Designing Visual Markers for Continuous Artificial Intelligence Support: A Colonoscopy Case Study

NIELS VAN BERKEL, UCL Interaction Centre, United Kingdom and Aalborg University, Denmark
OMER F AHMAD, Wellcome/EPSRC Centre for Interventional & Surgical Sciences, University College London, United Kingdom
DANAIL STOYANOV, Medical Physics and Bioengineering, University College London, United Kingdom
LAURENCE LOVAT, University College London Hospitals, United Kingdom
ANN BLANDFORD, UCL Interaction Centre, United Kingdom

Colonoscopy, the visual inspection of the large bowel using an endoscope, offers protection against colorectal cancer by allowing for the detection and removal of pre-cancerous polyps. The literature on polyp detection shows widely varying miss rates among clinicians, with averages ranging around 22–27%. While recent work has considered the use of AI support systems for polyp detection, how to visualise and integrate these systems into clinical practice is an open question. In this work, we explore the design of visual markers as used in an AI support system for colonoscopy. Supported by the gastroenterologists in our team, we designed seven unique visual markers and rendered them on real-life patient video footage. Through an online survey targeting relevant clinical staff (N = 36), we evaluated these designs and obtained initial insights and understanding into the way in which clinical staff envision AI to integrate in their daily work-environment. Our results provide concrete recommendations for the future deployment of AI support systems in continuous, adaptive scenarios.

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

Additional Key Words and Phrases: human-centered AI, video, endoscopy, continuous AI, markers, annotation, visualisation, support system, artificial intelligence, AI, machine learning, ML, colonoscopy, medical imaging

1 INTRODUCTION

Colonoscopy is a medical procedure in which a trained endoscopist evaluates the inside of a patient's colon. This procedure is carried out using a long, flexible tube (colonoscope) containing a camera, light, and, when necessary, small instruments (*e.g.*, for removing polyps). Colonoscopy is widely considered to be the gold standard for investigation of the large bowel [45], and allows for simultaneous diagnosis and, if required, removal of bowel polyps. Colonoscopy is an important screening instrument for the prevention and detection of colorectal cancer, also known as bowel cancer. Colorectal cancer is fairly common, with an estimated yearly total of 1.8 million new cases [7]; it is the second leading cause of cancer-related death both globally and within the US [7, 24]. Although polyps typically grow slowly, the removal of precancerous lesions and the early detection of cancers is key in increasing survival rate among patients [43].

Recent clinical work has highlighted that a high number of polyps are being missed, with meta-analyses reporting averages between 22 and 27% [51, 61] – as determined through tandem colonoscopy. Polyps can be missed during the procedure for a variety of reasons, *e.g.*, shifts in operator attention, partial concealment, or obstructed from view due to bowel contents [31]. To assist endoscopists in the challenging task of identifying polyps, novel technologies – sometimes referred to as Computer-Aided Diagnosis (CAD) systems – have been

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

developed to automatically detect polyps based on the colonoscope's video source [1, 36, 41]. While promising in their ability to detect polyps, these applications raise questions on the design and integration of Artificial Intelligence (AI) support systems in daily clinical practice. Previous work has highlighted that problems introduced by a lack of HCI considerations in deploying healthcare technology may result in abandonment by the end-user [37]. Recent work highlights that additional concerns arise when considering the use of AI in day-to-day practice [9, 59], where incorrect and missing recommendations can cause patient harm.

To inform the design of a specific aspect of AI support systems, namely the visual marker used to highlight an object of interest, we systematically explore the design space of visual markers in live video streams. Based on many conversations with gastroenterologists, we established that they consider this to be a crucial element of the interaction with future AI systems. In this paper, we identify distinct visualisations currently adopted in the literature, and devise alternative designs using the literature on visual attention [58] and designing for AI systems [2] in a collaborative effort with gastroenterologists. The aim of this paper is not to extend the visual attention literature, but instead to assess how different designs of visual markers are perceived by clinical staff in the context of continuous AI support. We conceptualise a total of seven distinct visual markers and develop these designs into operational video markers. Through an online survey (N = 36), we present these designs as overlaid on patients' colonoscopy video footage to endoscopists and assistants. Participants rated our designs on a number of aspects, including their ability to locate polyps as well as their potential to interfere with clinical work. Furthermore, we investigated the perceptions of the clinical staff in relation to the (future) integration of AI-based support systems into clinical practice.

Our results indicate that currently used visual markers are perceived as significantly less useful in detecting and locating polyps in comparison to some of our alternative designs. We identify clear preferences towards marker colour and system integration. We contrast our findings of AI support in a continuously adapting context (*e.g.*, video footage) to recent design guidelines and recommendations for AI interaction as brought forward by the HCI community. Our work contributes to these guidelines by highlighting the disparate requirements of users in continuously adapting scenarios.

2 RELATED WORK

The Global Cancer Observatory[1] lists a yearly 1.8 million new cases of colorectal cancer globally (colorectal groups cancers in the colon, rectum, and anus – all of which are inspected during colonoscopy) [7]. Surpassed only by lung cancer, an estimated 880.000 people die as a direct result of colorectal cancers (9.2% of total cancers). Furthermore, incidence rates are expected to increase to 3.2 million and casualty rates to 1.6 million by 2040 [7] – an increase of 71.5% and 81.2% respectively. Stage of the disease at diagnosis is the most important prognostic factor for colorectal cancer [32, 43]. The primary method to detect pre-cancerous developments of colorectal cancer is through visual inspection of the colon. According to a 2012 survey, an estimated 15 million colonoscopies are performed annually in the United States [25]. Furthermore, the European Union and its member states have implemented colorectal cancer screening programs for early detection of cancer – typically consisting of a self-test and a follow-up colonoscopy if required [16]. Given the extensive number of polyps that are missed during colonoscopy [51, 61], additional support of clinical staff during this procedure is critical.

Colonoscopies are typically carried out at specialised endoscopy unit by a medical team consisting of one or two gastroenterologists or nurse endoscopists, as well as up to three nurse/healthcare assistants. The endoscopy itself is performed by the gastroenterologist or (nurse) endoscopist, controlling the endoscope and being primarily responsible for polyp detection and clinical decision making. This task is supported by the nurse/healthcare assistants, who are responsible for maintaining patient sedation, assisting in any required movement of the patient, and delivering medical instruments through a channel in the endoscope upon request. Although the operating

[1]Part of the World Health Organization's International Agency for Research on Cancer: https://gco.iarc.fr/.

gastroenterologist or endoscopist is primarily responsible for polyp detection they are often supported in this task by the other medical staff. The withdrawal phase of colonoscopy, in which the endoscope is manoeuvred from the cecum to the anus, is a critical component of the examination where major attention needs to be paid to the screen to identify often subtle visual cues that may indicate the presence of a polyp. Prior work shows that nurses participating as second observers during colonoscopy withdrawal can improve the adenoma detection rate [30]. Polyp detection is complicated by the fact that polyps can be difficult to see, can be covered in mucosa and therefore hidden from direct view, or can be missed when moving the camera through the colon.

In addition to the endoscope's camera, other frequently embedded tools are a light source and a water jet, respectively used for illumination and the removal of mucus or stool inside the colon. The aforementioned instrument channel can be used to navigate small medical instruments inside the colon. An example of a commonly used tool is the snare, a contractible wire loop used for polyp removal (*i.e.*, polypectomy).

2.1 Challenges in Colonoscopy

Whilst colonoscopy is effective at preventing colorectal cancer, the procedure is highly operator dependent. The key metric for determining the quality of colonoscopy is the adenoma detection rate (ADR). An adenoma is a subtype of polyp that is a common precursor to colorectal cancer. Unfortunately, ADRs vary considerably between different endoscopists even when accounting for patient related factors [11]. Polyps are missed broadly due to two reasons. Firstly, polyps may not be exposed by the operator due to poor withdrawal technique *e.g.* not manipulating the colonoscope effectively to look behind folds or inadequate washing of bowel contents [46]. Secondly, it is now appreciated that many polyps can be overlooked even when they are in the field of view particularly due to subtle appearances. Crucially, flat or depressed type polyps are most likely to be missed, which in turn also are more likely to have an increased risk of cancer progression [15]. Therefore, novel solutions to overcome both of these problems are essential.

The use of AI based software acting as a 'second observer' of the screen and highlighting suspected polyps in real-time, could potentially help overcome the issue of polyps that may otherwise be overlooked. Recent evidence suggests that AI assistance could lead to an increase in ADR [55]. Before widespread implementation and deployment of such AI systems can be considered, it is crucial that their design and integration into clinical workflow are explored further.

2.2 Artificial Intelligence in Healthcare

As AI support systems enter the medical domain, HCI plays a critical role in integrating these technological possibilities with clinical practice. Clinical decision-support systems (CDSS), perhaps the most widespread and enduring effort to bring AI into the hands of health professionals, have long suffered from a lack of successful integration into daily practice [37]. "*There is perhaps no omission that, historically, accounts more fully for the impracticality of many clinical decision tools than the failure of developers to deal adequately with the logistical, mechanical, and psychological aspects of system use*" [37]. Rather than deploying systems that are disconnected from existing workflows and practice, Yang et al. describe the concept of an unremarkable AI – suggesting that AI is most beneficial to clinicians when used to unobtrusively augment existing routines rather than developing stand-alone tools [59].

Already in 1990, Miller and Masarie critique the 'Greek Oracle' model present in diagnosis support systems [35]. These systems work independently from clinicians by first collecting all relevant patient information and subsequently producing an unintelligible diagnostic recommendation. Given the legislative implications faced in the healthcare domain, these type of recommendations are deemed unusable [60]. The explainability of AI systems has since been raised as a critical component for healthcare applications [21, 22].

Recent pioneering work by Cai et al. has explored the interaction between HCI and AI in medical decision making, specifically investigating the design of an interface that deals with imperfect algorithms and explainability in the context of pathology [9]. The system can be used to identify and present images of tissues with visual similarity to a tissue currently inspected by a pathologist, potentially assisting in their decision making. However, these images may not necessarily be medically relevant or meet diagnostic needs. In order to combat this problem, Cai et al. introduce a functionality called 'refine-by-concept', allowing pathologists to increase or decrease the representation of a selected clinical concept in the search results. The refinement tool was reported to increase both diagnostic utility and user trust in the support system [9]. Wang et al. design a clinical diagnostic tool for intensive care phenotyping which incorporates counterfactual rules to increase the explainability of AI conclusions [54]. By contrasting a patient's state with the parameters of a certain event (*e.g.*, blood pressure and shock), users can ask 'why not' questions and subsequently investigate the system's behaviour.

While the aforementioned work presents compelling examples of the positive role of HCI in integrating AI in the healthcare domain, they all focus on scenarios in which the user requires intermittent AI support. In this work, we explore how to design AI support for live video footage in a clinical context, requiring continuous adaptation of both user and AI system.

2.3 Video Annotation

Annotation of live video is an increasingly common practice in a number of different fields. For example, the use of augmented reality allows for the overlay of information on top of real world footage. These systems can provide support without forcing the user to change their visual focus, and are explored in the clinical and aerospace domain [14]. Other examples of video annotation can be found in sport broadcasting, where factors such as camera alignment and player pose estimation directly influence the presented information overlay [48]. Similarly, recent work in esports presents the use of visual overlays to *e.g.* compare an individual's performance against historical data or to explain strategies as they unfold [5].

Despite these existing applications, how to effectively communicate the results of computer vision systems to end-users is an open research question in HCI [28]. Work on intelligent image overlays in the medical domain has predominantly focused on still imaging, with radiology being a key area of development [23]. For example, Rajpurkar et al. show how AI can be used to both detect and localise pneumonia from x-ray images – producing a heatmap annotation to indicate to the clinician the area most indicative of pneumonia [44]. Their results achieve a better sensitivity (higher number of true positives) and specificity (lower number of false positives) as compared to human radiologists. Yet, as these algorithms are not infallible, the use of localised visualisations serves to inform the clinician and allow them to verify the AI's classification.

Although existing work on still image annotation is promising [23], the use of AI support for continuously adapting scenarios (*e.g.*, colonoscopy footage) remains under-explored. The live annotation of medical footage introduces several new challenges, such as the chance to miss AI annotations, to obscure the clinician's view of identified points of interest, and to process images in real time. This paper provides exploration of AI annotation on live footage by identifying the preferences and concerns of domain experts.

3 STUDY DESIGN

In order to systematically investigate user preferences of the visual marker designs (shown in Figure 1), we developed an online survey to be distributed among the colonoscopy community. Rather than presenting still images, we incorporated our designs into endoscopic footage as captured during patient inspection. Although this enhances the ecological validity of the study as compared to still images, we stress that potential real-world effects such as end-user fatigue, risk of deskilling in polyp identification, and false positives are outside the scope of the study. Furthermore, as we expect that AI support will be deemed as more useful when presented with

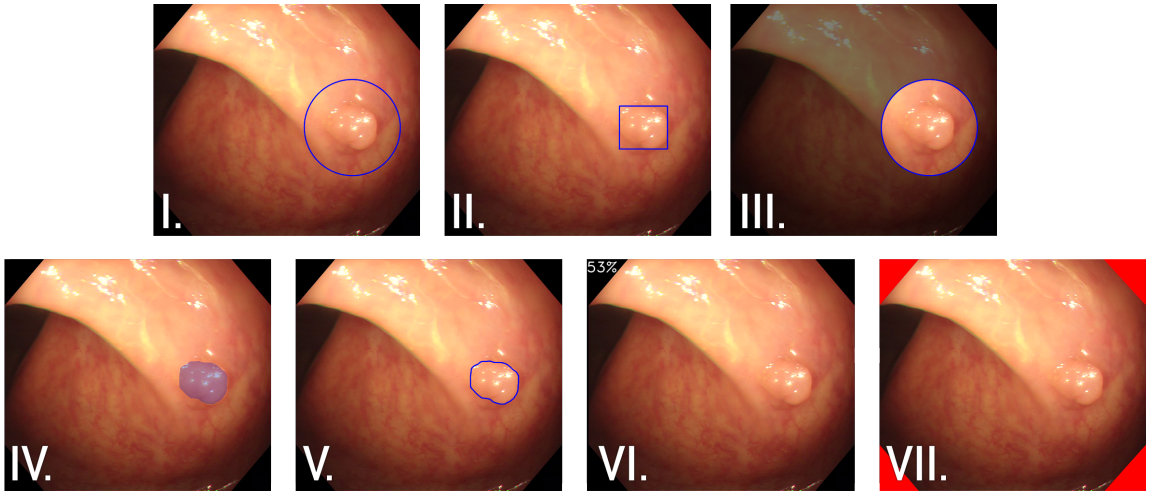

Fig. 1. Still frame of visual markers videos (showing apparent polyp). Visual markers are described in detail in Section 3.1.

challenging scenarios, we made use of two unique videos, one video containing an apparent polyp and one video containing a challenging polyp – a still image from each video is shown in Figure 2. We obtained ethical approval for the use of these (anonymised) videos for research purposes prior to colonoscopy. Following manual annotation of the polyps[2], the visual markers are subsequently rendered on top of the original videos. As such, we end up with a total of 14 videos, with each design presented twice to each participant (once for an apparent polyp and once for a challenging polyp). We randomise the order of these videos (both the visual marker and the video difficulty) for all participants. Participants are shown both the apparent and challenging video without any overlay prior to the presentation of the visual markers, allowing participants to (attempt to) identify the polyp on their own.

3.1 Materials

Through extensive ideation and iteration with two gastroenterologist team members, combined with a total of five observation sessions during colonoscopy, we identified an initial set of designs, which we then turned into visual mockups. We employed a structured approach to the design of the visual markers, as described

[2]Frame-by-frame annotation by a domain expert using *PixelAnnotationTool* [8].

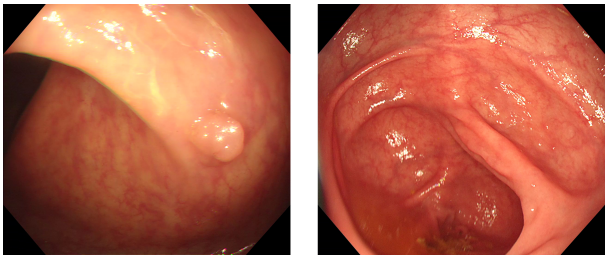

Fig. 2. Still frames of the apparent (left) and challenging (right) polyp. Both videos are 9 seconds in length.

	I. Wide bounding circle	II. Tight bounding box	III. Spotlight	IV. Segmentation	V. Segmentation outline	VI. Detection confidence	VII. Detection signal	Rationale
Localise attention	✓	✓	✓	✓	✓			"*Display information relevant to the user's current task and environment*" [2].
Contrast colours	✓	✓	✓	✓			✓	Guiding attribute in visual attention [58], colour visible to colour blind operators.
Highlight target				✓	✓			Highlighting the target area with contrasting colour guides visual attention [58].
Blurred guidance			✓					Blurring of the non-target area guides visual attention to target area [20].
Indicate confidence						✓		Communicate to the user the AI's current level of certainty [2].
Size	✓	✓	✓					Minimum size to ensure visibility [58], otherwise dependent on identified polyp.
Unobtrusive	✓	✓	✓		✓	✓	✓	Displayed only when necessary (*i.e.*, on detection) and avoids overlapping of target area [2].
Motion								Although a guiding attribute [58], motion would further impede the already unstable camera feed.
Orientation								Although a guiding attribute [58], orientation is irrelevant for our circular and shape-dependent designs.

Table 1. Design space for real-time visualisations in AI-supported colonoscopy. Recommendations based on guidelines from visual attention [20, 57, 58] and Human-AI interaction [2].

in [40]. In particular, we identified a suitable design space through analysis of relevant guidelines on visual attention [20, 57, 58] and AI interaction [2] – as well as an identification of visual designs used in existing systems [36].

Following this, another round of selections was made in consultation with the aforementioned gastroenterologists – leading to the exclusion of visual designs which were either considered to have a clear undesirable effect on the endoscopic procedure or contained too much overlap with the other designs. Rather than fully extending all design options (*e.g.*, include combinations of individual designs, bounding boxes using different geometric shapes), we selected designs deemed most viable and differentiating as based on the identified design-space and input from domain experts. This reduced the number of options presented in the survey and helped to manage participant strain. The selected seven designs were then converted from a stationary visual mockup into working video-based visual prototypes. For this, we annotated the individual frames of the two aforementioned colonoscopy examination videos and programmed image overlays using *OpenCV* [6]. Although the highlights are based on a frame-by-frame annotation, we apply a straightforward smoothing algorithm based on preceding frames to increase the fluency of the videos. Both the generated videos and the code required to generate each individual marker are included in the paper's auxiliary materials.

Following a third round of discourse, we made a number of minor modifications (*e.g.*, edge thickness). An overview of our final visual markers is shown in Figure 1 and discussed below. We include the respective rationale and literature recommendations that drove our design decisions in Table 1.

I. Wide bounding circle. A wide circle positioned around the polyp's centre point.
II. Tight bounding box. A rectangle enclosing the polyp's outline.
III. Spotlight. A wide circle positioned around the polyp's centre point, dims background display.
IV. Segmentation. A transparent overlay of the polyp's shape, as employed by Ganz et al. [18].
V. Segmentation outline. An outline of the polyp's shape, as employed by Bernal et al. [3].
VI. Detection confidence. A percentage sign (upper-left corner) indicating likelihood of polyp in the current frame. Width of annotated polyp used to mimic the confidence of a real-life system. Confidence level of 0% when no polyp is visible, between 50–75% when the polyp is most clearly visible. Based on a current deployment [36].
VII. Detection signal. Corners of the viewport highlight in clear red when a polyp is detected. Based on a current deployment [36].

For the visual markers A–E, we decided on the marker's colour (blue) by evaluating the visibility of the marker when positioned on a snapshot of a colon. We evaluate the contrast of marker and colon for trichromacy (*i.e.*, healthy colour vision), deuteranomaly (*i.e.*, weak green vision), and protanomaly (*i.e.*, weak red vision). Deuteranomaly and protanomaly account for the majority of colour blindness (known as red-green colour blindness), affecting 8% percent of men and 0.5% percent of women of Northern European ancestry [13].

3.1.1 Survey questions. Participants were shown a total of 14 videos (seven designs across two different videos) – with each video totalling nine seconds in length. For each video, participants were asked to answer four questions concerning the presented design ((1) allow to detect more polyps, (2) locate polyps faster, (3) interference during polyp removal, and (4) interference with the regular display) as compared to baseline (no visual marker). Participant responses were restricted to a Likert-item format. We incorporate design recommendations from the literature such as the use of a 7-point scale [17], including a mid-point option ('neither agree nor disagree') [10, 12], and labelling all answer options as opposed to only the end-points [10]. Furthermore, participants were given the ability to provide additional commentary on the marker via an open textfield. After the participants had answered the aforementioned questions for all 14 videos, they were asked to rank both the full set of design options (seven images) as well as a range of colour options (seven images) in order of preference.

Following this, participants answered a range of questions concerning the design and implementation of an AI support system. This included questions on alert modalities (*e.g.*, sound) and the general integration of such a system in the endoscopy room. We included these questions to obtain additional insight into the workplace integration of AI support systems. The survey also collected the participant's professional role and demographic information. We pilot tested our survey with three local endoscopists independent from the project. We include the survey in full in Appendix A.

3.2 Participants

Following ethical approval, we disseminated the survey using national (anonymised for review) mailing lists targeting the colonoscopy community. Furthermore, we visited a local hospital to recruit participants not easily reached through professional mailing lists (*i.e.*, assistants). As colonoscopy is performed by a clinical team, we aimed to collect responses from the broadest variety of roles present in the endoscopy room – all of whom are inspecting the video source of the colonoscope, including both assistants (staff nurse, healthcare assistant) and endoscopists (gastroenterologists, colorectal surgeons, and nurse endoscopists). In order to incentivise participation among our difficult to reach target group [52], we included a £100 voucher as a raffle.

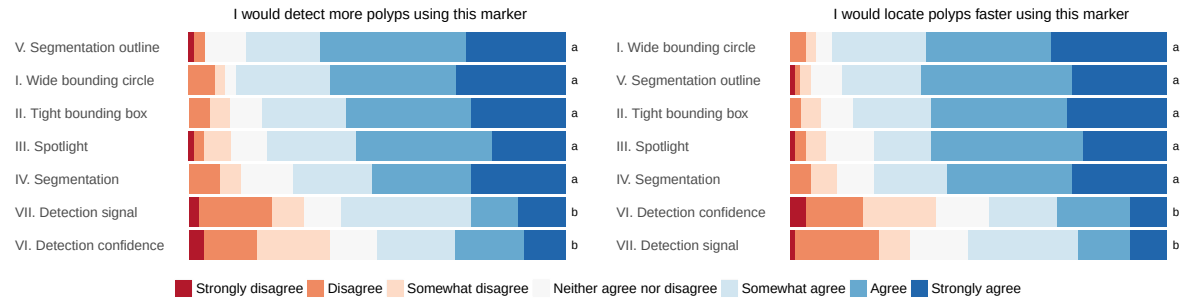

Fig. 3. Responses to 7-point Likert items on polyp detection and localisation using the visual markers (combined apparent and challenging videos).

3.3 Analysis

The analysis of Likert items is a widely debated topic, both within the field of HCI [10, 26] and the larger academic community [29]. A major point of discussion is the interpretation of responses as either ordinal or interval data, which subsequently informs the selection of the applied statistical test. The use of interval data assumes that Likert item-answers are equidistant in nature (*i.e.*, the difference between 'Strongly agree' and 'Agree' is equal to that of 'Neither agree nor disagree' and 'Somewhat agree'), a contested perspective [29]. In this work, we follow an ordinal interpretation for the analysis of our dependent variables (*i.e.*, perceived ability to detect more polyps, locate polyps faster, and interference of the visual markers).

Although a Friedman test is suitable for the analysis of non-parametric ordinal data with repeated measures (*i.e.*, different conditions, in our case visual markers), it does not allow for the analysis of multiple factors or interactions (*e.g.*, difficulty of the video, role of the respondent). We therefore analysed both our Likert-response and rank-order data using Aligned Rank Transform (ART), as described by Wobbrock et al. [56], who explain that "*The ART relies on a preprocessing step that 'aligns' data before applying averaged ranks, after which point common ANOVA procedures can be used.*" [56]. We utilised the *R* package *ARTool* [27].

In case we reject the null hypothesis, we complete a post-hoc test to identify differences between groups. We apply a Bonferroni correction to all pairwise comparisons to protect against the increased risk of Type I errors. In the interests of space, we summarise all pair-wise comparisons on the right-hand margin of the figures (see *e.g.* Figure 3). This so called 'compact letter display' assigns unique letters for conditions which are significantly different from one another [42]. Conditions that are not significantly different are followed by a common letter. A condition assigned the letter 'a' is significantly different from a condition marked with a different letter. A condition marked with two or more letters (*e.g.*, 'ab') is not significantly different from conditions labelled with one or more of the same letters (*e.g.*, 'a' or 'bc'). Our analysis files (survey responses and *R* scripts) are included as supplementary material.

4 RESULTS

We collected a total of 36 completed surveys. Our respondents were 8 assistants and 28 endoscopists. From the 36 completed surveys, 31 were collected as a response to our online dissemination, with the remaining five surveys having been collected from a local hospital (all were assistants). Median survey completion time was 19.02 minutes. Our sample consists of 23 men and 13 women, with an average age of 38.28 (SD = 9.66). The endoscopists had a varying degree of experience, with seven respondents performing colonoscopy for over 15 years, three respondents between 5–15 years, twelve respondents 2–5 years, and six respondents less than 2 years.

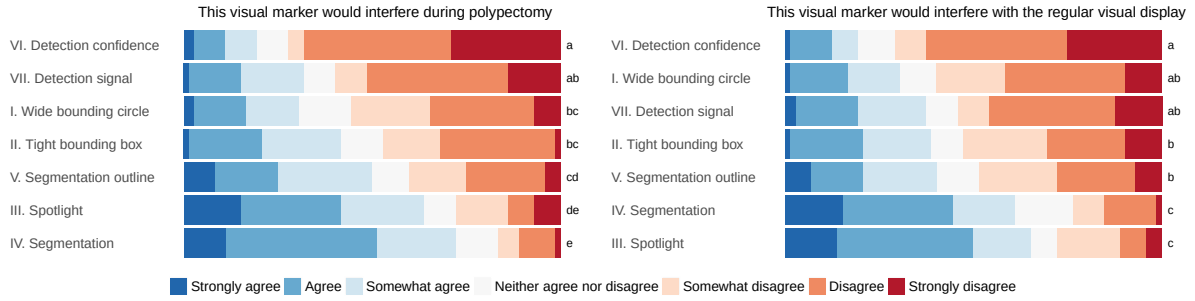

Fig. 4. Responses to 7-point Likert items on perceived interference of the visual markers (combined apparent and challenging videos).

Inspired by Sarwar et al.'s analysis of AI usage in pathology [47], we collected the participant's excitement, concern, and self-reported likelihood of uptake of AI technology for colonoscopy. Participants were mostly excited about the use of AI for colonoscopy (14 'moderately', 8 'very', and 9 'extremely' excited), with two participants reporting to be 'slightly' excited and three participants reporting to be 'not excited at all'. In terms of concerns surrounding AI for colonoscopy, we find eight participants who are 'not concerned at all', 13 who are 'moderately concerned', another 13 'slightly concerned', and two participants who are 'very concerned'. Finally, participants report high willingness to use (validated) AI technology in their job, with 27 respondents indicating either 'likely' or 'very likely', six 'neutral', and three either 'unlikely' or 'very unlikely'.

4.1 Detection & Localisation

We distinguish between detecting and locating a polyp within the colon. Detection concerns identification of a polyp 'somewhere' within the frame, whereas localisation additionally considers the exact location of the polyp. We present an overview of participants' responses to the two Likert questions concerning detection and localisation in Figure 3. We note mostly positive responses among participants for both these questions across all visual markers.

Using the aforementioned Aligned Rank Test (ART) for data alignment [56], an ANOVA reveals a significant main effect of *Visual Marker* on *Perceived Increase in Polyp Detection* ($F_{6,490} = 9.40$, $p < 0.001$). A post-hoc test using pairwise comparison with Bonferroni correction showed significant differences between the rating of the visual markers, as indicated using a compact letter display in Figure 3. We find no significant effect of *Video difficulty* on *Perceived Increase in Polyp Detection* ($F_{1,490} = 1.80$, $p = 0.180$). We discuss the effect of video difficulty in more detail in Section 4.2.1. Summarising these results, we find a significant difference between two clusters of visual markers. Markers visualised in the corner (*i.e.*, 'VI. Detection confidence', 'VII. Detection signal') are rated as significantly less useful in detecting more polyps as compared to the remaining set of markers. Participants experienced the 'Detection confidence' design as distracting; "*This [design] takes the attention away from the mucosa for a number that is almost continuously changing. [...] This display is very inefficient on its own and with continuous variation.*" (P03).

Similarly, we evaluate the effect of visual marker on the perceived ability to locate polyps faster. We find a significant effect of *Visual Marker* on *Perceived Faster Polyp Detection* ($F_{6,490} = 12.58$, $p < 0.001$). A post-hoc test using pairwise comparison with Bonferroni correction showed significant differences between the rating of the visual markers, indicated using a compact letter display in Figure 3. We find no significant effect between *Video difficulty* and *Perceived Faster Polyp Detection* ($F_{1,490} = 1.75$, $p = 0.187$). Similar to the previous results, we find a significant difference between the 'VI. Detection confidence' and 'VII. Detection signal' markers and the

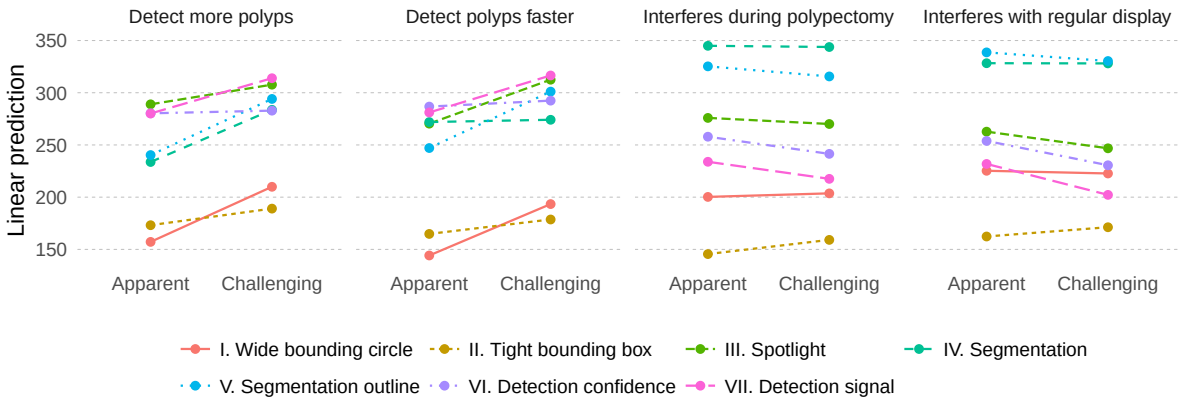

Fig. 5. Effect of video condition (Apparent or Challenging to detect polyp) on participant ratings across four Likert items.

remaining markers, with participants assessing the former least positively. Participants noted how the use of *e.g.* 'VI. Detection confidence' would not help in actually finding the target of interest, while pointing out that AI systems may not be entirely accurate – thereby wasting valuable time; "*I don't see the benefit in this - doesn't help you localise the lesion and would be desperately irritating when managing a false positive*" (P23). The speed at which a design helps in locating polyps was raised as an important benefit by some participants, *e.g.* regarding the 'III. Spotlight' design; "*This was much clearer - helped locate the polyp much faster*" (P07).

4.2 Interference

As AI support systems will not be 100% foolproof, it is critical that these systems do not interfere with the clinical staff's ability to perform their task. In addition to a visual marker's ability to support in detecting and locating polyps, we therefore also explore how obstructing the proposed designs are. Here, we distinguish between interference of the design during a specific task (removal of a polyp) and general interference with the visual display.

We find a significant main effect of *Visual Marker* on *Perceived Interference during Polypectomy* ($F_{6,490} = 18.07$, $p < 0.001$). A post-hoc test using pairwise comparison with Bonferroni correction showed significant differences between the visual markers, as indicated in Figure 4. We find an indicative significant effect of *Video difficulty* on *Perceived Interference during Polypectomy* ($F_{1,490} = 3.69$, $p = 0.055$). Finally, we evaluate the perceived effect of the visual marker on interference with the regular visual display. We find a significant effect between *Visual Marker* and *Perceived Interference during Regular Display* ($F_{6,490} = 14.29$, $p < 0.001$), with significant post-hoc results (Bonferroni correction applied) indicated in Figure 4. We find an indicative significant interaction effect between *Video difficulty* and *Perceived Interference during Regular Display* ($F_{1,490} = 3.34$, $p = 0.068$).

Both with regards to interference during polypectomy and interference with the regular visual display, participants considered the two markers positioned in the corner of the video (*i.e.*, 'VI. Detection confidence' and 'VII. Detection signal') as significantly less interfering than the other designs, followed by the 'I. Wide bounding circle' and the 'II. Tight bounding box'. The 'III. Spotlight', and 'IV. Segmentation' markers were rated as the most interfering designs. The 'III. Spotlight' design in particular was critiqued on the fact that it hides part of the display from the operator; "*It looks nice but VERY distracting and would mean I couldn't spot other polyps missed by the AI (which I'm sure is possible!)*" (P23). Furthermore, participants expressed on their own accord how some of these designs would require the ability to turn the device off, an opinion expressed multiple times for

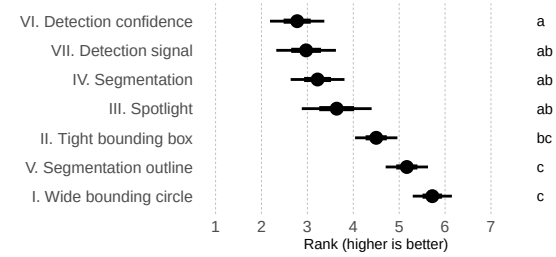

Fig. 6. Rank order of visual markers. Graph indicates the mean rank values, as well as one and two SD.

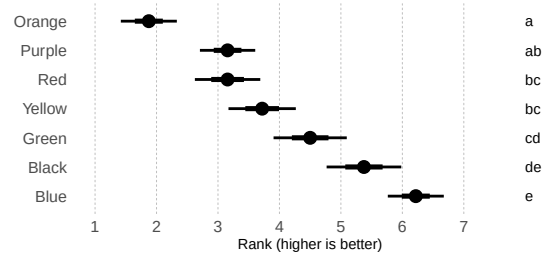

Fig. 7. Rank order of colour preference. Graph indicates the mean rank values, as well as one and two SD.

the 'Segmentation' design; "*After identification there should be an option to disable the marker so the operator can proceed with polypectomy*" (P03).

4.2.1 Effect of video difficulty. Across two of the four analysed question items, we find an indicative significant effect of video difficulty on participant answers (with the two remaining questions showing a similar trend). Investigating these results in more detail, we plot the effect of video type for each question in Figure 5. Although our tests did not reveal a significant effect of video difficulty, we observe a clear direction of effect for both the increased detection of polyps and their faster localisation. In videos in which the polyp is more challenging to spot, our visual markers are rated as more positively as compared to their respective ratings on the video with a more apparent polyp.

4.3 Marker design

In order to assess the participant's overall preference of the different designs, our survey included a score sorting question. Participants ranked each visual marker by assigning a unique number from 1–7. Figure 6 visualises the distribution of participant rankings, with distinct differences visible between conditions. An ANOVA on our aligned rank transformed data reveals a significant main effect of *Visual Marker* on participant *Ranking* ($F_{6,238} = 9.88$, $p < 0.001$). No significant interaction was found between *Professional Role* and *Ranking* ($F_{1,238} = 0.06$, $p = 0.809$). A post-hoc test using pairwise comparison with Bonferroni correction showed significant differences between the rating of the visual markers. From these results, we observe that 'I. Wide bounding circle' and 'V. Segmentation outline' are the most popular designs. Reflecting the most common critiques expressed by our participants, P23 states; "*[The] box and circle are very clear - a geographical outline changes too often and is too distracting. The dimming thing is a pain and the corner flashes and the numbers are too distracting*".

4.3.1 Colour. Similar to the ordering of preferred visual markers, participants indicated their preferences of the colour of the marker. Figure 7 visualises the distribution of participant colour preferences. An ANOVA on aligned rank transformed data reveals a significant effect of *Colour* on participant *Ranking* ($F_{6,238} = 9.88$, $p < 0.001$). No significant effect was found for *Professional Role* on *Colour* ($F_{1,238} = 0.06$, $p = 0.809$). We label the significant differences between groups in Figure 7. The colours 'Blue' and 'Black' rank significantly higher than the other colours. Although two of our participants do not express a strong opinion on the colour of the visual marker, "*No real preference here*" (P30), the majority of the participants comment on the need for contrast and ability to easily distinguish the visual marker from other entities; "*I would favour maximizing contrast between marker and background*" (P03) and "*Catches the eye and will usually be alien to see that colour in the colon*" (P13).

4.4 Workplace Integration

We examined participants' beliefs on how to integrate AI support systems in current practice, focusing on practical aspects such as the ability to disable the support system and the use of additional modalities in communicating with the AI.

Participants almost unanimously agreed that it should be possible to disable (and subsequently enable) the visual markers, with 33 out of 36 participants responding in favour. Participants listed a number of reasons for wanting to disable the visual markers, such as being able to focus on a polyp during polypectomy, bowel content requiring cleanup, or in case of poor classification performance by the AI system. A minority of participants are in favour of an automated disable functionality, for example after signalling the likely presence of a polyp "*I think the marker should only appear when polyp is identified once the polyp is detected and highlighted then the marker should automatically be turned off*" (P26), however the majority of participants expressed desire for a manual way to turn off the visual markers. Reasons for turning the visual markers off include training sessions, occlusion during polyp interrogation and polypectomy, and annoyance with false positives in a specific area of the colon. Although there appears to be wide support for being able to disable the support system, a small number of participants warned that turning the system off may have unintended consequences; "*If its turned off then [the] endoscopist may forget to turn it back on.*" (P26), and "*I don't want to be able to turn it off by mistake without realising it, but it does need to be quick and easy and reliable to do – I would suggest a reminder to turn it on again after a set time (e.g., 1 minute) if it is silenced/switched off*" (P22). When asked how to disable and enable the system, the most popular answer option – with 27 respondents in favour – was the use of a physical button on the colonoscope.

When considering the integration of additional modalities in communicating the results of the AI support tool, our results indicate differences in opinion between participants. We allowed participants to select (multiple choice) from a list of pre-defined options (sound, vibration, no additional modalities, unsure) as well as free text input. A total of 14 participants indicated that no additional modalities should be included, and seven participants expressed that they did not know whether any additional modalities should be included. Remaining participants primarily supported the inclusion of an audible alert (16 participants). Three participants believed a vibration alert (*e.g.,* wristband) would be useful. One participant commented that the image could 'freeze' for a few seconds after detection.

5 DISCUSSION

The growing use of AI across a range of application domains has led to an increased focus on usable [9, 39], fair [49, 53], and transparent [4] systems within the HCI community as well as the wider public debate [33, 50]. The successful uptake and deployment of these algorithmic systems does, however, rely on the incorporation of stakeholders' knowledge and feedback [62]. As expressed by Zhu et al., "*algorithm developers [need] to think not just about 'solving problems' but also to respect the needs, motivations, and values of stakeholders*" [62].

In this work, we have used colonoscopy as a case study for the design of visual markers for AI detection systems in continuously adapting tasks. Colonoscopy is a highly-skilled task, which – despite careful training and increased quality standards [45] – features extensive miss rates in polyp detection [51, 61]. Furthermore, colonoscopy is a multifaceted task, involving a team of specialists to *inter alia* detect, locate, and remove polyps while navigating a complex organ. As such, we provide a real-world case study for the integration of AI-technology. Our results highlight clear differences in end-user preference across the seven presented visual markers, with a wide bounding circle design (Figure 1-I) being the most popular among our respondents. In addition, blue was the preferred colour option for the visual marker due to its contrasting nature when overlaid on colon footage.

Participants reported positively on their expected adoption of AI support tools in colonoscopy. These results are in line with participants' generally positive take on the use of AI support systems in detecting and locating

polyps (see Figure 3). Although these results appear promising for the future development of clinical AI support systems, we note that a survey on AI is likely to attract those who are already invested in the topic (participation bias).

5.1 Designing AI for Continuous Adaptation

Previous work on the design of AI systems has typically modelled the interaction between user and AI as a turn-taking process (see *e.g.*, [2, 9, 49]). Following this paradigm, users are presented with a single AI recommendation following an explicit cue (*e.g.*, an on-screen button, "*Hey Siri*"). This process is repeated as required, with the user actively in control of changes in the commands given to the AI (*e.g.*, parameter adjustments made on screen). In the use case explored in this work, however, interaction between user and AI support follows a different paradigm. Rather than actively requesting the system's support, the user provides continuous input and, as such, relies on a more ubiquitous AI system. This results in a continuous interaction between (sustained) user input (*e.g.*, moving a camera feed) and AI suggestions. Other recent examples from the HCI literature include the use of gaze-enabled intention recognition [38] and collaborative music improvisation [34] – showcasing how both implicit and explicit user input can drive continuous AI support systems. We illustrate these different interaction paradigms in Figure 8.

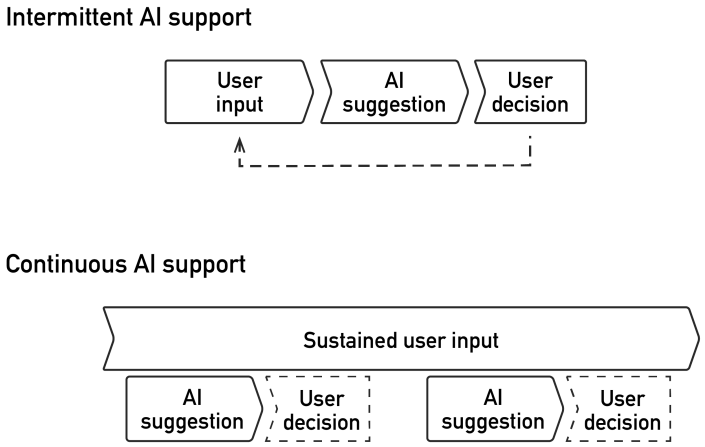

Fig. 8. Two distinct models of AI support systems. Top: intermittent AI support in which AI suggestions are pro-actively requested. Bottom: continuous AI support in which the user's actions and AI support are intertwined.

Continuous adaptation introduces new challenges in the design of AI systems. The HCI and digital health community therefore needs 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. Here, we discuss some of the challenges in relation to current AI-interaction guidelines when considering continuous AI support applications.

Our results clearly indicate that users want the ability to turn off the AI system as required. This request to turn off the AI support follows the 'Support efficient dismissal' guideline proposed by Amershi et al. [2]. In our use case, reasons for turning off the AI support may *e.g.* be when removing a polyp or when the occurrence of stool in the colon inhibits the AI from accurately detecting elements of interest. However, such use cases have universal counterparts; respectively the requirement to switch to a non-AI-supported subtask, and the presentation of data which the AI system is unable to handle. For continuously adapting scenarios, it is critical

Design Guidelines for Continuous AI	Rationale	Contrast with Intermittent AI
1. Position support information in close visual proximity to the area of interest to the user.	Avoid forcing users to switch their focus during sustained user input, subsequently missing critical information.	In intermittent AI, the recommendation from the system is typically the primary content – not the case for continuous AI.
2. Do not overlay the user's area of interest with additional information.	Avoid obfuscating the view and prevent further analysis by the user (*e.g.*, in case of false positives).	The recommendation "*Make clear what the system can do.*" [2] can result in an overabundance of information.
3. Minimise the amount of visual information presented outside of the identified 'area of interest'.	Allow the user to pick up on information failed to be identified by the system (*e.g.*, false negatives). Prevents distraction.	See above.
4. Allow the user to manually disable and enable the AI support system.	The user may wish to switch to a directly related subtask which is unsupported and potentially frustrated by the AI system.	Aligns with previous guidelines [2].
5. Display the current operational state (*i.e.*, 'on/off') of the AI support system.	Ensure that users do not involuntarily leave the AI system turned off after disabling AI support.	Displaying the operational state is noncritical for intermittent AI as the turn-taking interaction immediately reveals the state.
6. Allow retrospective access to the explanations of the AI's recommendations.	Avoid user interruption with noncritical information and allow user to obtain explanation following task completion.	"*Make clear why the system did what it did.*" [2] is typically not appropriate in-the-moment. Continuous scenario requires user to focus elsewhere.
7. Use contrasting colour in the presentation of support elements.	Ensure AI recommendation stands out from user content.	N/A

Table 2. Overview of design guidelines for continuous AI. Presented as an adaptation to the 'Guidelines for Human-AI Interaction' by Amershi et al. [2] which focused primarily on intermittent AI systems.

for the system to inform the user that AI support has been turned off, for example by displaying an icon on the screen. Involuntarily completing *e.g.* the remainder of a colonoscopy without AI support could negatively affect the user's performance.

Another recommendation from the literature on AI interaction describes the need for explainability; 'Make clear why the system did what it did' [2]. While explainability (and subsequent causability) is key to increasing the user's trust and understanding of an AI system [22, 54], the nature of continuous tasks does not allow for immediate extensive interactions between user and system. Both the available visual space (*e.g.*, computer screen) and cognitive space (*i.e.*, user's focus on the task at hand) do not allow for interruption with additional explanations. Integrating explainability in these applications therefore has to follow a different model. For example, in the case of colonoscopy, the clinician can be presented with an explanation of individual AI suggestions in cases of uncertainty while completing the patient's operative report.

Finally, the guidelines by Amershi et al. suggest that the interface should 'Make clear how well the system can do what it can do' and 'Show contextually relevant information' [2]. These features most closely resembled our 'Detection confidence' design, in which a percentage sign – positioned in the top-left corner of the video – indicated the certainty with which the AI had detected a polyp. This design, which was voted as least popular, was considered as distracting and forcing the operator to focus on two locations simultaneously.

Although our results highlight some of the difficulties in applying existing guidelines in novel scenarios, this attests to the challenges of designing real-life applications. We call on the HCI community to continue to develop and revisit our shared guidelines as we learn from evaluations of 'real world' AI applications. As stressed by Amershi et al., "*Further research is necessary to understand the implications of these potential interactions and trade-offs for the design of AI systems and to understand how designers employ these guidelines 'in the wild'.*" [2]. We summarise our contributions to human-AI interaction design guidelines in Table 2.

5.2 Integrating AI in Medical Practice

Recent work has highlighted the disconnect between technological advances in AI and the integration of these breakthroughs in clinical practice [59]. The use of AI technology in a clinical setting raises numerous questions regarding *e.g.* data ownership, anonymity, and representation. These issues are not unique to medical practice, but such issues come to the fore in the medical context, where patient safety, privacy, etc. are paramount, and where team working and the uniqueness of each patient increase the inherent complexity of the work. In an assessment of clinical decision-making processes, Yang et al. attribute the repeated failure of AI systems in healthcare to a "*lack of contextual integration in the design of these systems.*" [59]. Rather than developing novel stand-alone tools, Yang et al. suggest that AI should unobtrusively augment existing routines. We have presented an evaluation of different visual markers as presented on real-world patient footage, increasing ecological validity without compromising current clinical practice and patient safety. Our methodological approach highlights solutions deemed clinically acceptable by our target audience; this approach can be readily applied to other research questions. For example, while we expect similar results in related applications of endoscopy (*e.g.*, upper gastrointestinal tract), we anticipate that AI-supported imaging during surgery (*e.g.*, in laparoscopy or thoracoscopy) or endoscopy under distinct circumstances (*e.g.*, small size during fetoscopy) may result in significantly different end-user requirements. Our findings show that the design of recently developed and deployed colonoscopy support systems (*e.g.*, [3]) fails to align with end-user requirements – potentially distracting rather than supporting in the goal of polyp detection. These results can be used to inform future clinical trials of AI support systems.

Although not the primary focus of this study, we note that the overall positive attitude of our participants towards AI support was complemented by a number of concerns. First, clinicians were aware of the limitations of AI, and used appropriate jargon to describe *inter alia* the occurrence of false positives and false negatives. An analysis of the effect of false positives and false negatives on endoscopists was out of scope for this study, though our results highlight the importance of studying this topic in more detail from an HCI and system-integration perspective. Second, as the responsibility of polyp detection lies with the clinicians rather than the support system, participants were resolute to dismiss visual markers which may interfere with their own ability to identify polyps. Simultaneously, visual markers which did not provide a sufficient level of support were not positively received (see Figure 3). Third, medical practitioners are inevitably going to display signs of fatigue during lengthy procedures. This may affect the level of support that is required by the user and therefore require a change in the way that the AI system is presented. Our study design could not capture these changes in end-user fatigue and we therefore cannot assess how continued usage of the system under real-world circumstances would alter our study outcomes. Fourth, whether or how AI systems affect a clinician's level of skill and concentration over time is an unexplored question – raising both opportunities for education and challenges around patient safety. These four aforementioned concerns highlight opportunities for future HCI research in this domain.

5.3 Limitations and Future work

The current study presents a first assessment of the design of visual markers through an online survey. Our study did not capture the actual behaviour of participants in relation to the different designs. Running the study in a laboratory environment on *e.g.* a dummy colon was deemed unfeasible for a multitude of reasons. First, the duration and complexity of such an experiment would have severely limited our population sample given busy clinical work schedules. Capturing operator behaviour (*e.g.*, detection of polyps) would require the evaluation of an entire colon per design and compensate for differences between operator detection rate and the variability of polyp discoverability. Second, we were interested in collecting insights from a variety of end-user roles (*e.g.*, assistants) as opposed to solely focusing on gastroenterologists. Third, the use of a survey allowed us to collect responses from a national population of experts rather than limiting ourselves to a local colonoscopy centre. The subjects' specialist knowledge and the required length of the survey were expected limiting factors in the study's sample size [52]. However, we argue that by tapping into the domain specific knowledge of experts we were able to obtain valuable insights. These insights were often not in line with what we had expected, highlighting the importance of inviting subject specialists. By showing our designs on video footage of patients as opposed to the use of still images we more closely resemble the real-life interactions of clinical staff, thereby increasing the study's ecological validity, while simultaneously providing a practical approach to reach a difficult target audience while taking into account patient safety.

A second limitation of our study is the number of visual markers we were able to present to participants. Although we followed a systematic way to generate the visual markers, we did not include complementary visual modalities (*e.g.*, combining 'II. Tight bounding box' and 'VI. Detection confidence') in order to limit the time required to complete the survey. Interestingly enough, none of our participants suggested the combination of visual markers.

AI-based detection systems will inevitably register false positives. Future work should investigate the effect of these false positives on both subsequent (clinical) error (*i.e.*, clinician following an erroneous system suggestion) and a potential decrease in system uptake. This draws a parallel with the more widely investigated concept of alarm fatigue in hospitals [19], in which system warnings are ignored due to their prevalence. Similarly, the occurrence of false negatives (*i.e.*, failure to detect) remains an area of concern which can occur due to a variety of factors, *e.g.* poor bowel preparation, incomplete coverage of the bowel, or an over reliance on the AI support system – especially in light of current omission rates (22–27% [51, 61]).

Our study is focused on the use of AI support for colonoscopy. Given the essential role of live imaging in today's clinical practice, AI support systems will be embedded in a range of video examination instruments. While each area faces its unique challenges, how to best present AI results within existing medical imaging applications is a shared research question. To support the future systematic study of this question in other (medical) imaging applications we publicly release the source code required to generate the visual markers on any annotated video material[3].

6 CONCLUSION

In this paper we reported on an online survey study examining the design and implementation of a support system for colonoscopy. Thirty-six domain experts were shown seven unique designs for visual markers overlaid on real-world patient footage and asked for their opinion on the integration of AI technology in their daily work. Our findings reflect the trade-offs present in the selected range of visual markers, and provides concrete recommendations on the implementation of AI for continuously adapting scenarios. Concretely, our results identify a wide bounding circle with a high-contrast colour (*i.e.*, blue) as the most preferred visual marker design. While our work focuses on colonoscopy, the implications of our results extend beyond this application area.

[3]Available in the paper's supplementary material and at https://github.com/nielsvanberkel/Visual-Markers.

Future AI applications will be deployed in continuously adapting scenarios (*e.g.*, endoscopic imaging, autonomous driving, robot-assisted surgery), where the interaction between user and system is not intermittent. This model of continuous AI support presents novel challenges around localisation of user attention, the tradeoff between missing critical content or AI recommendations, and the manual interruption and subsequent continuation of automated support. Balancing these user needs and the novel opportunities provided by AI in daily practice is a key avenue for HCI and AI researchers.

ACKNOWLEDGMENTS

This work was supported through the Wellcome/EPSRC Centre for Interventional and Surgical Sciences (WEISS) [NS/A000050/1] at UCL. We are grateful to Daniel Toth for his technical support.

REFERENCES

[1] Omer F. Ahmad, Antonio S. Soares, Evangelos Mazomenos, Patrick Brandao, Roser Vega, Edward Seward, Danail Stoyanov, Manish Chand, and Laurence B. Lovat. 2019. Artificial intelligence and computer-aided diagnosis in colonoscopy: current evidence and future directions. *The Lancet Gastroenterology & Hepatology* 4, 1 (2019), 71 – 80. https://doi.org/10.1016/S2468-1253(18)30282-6

[2] 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 *(CHI '19)*. ACM, New York, NY, USA, Article 3, 13 pages. https://doi.org/10.1145/3290605.3300233

[3] J. Bernal, J. Sánchez, and F. Vilariño. 2012. Towards automatic polyp detection with a polyp appearance model. *Pattern Recognition* 45, 9 (2012), 3166 – 3182. https://doi.org/10.1016/j.patcog.2012.03.002

[4] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions *(CHI '18)*. ACM, New York, NY, USA, Article 377, 14 pages. https://doi.org/10.1145/3173574.3173951

[5] Florian Block, Victoria Hodge, Stephen Hobson, Nick Sephton, Sam Devlin, Marian F. Ursu, Anders Drachen, and Peter I. Cowling. 2018. Narrative Bytes: Data-Driven Content Production in Esports *(TVX '18)*. ACM, New York, NY, USA, 29–41. https://doi.org/10.1145/3210825.3210833

[6] G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools* (2000).

[7] F. Bray, J. Ferlay, I. Soerjomataram, R. L. Siegel, L. A. Torre, and A. Jemal. 2018. Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA: A Cancer Journal for Clinicians* 68, 6 (11 2018), 394–424.

[8] Amaury Bréhéret. 2017. Pixel Annotation Tool. https://github.com/abreheret/PixelAnnotationTool.

[9] 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 *(CHI '19)*. ACM, New York, NY, USA, Article 4, 14 pages. https://doi.org/10.1145/3290605.3300234

[10] Paul Cairns. 2019. *Doing Better Statistics in Human-Computer Interaction.* Cambridge University Press. 253 pages. https://doi.org/10.1017/9781108685139

[11] S. C. Chen and D. K. Rex. 2007. Endoscopist can be more powerful than age and male gender in predicting adenoma detection at colonoscopy. *The American Journal of Gastroenterology* 102, 4 (Apr 2007), 856–861.

[12] Eli P. Cox. 1980. The Optimal Number of Response Alternatives for a Scale: A Review. *Journal of Marketing Research* 17, 4 (1980), 407–422. http://www.jstor.org/stable/3150495

[13] S. S. Deeb. 2005. The molecular basis of variation in human color vision. *Clinical Genetics* 67, 5 (May 2005), 369–377.

[14] Alan Dix, Janet Finlay, Gregory D. Abowd, and Russell Beale. 2004. *Human-Computer Interaction.* Pearson/Prentice-Hall, Harlow, England New York.

[15] Endoscopic Classification Review Group. 2005. Update on the Paris Classification of Superficial Neoplastic Lesions in the Digestive Tract. *Endoscopy* 37, 06 (2005), 570–578. https://doi.org/10.1055/s-2005-861352 570.

[16] European Colorectal Cancer Screening Guidelines Working Group. 2013. European guidelines for quality assurance in colorectal cancer screening and diagnosis: Overview and introduction to the full Supplement publication. *Endoscopy* 45, 01 (2013), 51–59. https://doi.org/10.1055/s-0032-1325997 51.

[17] Kraig Finstad. 2010. Response Interpolation and Scale Sensitivity: Evidence Against 5-point Scales. *Journal of Usability Studies* 5, 3 (May 2010), 104–110.

[18] M. Ganz, X. Yang, and G. Slabaugh. 2012. Automatic Segmentation of Polyps in Colonoscopic Narrow-Band Imaging Data. *IEEE Transactions on Biomedical Engineering* 59, 8 (Aug 2012), 2144–2151. https://doi.org/10.1109/TBME.2012.2195314

[19] Kelly Creighton Graham and Maria Cvach. 2010. Monitor Alarm Fatigue: Standardizing Use of Physiological Monitors and Decreasing Nuisance Alarms. *American Journal of Critical Care* 19, 1 (2010), 28–34. https://doi.org/10.4037/ajcc2010651

[20] Hajime Hata, Hideki Koike, and Yoichi Sato. 2016. Visual Guidance with Unnoticed Blur Effect *(AVI '16)*. ACM, New York, NY, USA, 28–35. https://doi.org/10.1145/2909132.2909254

[21] Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. 2017. What do we need to build explainable AI systems for the medical domain? *arXiv e-prints* (Dec 2017), arXiv:1712.09923.

[22] Andreas Holzinger, Georg Langs, Helmut Denk, Kurt Zatloukal, and Heimo Müller. 2019. Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 9, 4 (2019), e1312. https://doi.org/10.1002/widm.1312

[23] 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

[24] N Howlader, AM Noone, M Krapcho, D Miller, A Brest, M Yu, J Ruhl, Z Tatalovich, A Mariotto, DR Lewis, HS Chen, EJ Feuer, and KA Cronin. 2018. SEER Cancer Statistics Review, 1975–2016, National Cancer Institute. https://seer.cancer.gov/csr/1975_2016/. [Online; accessed 29-May-2019].

[25] Djenaba A. Joseph, Reinier G. S. Meester, Ann G. Zauber, Diane L. Manninen, Linda Winges, Fred B. Dong, Brandy Peaker, and Marjolein van Ballegooijen. 2016. Colorectal cancer screening: Estimated future colonoscopy need and current volume and capacity. *Cancer* 122, 16 (2016), 2479–2486. https://doi.org/10.1002/cncr.30070 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/cncr.30070

[26] Maurits Clemens Kaptein, Clifford Nass, and Panos Markopoulos. 2010. Powerful and Consistent Analysis of Likert-type Ratingscales *(CHI '10)*. ACM, New York, NY, USA, 2391–2394. https://doi.org/10.1145/1753326.1753686

[27] Matthew Kay and Jacob O. Wobbrock. 2019. *ARTool: Aligned Rank Transform for Nonparametric Factorial ANOVAs.* https://doi.org/10.5281/zenodo.594511 R package version 0.10.6.

[28] Jacob Kittley-Davies, Ahmed Alqaraawi, Rayoung Yang, Enrico Costanza, Alex Rogers, and Sebastian Stein. 2019. Evaluating the Effect of Feedback from Different Computer Vision Processing Stages: A Comparative Lab Study *(CHI '19)*. ACM, New York, NY, USA, Article 43, 12 pages. https://doi.org/10.1145/3290605.3300273

[29] Thomas R. Knapp. 1990. Treating ordinal scales as interval scales: an attempt to resolve the controversy. *Nursing research* 39, 2 (1990), 121–123. https://doi.org/10.1097/00006199-199003000-00019

[30] 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

[31] A. M. Leufkens, M. G. van Oijen, F. P. Vleggaar, and P. D. Siersema. 2012. Factors influencing the miss rate of polyps in a back-to-back colonoscopy study. *Endoscopy* 44, 5 (May 2012), 470–475. https://doi.org/10.1055/s-0031-1291666

[32] J. S. Mandel, J. H. Bond, T. R. Church, D. C. Snover, G. M. Bradley, L. M. Schuman, and F. Ederer. 1993. Reducing mortality from colorectal cancer by screening for fecal occult blood. Minnesota Colon Cancer Control Study. *The New England Journal of Medicine* 328, 19 (May 1993), 1365–1371.

[33] Gary Marcus and Ernest Davis. 2019. *Rebooting AI: Building artificial intelligence we can trust.* Pantheon.

[34] Jon McCormack, Toby Gifford, Patrick Hutchings, Maria Teresa Llano Rodriguez, Matthew Yee-King, and Mark d'Inverno. 2019. In a Silent Way: Communication Between AI and Improvising Musicians Beyond Sound *(CHI '19)*. ACM, New York, NY, USA, Article 38, 11 pages. https://doi.org/10.1145/3290605.3300268

[35] R. A. Miller and F. E. Masarie. 1990. The demise of the "Greek Oracle" model for medical diagnostic systems. *Methods of Information in Medicine* 29, 1 (Jan 1990), 1–2.

[36] 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.

[37] Mark A. Musen, Blackford Middleton, and Robert A. Greenes. 2014. *Clinical Decision-Support Systems.* Springer London, London, 643–674. https://doi.org/10.1007/978-1-4471-4474-8_22

[38] Joshua Newn, Ronal Singh, Fraser Allison, Prashan Madumal, Eduardo Velloso, and Frank Vetere. 2019. Designing Interactions with Intention-Aware Gaze-Enabled Artificial Agents. In *Human-Computer Interaction – INTERACT 2019*. Springer International Publishing, Cham, 255–281.

[39] Changhoon Oh, Jungwoo Song, Jinhan Choi, Seonghyeon Kim, Sungwoo Lee, and Bongwon Suh. 2018. I Lead, You Help but Only with Enough Details: Understanding User Experience of Co-Creation with Artificial Intelligence *(CHI '18)*. ACM, New York, NY, USA, Article 649, 13 pages. https://doi.org/10.1145/3173574.3174223

[40] Minna Pakanen, Jussi Huhtala, and Jonna Häkkilä. 2011. Location Visualization in Social Media Applications *(CHI '11)*. ACM, New York, NY, USA, 2439–2448. https://doi.org/10.1145/1978942.1979298

[41] S. Y. Park, D. Sargent, I. Spofford, K. G. Vosburgh, and Y. A-Rahim. 2012. A Colon Video Analysis Framework for Polyp Detection. *IEEE Transactions on Biomedical Engineering* 59, 5 (May 2012), 1408–1418. https://doi.org/10.1109/TBME.2012.2188397

[42] Hans-Peter Piepho. 2004. An Algorithm for a Letter-Based Representation of All-Pairwise Comparisons. *Journal of Computational and Graphical Statistics* 13, 2 (2004), 456–466.

[43] C. Pox, W. Schmiegel, and M. Classen. 2007. Current status of screening colonoscopy in Europe and in the United States. *Endoscopy* 39, 02 (2007), 168–173.

[44] Pranav Rajpurkar, Jeremy Irvin, Kaylie Zhu, Brand on Yang, Hershel Mehta, Tony Duan, Daisy Ding, Aarti Bagul, Curtis Langlotz, Katie Shpanskaya, Matthew P. Lungren, and Andrew Y. Ng. 2017. CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. *arXiv e-prints* (Nov 2017), arXiv:1711.05225.

[45] Colin J. Rees, Siwan Thomas Gibson, Matt D. Rutter, Phil Baragwanath, Rupert Pullan, Mark Feeney, and Neil Haslam. 2016. UK key performance indicators and quality assurance standards for colonoscopy. *Gut* 65, 12 (2016), 1923–1929. https://doi.org/10.1136/gutjnl-2016-312044

[46] Douglas K. Rex. 2017. Polyp detection at colonoscopy: Endoscopist and technical factors. *Best Practice & Research Clinical Gastroenterology* 31, 4 (2017), 425 – 433. https://doi.org/10.1016/j.bpg.2017.05.010

[47] Shihab Sarwar, Anglin Dent, Kevin Faust, Maxime Richer, Ugljesa Djuric, Randy Van Ommeren, and Phedias Diamandis. 2019. Physician perspectives on integration of artificial intelligence into diagnostic pathology. *npj Digital Medicine* 2, 1 (2019), 28. https://doi.org/10.1038/s41746-019-0106-0

[48] Graham Thomas. 2007. Real-time camera tracking using sports pitch markings. *Journal of Real-Time Image Processing* 2, 2 (01 Nov 2007), 117–132. https://doi.org/10.1007/s11554-007-0041-1

[49] 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. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 21. https://doi.org/10.1145/3359130

[50] Niels van Berkel, Lefteris Papachristos, Anastasia Giachanou, Simo Hosio, and Mikael B. Skov. 2020. A Systematic Assessment of National Artificial Intelligence Policies: Perspectives from the Nordics and Beyond. In *Proceedings of the 11th Nordic Conference on Human-Computer Interaction (NordiCHI '20)*. 1–18.

[51] 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).

[52] 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

[53] Michael Veale, Max Van Kleek, and Reuben Binns. 2018. Fairness and Accountability Design Needs for Algorithmic Support in High-Stakes Public Sector Decision-Making *(CHI '18)*. ACM, New York, NY, USA, Article 440, 14 pages. https://doi.org/10.1145/3173574.3174014

[54] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI *(CHI '19)*. ACM, New York, NY, USA, Article 601, 15 pages. https://doi.org/10.1145/3290605.3300831

[55] Pu Wang, Tyler M Berzin, Jeremy Romek Glissen Brown, Shishira Bharadwaj, Aymeric Becq, Xun Xiao, Peixi Liu, Liangping Li, Yan Song, Di Zhang, Yi Li, Guangre Xu, Mengtian Tu, and Xiaogang Liu. 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

[56] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The Aligned Rank Transform for Nonparametric Factorial Analyses Using Only Anova Procedures *(CHI '11)*. ACM, New York, NY, USA, 143–146. https://doi.org/10.1145/1978942.1978963

[57] Jeremy M. Wolfe and Todd S. Horowitz. 2004. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience* 5, 6 (2004), 495–501. https://doi.org/10.1038/nrn1411

[58] Jeremy M. Wolfe and Todd S. Horowitz. 2017. Five factors that guide attention in visual search. *Nature Human Behaviour* 1, 3 (2017), 0058.

[59] Qian Yang, Aaron Steinfeld, and John Zimmerman. 2019. Unremarkable AI: Fitting Intelligent Decision Support into Critical, Clinical Decision-Making Processes *(CHI '19)*. ACM, New York, NY, USA, Article 238, 11 pages. https://doi.org/10.1145/3290605.3300468

[60] Kun-Hsing Yu, Andrew L Beam, and Isaac S Kohane. 2018. Artificial intelligence in healthcare. *Nature Biomedical Engineering* 2, 10 (2018), 719.

[61] 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 (01 May 2019), 1661–1674.e11. https://doi.org/10.1053/j.gastro.2019.01.260

[62] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW, Article 194 (Nov. 2018), 23 pages. https://doi.org/10.1145/3274463

A SURVEY QUESTIONS

A.1 Demographics

Age _____

Gender

- Woman
- Man
- Non-binary
- Prefer not to disclose
- Prefer to self-describe _____

What is your professional role?

- Gastroenterology Consultant
- Gastroenterology SpR
- Surgical Consultant
- Surgical SpR
- Nurse Endoscopist
- Assistant - Staff Nurse (not performing endoscopy)
- Assistant - Healthcare Assistant (not performing endoscopy)
- Other _____

What is your JAG certification level in colonoscopy? Please indicate:

- None
- Provisional
- Independent (Full certification)
- Other - I am not a UK based colonoscopist

Are you a BCSP accredited colonoscopist?

- Yes
- No
- Other - I am not a UK based colonoscopist

How many years have you been performing colonoscopy? Please indicate:

- 0-1
- 1-2
- 2-5
- 5-10
- 10-15
- >15

Approximately how many colonoscopies have you performed in total? Please indicate:

- ○ 0-100
- ○ 100-200
- ○ 200-500
- ○ 500-1000
- ○ 1000-2500
- ○ >2500

How many colonoscopies do you perform per week on average (diagnostic and therapeutic combined)? Please indicate:

- ○ <5
- ○ 5-10
- ○ 11-15
- ○ 16-20
- ○ 21-25
- ○ >25

Do you know your current polyp detection rate?

- ○ Yes
- ○ No

If yes - please type below: _____

Do you know your current adenoma detection rate?

- ○ Yes
- ○ No

If yes - please type below: _____

How much time do you spend on average during a colonoscopy withdrawal? (approximate answer in minutes)
_____

How often do you spend more than six minutes for a colonoscopy withdrawal?

- ○ Never
- ○ Rarely
- ○ Sometimes
- ○ Usually
- ○ Always

How often do you use dynamic position change during a colonoscopy withdrawal?

- ○ Never
- ○ Rarely
- ○ Sometimes
- ○ Usually

- Always

How often do you use Buscopan (hyoscine butylbromide) during a colonoscopy withdrawal?

- Never
- Rarely
- Sometimes
- Usually
- Always

A.2 AI Perception

How would you describe your willingness to generally incorporate new technology into endoscopy?

- I am not willing to incorporate new advances in the field that were not part of my formal training
- I incorporate new advances in the field once they are incorporated into international guidelines
- I incorporate new advances in the field once I see they have been adopted by leaders in my field
- I incorporate new advances in my field if they are published in highly respected scientific journals
- I try to incorporate new technology into my practice where I can, regardless of where it's published I actively seek out new technology to use in my practice

How excited are you about the development of Artificial Intelligence technology for endoscopy?

- Not excited at all
- Slightly excited
- Moderately excited
- Very excited
- Extremely excited

How concerned are you about the development of Artificial Intelligence technology for endoscopy?

- Not concerned at all
- Slightly concerned
- Moderately concerned
- Very concerned
- Extremely concerned

How likely are you to use validated Artificial Intelligence or Computer-Aided Diagnosis software to help detect polyps if it was available?

- Very unlikely
- Unlikely
- Neutral
- Likely
- Very likely

A.3 Videos previews

In this survey you will be presented with a number of visual markers overlaid on footage obtained from two separate colonoscopies. The videos are about 9 seconds in duration. You are asked to watch each video from start to finish. We have included one of the videos below without overlay. Please watch the video before proceeding to the next page. **This video contains one polyp.**

[Video]

We have included the other video below without overlay. Please watch the video before proceeding to the next page. **This video contains one polyp.**

[Video]

A.4 Videos markers (14 videos – 7 apparent, 7 challenging)

Please watch the above video, showing [design], in its entirety. Please indicate how much you agree with each of the following statements in relation to the video shown above;

[Video]

I would detect more polyps using this visual marker as compared to without this marker.

- Strongly disagree
- Disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Agree
- Strongly agree

I would locate polyps faster using this visual marker as compared to without this marker.

- Strongly disagree
- Disagree
- Somewhat disagree
- Neither agree nor disagree
- Somewhat agree
- Agree
- Strongly agree

This visual marker would interfere during polypectomy.

- Strongly disagree
- Disagree
- Somewhat disagree

- ○ Neither agree nor disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

This visual marker would interfere with the regular visual display.

- ○ Strongly disagree
- ○ Disagree
- ○ Somewhat disagree
- ○ Neither agree nor disagree
- ○ Somewhat agree
- ○ Agree
- ○ Strongly agree

Please provide any thoughts, suggestions, or comments you may have on this visual marker (optional).

_____

A.5 Rank

Rank the following visual markers in order of preference by dragging them up or down (most preferred marker at the top). [7 still images]

Please explain your choice of preferred visual marker (optional). _____

Would this order of preference remain the same for difficult to detect polyps?

- ○ Yes, order of preference remains the same.
- ○ No, I would have a different order of preference.

A.6 Rank colour

Rank the following colour options in order of preference by dragging them up or down (most preferred colour at the top) irrespective of the design of the visual marker. [7 still images]

Please explain your choice of preferred colour option (optional). _____

A.7 Interaction

In addition to the visual marker shown on the screen, should any additional modalities be activated when a potential polyp is detected? Select as many options as desired.

- ○ Yes, in addition to the visual marker I want to hear a sound when a potential polyp is detected.
- ○ Yes, in addition to the visual marker I want to feel a vibration on my skin (e.g., through an armband) when a potential polyp is detected.

- ○ Yes, in addition to the visual marker I want _____
- ○ No, no additional modalities should be included.
- ○ I don't know whether any additional modalities should be included.

Please explain your choice of preferred / no additional modalities (optional). _____

Should the system allow you to turn the visual markers off (and back on again) during colonoscopy?

- ○ Yes, it should be possible to turn the visual markers on/off during colonoscopy.
- ○ No, the system should remain on throughout the entire colonoscopy.
- ○ I don't know whether it should be possible to turn the visual markers on/off system should be able to be turned on and off.

How should you be able to turn the visual markers on and off? Select as many options as desired.

- ○ I want to turn on/off the visual markers using a physical button on the colonoscope.
- ○ I want to turn on/off the visual markers using a physical button on a separate box.
- ○ I want to turn on/off the visual markers using a foot pedal on the floor.
- ○ I want to turn on/off the visual markers using a voice command (as used with e.g. Siri on an iPhone).
- ○ I want the nurse assistant to be able to turn on/off the visual markers.
- ○ I want to be able to turn on/off the visual markers using _____

Please explain your choice with regards to turning the visual markers on and off (optional). _____

Please list situations in which you would turn the visual markers off. _____
What would be your preferred location for displaying the visual marker?

- ○ The visual marker should be overlaid (integrated) on the same one monitor that is used during standard colonoscopy.
- ○ The visual marker should be shown on a separate and directly adjacent monitor, copying over the video from the colonoscope and overlaying the visual marker.
- ○ The visual marker should be shown on _____
- ○ No visual marker should be displayed. Instead, _____
- ○ I don't know where the visual marker should be displayed.