy* (2020). https://doi.org/10.1055/a-1306-7590

[2] Omer F. Ahmad, Danail Stoyanov, and Laurence B. Lovat. 2020. Barriers and pitfalls for artificial intelligence in gastroenterology: Ethical and regulatory issues. *Techniques and Innovations in Gastrointestinal Endoscopy* 22, 2 (2020), 80–84. https://doi.org/10.1016/j.tgie.2019.150636

[3] Sang Bong Ahn, Dong Soo Han, Joong Ho Bae, Tae Jun Byun, Jong Pyo Kim, and Chang Soo Eun. 2012. The Miss Rate for Colorectal Adenoma Determined by Quality-Adjusted, Back-to-Back Colonoscopies. *Gut and liver* 6, 1 (2012), 64–70. https://doi.org/10.5009/gnl.2012.6.1.64

[4] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. 2019. Guidelines for Human-AI Interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Article 3, 13 pages. https://doi.org/10.1145/3290605.3300233

[5] Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software* 67, 1 (2015), 1–48. https://doi.org/10.18637/jss.v067.i01

[6] A. M. Buchner, M. W. Shahid, M. G. Heckman, N. N. Diehl, R. B. McNeil, P. Cleveland, K. R. Gill, A. Schore, M. Ghabril, M. Raimondo, S. A. Gross, and M. B. Wallace. 2011. Trainee participation is associated with increased small adenoma detection. *Gastrointest. Endosc.* 73, 6 (2011), 1223–1231.

[7] Carrie J. Cai, Emily Reif, Narayan Hegde, Jason Hipp, Been Kim, Daniel Smilkov, Martin Wattenberg, Fernanda Viegas, Greg S. Corrado, Martin C. Stumpe, and Michael Terry. 2019. Human-Centered Tools for Coping with Imperfect Algorithms During Medical Decision-Making. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Article 4, 14 pages. https://doi.org/10.1145/3290605.3300234

[8] Douglas A. Corley, Christopher D. Jensen, Amy R. Marks, Wei K. Zhao, Jeffrey K. Lee, Chyke A. Doubeni, Ann G. Zauber, Jolanda de Boer, Bruce H. Fireman, Joanne E. Schottinger, Virginia P. Quinn, Nirupa R. Ghai, Theodore R. Levin, and Charles P. Quesenberry. 2014. Adenoma Detection Rate and Risk of Colorectal Cancer and Death. *New England Journal of Medicine* 370, 14 (2014), 1298–1306. https://doi.org/10.1056/NEJMoa1309086

[9] Thomas Davenport and Ravi Kalakota. 2019. The potential for artificial intelligence in healthcare. *Future healthcare journal* 6, 2 (2019), 94–98. https://doi.org/10.7861/futurehosp.6-2-94

[10] Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. 2017. UX Design Innovation: Challenges for Working with Machine Learning as a Design Material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17)*. 278–288. https://doi.org/10.1145/3025453.3025739

[11] Stephan Dreiseitl and Michael Binder. 2005. Do physicians value decision support? A look at the effect of decision support systems on physician opinion. *Artificial Intelligence in Medicine* 33, 1 (2005), 25–30. https://doi.org/10.1016/j.artmed.2004.07.007

[12] John J. Dudley and Per Ola Kristensson. 2018. A Review of User Interface Design for Interactive Machine Learning. *ACM Trans. Interact. Intell. Syst.* 8, 2, Article 8 (2018), 37 pages. https://doi.org/10.1145/3185517

[13] M. K. El-Najdawi and Anthony C. Stylianou. 1993. Expert Support Systems: Integrating AI Technologies. *Commun. ACM* 36, 12 (1993), 55–ff. https://doi.org/10.1145/163298.163306

[14] Charles P. Friedman, Arthur S. Elstein, Fredric M. Wolf, Gwendolyn C. Murphy, Timothy M. Franz, Paul S. Heckerling, Paul L. Fine, Thomas M. Miller, and Vijoy Abraham. 1999. Enhancement of Clinicians' Diagnostic Reasoning by Computer-Based Consultation - A Multisite Study of 2 Systems. *JAMA* 282, 19 (1999), 1851–1856. https://doi.org/10.1001/jama.282.19.1851

[15] Joseph F Hair, William C Black, Barry J Babin, Rolph E Anderson, and RL Tatham. 2010. *Multivariate Data Analysis*. Pearson.

[16] Russell P. Harris, Stacey L. Sheridan, Carmen L. Lewis, Colleen Barclay, Maihan B. Vu, Christine E. Kistler, Carol E. Golin, Jessica T. DeFrank, and Noel T. Brewer. 2014. The Harms of Screening: A Proposed Taxonomy and Application to Lung Cancer Screening. *JAMA Internal Medicine* 174, 2 (2014). https://doi.org/10.1001/jamainternmed.2013.12745

[17] Daniel A. Hashimoto, Guy Rosman, Daniela Rus, and Ozanan R. Meireles. 2018. Artificial Intelligence in Surgery: Promises and Perils. *Annals of surgery* 268, 1 (2018), 70–76. https://doi.org/10.1097/SLA.0000000000002693

[18] Cesare Hassan, Michael B Wallace, Prateek Sharma, Roberta Maselli, Vincenzo Craviotto, Marco Spadaccini, and Alessandro Repici. 2020. New artificial intelligence system: first validation study versus experienced endoscopists for colorectal polyp detection. *Gut* 69, 5 (2020), 799–800. https://doi.org/10.1136/gutjnl-2019-319914

[19] Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts. 2018. Artificial intelligence in radiology. *Nature Reviews Cancer* 18, 8 (2018), 500–510. https://doi.org/10.1038/s41568-018-0016-5

[20] Michal F. Kaminski, Jaroslaw Regula, Ewa Kraszewska, Marcin Polkowski, Urszula Wojciechowska, Joanna Didkowska, Maria Zwierko, Maciej Rupinski, Marek P. Nowacki, and Eugeniusz Butruk. 2010. Quality Indicators for Colonoscopy and the Risk of Interval Cancer. *New England Journal of Medicine* 362, 19 (2010), 1795–1803. https://doi.org/10.1056/NEJMoa0907667

[21] Rafal Kocielnik, Saleema Amershi, and Paul N. Bennett. 2019. Will You Accept an Imperfect AI? Exploring Designs for Adjusting End-User Expectations of AI Systems *(CHI '19)*. 1–14. https://doi.org/10.1145/3290605.3300641

[22] Jennifer Elston Lafata, Janine Simpkins, Lois Lamerato, Laila Poisson, George Divine, and Christine Cole Johnson. 2004. The Economic Impact of False-Positive Cancer Screens. *Cancer Epidemiology and Prevention Biomarkers* 13, 12 (2004), 2126–2132. https://cebp.aacrjournals.org/content/13/12/2126

[23] Chang Kyun Lee, Dong Il Park, Suck-Ho Lee, Young Hwangbo, Chang Soo Eun, Dong Soo Han, Jae Myung Cha, Bo-In Lee, and Jeong Eun Shin. 2011. Participation by experienced endoscopy nurses increases the detection rate of colon polyps during a screening colonoscopy: a multicenter, prospective, randomized study. *Gastrointestinal Endoscopy* 74, 5 (2011), 1094–1102. https://doi.org/10.1016/j.gie.2011.06.033

[24] Yeonjoo Lee, Miyeon Ha, Sujeong Kwon, Yealin Shim, and Jinwoo Kim. 2019. Egoistic and altruistic motivation: How to induce users' willingness to help for imperfect AI. *Computers in Human Behavior* 101 (2019), 180–196. https://doi.org/10.1016/j.chb.2019.06.009

[25] Patrick Mair. 2018. *Factor Analysis*. Springer International Publishing, 17–61. https://doi.org/10.1007/978-3-319-93177-7_2

[26] Monica S. Millan, Perita Gross, Elena Manilich, and James M. Church. 2008. Adenoma Detection Rate: The Real Indicator of Quality in Colonoscopy. *Diseases of the Colon & Rectum* 51, 8 (2008), 1217–1220. https://doi.org/10.1007/s10350-008-9315-3

[27] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Tomonari Cho, Shinichi Kataoka, Akihiro Yamauchi, Yushi Ogawa, Yasuharu Maeda, Kenichi Takeda, Katsuro Ichimasa, et al. 2018. Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience. *Gastroenterology* 154, 8 (2018), 2027–2029.

[28] Masashi Misawa, Shin-ei Kudo, Yuichi Mori, Tomonari Cho, Shinichi Kataoka, Akihiro Yamauchi, Yushi Ogawa, Yasuharu Maeda, Kenichi Takeda, Katsuro Ichimasa, Hiroki Nakamura, Yusuke Yagawa, Naoya Toyoshima, Noriyuki Ogata, Toyoki Kudo, Tomokazu Hisayuki, Takemasa Hayashi, Kunihiko Wakamura, Toshiyuki Baba, Fumio Ishida, Hayato Itoh, Holger Roth, Masahiro Oda, and Kensaku Mori. 2018. Artificial Intelligence-Assisted Polyp Detection for Colonoscopy: Initial Experience. *Gastroenterology* 154, 8 (2018), 2027–2029.e3. https://doi.org/10.1053/j.gastro.2018.04.003

[29] Jesper Molin, Paweł W. Woundefinedniak, Claes Lundström, Darren Treanor, and Morten Fjeld. 2016. Understanding Design for Automated Image Analysis in Digital Pathology. In *Proceedings of the 9th Nordic Conference on Human-Computer Interaction (NordiCHI '16)*. Association for Computing Machinery, Article 58, 10 pages. https://doi.org/10.1145/2971485.2971561

[30] Y. Mori, S. E. Kudo, T. M. Berzin, M. Misawa, and K. Takeda. 2017. Computer-aided diagnosis for colonoscopy. *Endoscopy* 49, 8 (2017), 813–819.

[31] Ahmed Nait Aicha, Gwenn Englebienne, Kimberley van Schooten, Mirjam Pijnappels, and Ben Kröse. 2018. Deep Learning to Predict Falls in Older Adults Based on Daily-Life Trunk Accelerometry. *Sensors* 18, 5 (2018), 1654. https://doi.org/10.3390/s18051654

[32] Muhammad Khalid Khan Niazi, Anil V Parwani, and Metin N Gurcan. 2019. Digital pathology and artificial intelligence. *The Lancet Oncology* 20, 5 (2019), e253 – e261. https://doi.org/10.1016/S1470-2045(19)30154-8

[33] Sun Young Park, Pei-Yi Kuo, Andrea Barbarin, Elizabeth Kaziunas, Astrid Chow, Karandeep Singh, Lauren Wilcox, and Walter S. Lasecki. 2019. Identifying Challenges and Opportunities in Human-AI Collaboration in Healthcare. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing (CSCW '19)*. 506–510. https://doi.org/10.1145/3311957.3359433

[34] A. N. Pettitt. 1979. A Non-Parametric Approach to the Change-Point Problem. *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28, 2 (1979), 126–135. https://doi.org/10.2307/2346729

[35] L. Preisler, M. B. Søndergaard Svendsen, B. Søndergaard, L. Brink, T. Nordentoft, L. B. Svendsen, and L. Konge. 2016. Automatic and unbiased assessment of competence in colonoscopy: exploring validity of the Colonoscopy Progression Score (CoPS). *Endosc Int Open* 4, 12 (2016), E1238–E1243.

[36] Alessandro Repici, Matteo Badalamenti, Roberta Maselli, Loredana Correale, Franco Radaelli, Emanuele Rondonotti, Elisa Ferrara, Marco Spadaccini, Asma Alkandari, Alessandro Fugazza, Andrea Anderloni, Piera Alessia Galtieri, Gaia Pellegatta, Silvia Carrara, Milena Di Leo, Vincenzo Craviotto, Laura Lamonaca, Roberto Lorenzetti, Alida Andrealli, Giulio Antonelli, Michael Wallace, Prateek Sharma, Thomas Rosch, and Cesare Hassan. 2020. Efficacy of Real-Time Computer-Aided Detection of Colorectal Neoplasia in a Randomized Trial. *Gastroenterology* (2020). https://doi.org/10.1053/j.gastro.2020.04.062

[37] Mike Schaekermann, Graeme Beaton, Elaheh Sanoubari, Andrew Lim, Kate Larson, and Edith Law. 2020. Ambiguity-Aware AI Assistants for Medical Data Analysis. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. 1–14. https://doi.org/10.1145/3313831.3376506

[38] Alison Smith-Renner, Ron Fan, Melissa Birchfield, Tongshuang Wu, Jordan Boyd-Graber, Daniel S. Weld, and Leah Findlater. 2020. No Explainability without Accountability: An Empirical Study of Explanations and Feedback in Interactive ML. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. 1–13. https://doi.org/10.1145/3313831.3376624

[39] B. J. Spier, M. Benson, P. R. Pfau, G. Nelligan, M. R. Lucey, and E. A. Gaumnitz. 2010. Colonoscopy training in gastroenterology fellowships: determining competence. *Gastrointest. Endosc.* 71, 2 (2010), 319–324.

[40] Niels van Berkel, Omer F. Ahmad, Danail Stoyanov, Laurence Lovat, and Ann Blandford. 2021. Designing Visual Markers for Continuous Artificial Intelligence Support: A Colonoscopy Case Study. *ACM Trans. Comput. Healthcare* 2, 1, Article 7 (2021), 24 pages. https://doi.org/10.1145/3422156

[41] Niels van Berkel, Matthew J. Clarkson, Guofang Xiao, Eren Dursun, Moustafa Allam, Brian R. Davidson, and Ann Blandford. 2020. Dimensions of ecological validity for usability evaluations in clinical settings. *Journal of Biomedical Informatics* 110 (2020), 103553. https://doi.org/10.1016/j.jbi.2020.103553

[42] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 28 (Nov. 2019), 21 pages. https://doi.org/10.1145/3359130

[43] Jeroen C. van Rijn, Johannes B. Reitsma, Jaap Stoker, Patrick M. Bossuyt, Sander J. van Deventer, and Evelien Dekker. 2006. Polyp Miss Rate Determined by Tandem Colonoscopy: A Systematic Review. *American Journal of Gastroenterology* 101, 2 (2006).

[44] Jonathan B. VanGeest, Timothy P. Johnson, and Verna L. Welch. 2007. Methodologies for Improving Response Rates in Surveys of Physicians: A Systematic Review. *Evaluation & the Health Professions* 30, 4 (2007), 303–321. https://doi.org/10.1177/0163278707307899

[45] Dakuo Wang, Liuping Wang, Zhan Zhang, Ding Wang, Haiyi Zhu, Yvonne Gao, Xiangmin Fan, and Feng Tian. 2021. "Brilliant AI Doctor" in Rural Clinics: Challenges in AI-Powered Clinical Decision Support System Deployment. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Article 697, 18 pages. https://doi.org/10.1145/3411764.3445432

[46] Pu Wang, Tyler M Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, et al. 2019. Real-time automatic detection system increases colonoscopic polyp and adenoma detection rates: a prospective randomised controlled study. *Gut* 68, 10 (2019), 1813–1819. https://doi.org/10.1136/gutjnl-2018-317500

[47] Pu Wang, Xiaogang Liu, Tyler M. Berzin, Jeremy R. Glissen Brown, Peixi Liu, Chao Zhou, Lei Lei, Liangping Li, Zhenzhen Guo, Shan Lei, Fei Xiong, Han Wang, Yan Song, Yan Pan, and Guanyu Zhou. 2020. Effect of a deep-learning computer-aided detection system on adenoma detection during colonoscopy (CADe-DB trial): a double-blind randomised study.

The Lancet Gastroenterology & Hepatology 5, 4 (2020), 343–351. https://doi.org/10.1016/S2468-1253(19)30411-X

[48] Pu Wang, Xiao Xiao, Jeremy R. Glissen Brown, Tyler M. Berzin, Mengtian Tu, Fei Xiong, Xiao Hu, Peixi Liu, Yan Song, Di Zhang, Xue Yang, Liangping Li, Jiong He, Xin Yi, Jingjia Liu, and Xiaogang Liu. 2018. Development and validation of a deep-learning algorithm for the detection of polyps during colonoscopy. *Nature Biomedical Engineering* 2, 10 (2018), 741–748. https://doi.org/10.1038/s41551-018-0301-3

[49] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, 1–14. https://doi.org/10.1145/3313831.3376813

[50] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. 143–146. https://doi.org/10.1145/1978942.1978963

[51] Robin Zachariah, Andrew Ninh, and William Karnes. 2019. Artificial intelligence for colon polyp detection: Why should we embrace this? *Techniques in Gastrointestinal Endoscopy* (2019), 150631. https://doi.org/10.1016/j.tgie.2019.150631

[52] Shengbing Zhao, Shuling Wang, Peng Pan, Tian Xia, Xin Chang, Xia Yang, Liliangzi Guo, Qianqian Meng, Fan Yang, Wei Qian, Zhichao Xu, Yuanqiong Wang, Zhijie Wang, Lun Gu, Rundong Wang, Fangzhou Jia, Jun Yao, Zhaoshen Li, and Yu Bai. 2019. Magnitude, Risk Factors, and Factors Associated With Adenoma Miss Rate of Tandem Colonoscopy: A Systematic Review and Meta-analysis. *Gastroenterology* 156, 6 (2019), 1661–1674.e11. https://doi.org/10.1053/j.gastro.2019.01.260