# Understanding the Role of Expert Intuition in Medical Image Annotation: A Cognitive Task Analysis Approach

Florian Leiser
Karlsruhe Institute of Technology
florian.leiser@kit.edu

Simon Warsinsky
Karlsruhe Institute of Technology
simon.warsinsky@kit.edu

Marie Daum
Heidelberg University Hospital
marie.daum@med.uni-heidelberg.de

Manuel Schmidt-Kraepelin
Karlsruhe Institute of Technology
manuel.schmidt-kraepelin@kit.edu

Scott Thiebes
Karlsruhe Institute of Technology
scott.thiebes@kit.edu

Martin Wagner
Heidelberg University Hospital
martin.wagner@med.uni-heidelberg.de

Ali Sunyaev
Karlsruhe Institute of Technology
sunyaev@kit.edu

**Abstract**

*To improve contemporary machine learning (ML) models, research is increasingly looking at tapping in and incorporating the knowledge of domain experts. However, expert knowledge often relies on intuition, which is difficult to formalize for incorporation into ML models. Against this backdrop, we investigate the role of intuition in the context of expert medical image annotation. We apply a cognitive task analysis approach, where we observe and interview six expert medical image annotators to gain insights into pertinent decision cues and the role of intuition during annotation. Our results show that intuition plays an important role in various steps of the medical image annotation process, particularly in the appraisals of very easy or very difficult images, and in case purely cognitive appraisals remain inconclusive. Overall, we contribute to a better understanding of expert intuition in medical image annotation and provide possible interfaces to incorporate said intuition into ML models.*

**Keywords:** medical image annotation, intuition, expert knowledge, cognitive task analysis, machine learning

## 1. Introduction

Contemporary machine learning (ML) models are increasingly able to assist healthcare professionals, for example, through automatic diagnosis of diseases (Pandl et al., 2021), or by assisting with camera control in surgeries (Wagner et al., 2021). To achieve such feats, ML models usually require large amounts of annotated images (Litjens et al., 2017). The annotation of these medical images is performed by medical experts with pertinent domain knowledge to ensure data quality (Litjens et al., 2017). Medical professionals hold a lot of expertise, which can be combined with ML models to improve the models' predictive accuracy (von Rueden et al., 2021). An exemplary approach included segmentation masks of unseen but related data sets which were provided by experts (Wang et al., 2020). Incorporating expert knowledge bears great potential to provide faster, more accurate or more robust predictions (Wang et al., 2020), and easier interpretation, especially if training data is limited (von Rueden et al., 2021). However, expert knowledge is usually tacit and has to be formalized for ML models (von Rueden et al., 2021).

A key aspect that characterizes expert judgments and inferences is intuition (Kahneman, 2011; Kahneman & Klein, 2009). Intuition broadly describes decision-making behavior where an individual is unable to describe in detail the reasoning or other processes that produced the answer (Simon, 1992). Research has argued that expertise is one of the main causes of intuition (Dane & Pratt, 2009), with some going as far as stating that it is precisely intuition that sets apart expert judgments from judgments made by non-experts or machines (Baylor, 2001; Kahneman, 2011). Thus, intuition also plays an important role in expert medical image annotation. This is further substantiated by looking at the nature of expert annotation tasks. Annotators often face tremendous workload in terms of the number of images they must annotate, and ground truth annotations are often not present (Ørting et al., 2020). Instead of being able to leisurely form a decision on the basis of ordered rational analyses, annotators are often forced to make fast decisions and complex judgments, where intuition thrives due to its relative speed compared to conscious cognitive processes (Akinci & Sadler-Smith, 2012; Simon, 1992). Thus far, research on expert medical image annotation tasks has mostly focused on identifying potentially mislabeled instances ex post (Rädsch et al., 2021) or improving the annotation process itself (Warsinsky et al., 2022). There

URI: https://hdl.handle.net/10125/102982
978-0-9981331-6-4

have also been some investigations into how expert annotators arrive at their judgments, for example, by identifying different annotation styles (i.e., cognitive vs. intuitive; Chang et al., 2022) or by having annotators mark the regions of interest they mainly looked at while annotating (Ørting et al., 2020). Some research also highlights the importance of a "gut feeling" (Freeman et al., 2021), thus hinting toward intuition. However, most of these studies stop at mentioning intuition and do not further tease out its role. Although intuition is an important contributor to experts' judgments, we currently lack knowledge of how expert (medical image) annotators draw on their intuition to make annotation judgments and inferences. Accordingly, investigating the role of intuition is an important step to better understand and formalize expert knowledge to ultimately apply it to ML models. We therefore ask: *How do experts rely on their intuition in medical image annotation?*

To answer this research question, we follow a cognitive task analysis (CTA) approach where we observe and interview six expert medical image annotators to identify cues that give us insights into their judgments and inferences about their use of intuition. We contribute to extant research in three ways. First, we contribute to ongoing research efforts on leveraging expert knowledge to improve ML models (von Rueden et al., 2021) by making intuition as a key component of expert knowledge more palpable and thus more readily applicable to ML pipelines. Second, we contribute to a better understanding of experts' judgments during medical image annotation, which may help shape annotation environments in a way that improves annotation quality and thus ultimately improve ML models in healthcare. Lastly, by using CTA methods to understand expert intuition in this particular context (i.e., expert medical image annotation) we contribute to a better understanding of how CTA methods can be applied to capture expert intuition.

## 2. Background

### 2.1. Expert Medical Image Annotation

In ML, annotation refers to the addition of metadata to existing data instances with the goal of making it easier for ML models to recognize patterns and make inferences (Pustejovsky & Stubbs, 2013). Annotation usually marks the first important step in an ML pipeline (von Rueden et al., 2021). Typical medical image annotation tasks include tracing anatomical structures (i.e., segmentation) in intra-surgical images or tagging pathologies in CT images (Ørting et al., 2020). Medical image annotation is challenging. Many medical images features exist on a continuum, which is why objective ground truths (e.g., "disease X is definitely present") are often difficult if not impossible to make (Litjens et al., 2017). Efforts to crowdsource or to automate annotation tasks exist and are somewhat successful on simple tasks like instrument annotation (Ørting et al., 2020). However, to achieve ample annotation quality, medical experts with some level of clinical experience should annotate complex medical images (Ward et al., 2021).

Existing research investigated how expert medical image annotators come to their judgments. These studies identified factors like confidence or the level of expertise (Ward et al., 2021) as important drivers. Some studies also hint at intuition as they speak of intuitive annotation styles (Chang et al., 2022) or how some medical image annotators rely on their "gut feeling" (Freeman et al., 2021). We seek to build on these indications and aim to further tease out the role of intuition in expert medical image annotation.

### 2.2. Expert Intuition

The concept of intuition is the subject of scholarly inquiry in various domains, including management (Akinci & Sadler-Smith, 2012), healthcare (Campbell & Angeli, 2019), and finance (Hensman & Sadler-Smith, 2011). Given the tacit nature of intuition, no unified definition exists. Dictionary defines the term 'intuition' as the "power or faculty of attaining to direct knowledge or cognition without evident rational thought and inference" (Merriam-Webster, n.d.). Beyond that, several schools of thought on intuition have formed (for an overview see Akinci & Sadler-Smith, 2012).

Intuition is notoriously difficult to formalize and measure (Dane & Pratt, 2009). Research has, however, identified attributes that characterize intuition (Chilcote, 2017). To begin with, intuition is usually subconscious, quick (Hammond, 1996), and "the person who is experiencing intuition does so without using a rational, analytical process" (Chilcote, 2017, p.64). In fact, there is an antithetical relationship between intuition and conscious thought processes; that is, they impede each other (Baylor, 2001). Some authors argue that intuition simply involves building patterns that enable rapid decisions (Simon, 1992). Others emphasize that intuition is more than pattern recognition, and involves creatively combining elements to produce new solutions (Dane & Pratt, 2009). Regarding the quality of decisions, research recognizes intuition as a troublesome decision tool, because "expert intuition is sometimes remarkably accurate and sometimes off the mark" (Kahneman & Klein, 2009, p.515). Yet, intuition is often accompanied by an overwhelming feeling of certainty (Hammond, 1996). A major source of intuition is expertise. With increasingly advanced knowledge in an area, individuals are able to make more higher order

intuitive connections—hence, intuition is often fueled by highly specific domain knowledge (Baylor, 2001). Overall, intuition is a tacit concept where judgments are done seemingly without rational thought but can still be remarkably accurate. To grasp intuition in expert medical image annotation tasks, this study draws on the presented attributes of intuition.

## 3. Methods

Our approach draws from Cognitive Task Analysis (CTA), a family of methods used to study and describe reasoning and knowledge (Crandall et al., 2006). CTA builds on the premise that one cannot expect decision makers to accurately explain why they made decisions, and accordingly provides methods for making inferences about pertinent judgment and decision processes (Crandall et al., 2006). The three phases of CTA are (1) knowledge elicitation (i.e., obtaining information on what people know and how they know it), (2) data analysis (i.e., structuring data, identifying findings and discovering meaning) and (3) knowledge representation (i.e., displaying data and communicating meaning). Various approaches exist for each phase. Our approach is visualized in Figure 1.

Our main goal was to identify cues, which are stimuli that trigger actions (Okoli et al., 2022). Relevant actions in our case were specific annotation decisions. Identifying cues is an effective practice in understanding experts' judgments (Okoli et al., 2022) that has been applied in trying to understand expert intuition (Crandall et al., 2006). We thus deem cues a suitable approach to better understand experts' judgment in medical image annotation and infer about their use of intuition.

### 3.1. Knowledge Elicitation

For the first step of a CTA inquiry, knowledge elicitation, we engaged with six experts in medical image annotation. We engaged with each participant in a 45-minute online session, during which we recorded audio and video. In the first 20 minutes, we asked them to annotate a prepared set of 20 medical images during which we observed them and had them self-report their behavior. We ended this first part if either 20 minutes elapsed, or all images were annotated. We then followed up with a brief questionnaire about their annotation experience and demographics (5 minutes), and a semi-structured interview (20 minutes) to gain additional insights on their judgments. Our participants were all medical students from the same annotation team, aged between 22 and 24 and annotated medical images at least once a week. Three of them had less than six months of experience in medical image annotation, whereas the other three had over one year of experience.

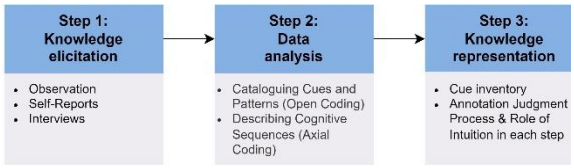


**Figure 1. Overview of applied CTA approach.**

All experts were initially introduced to the task by a supervisor and had an annotation protocol containing assisting guidelines during the task. The annotation task we gave our experts was the tagging of intraoperative images from robotically assisted esophagectomies. They used a web-based image annotation tool to choose one tag to signify the level of smokiness in an image (available tags: 0-no smoke, 1-smoke-small, 2-smoke-increased, 3-no visibility) and one tag to signify the level of bloodiness in an image (available tags: 0-no blood, 1-small amount of blood, 2-blood accumulation, 3-blood great, 4-intervention required). We chose this task as it was routinely performed by our experts and would thus allow us to observe annotation in a natural setting. Another reason was that the surgery is sufficiently complex to force annotators to make complex judgments, a situation where intuitive mental processes are common (Bastick, 1982).

Images were selected by a medical doctor experienced in surgical image annotation to represent a variety of difficulty levels. Images were considered difficult to annotate if structures were not clearly identifiable due to image distortion, light under- or overexposure or image artifacts like camera smudges or light reflection. Smoke was not considered an artifact. Images were assigned to one of three levels (see Fig. 2), with all levels of difficulty distributed across all images.

While annotating, we closely observed our participants' annotation behaviors regarding exhibited cues and encouraged our annotators to self-report their thought processes. After finishing the annotation task, we asked them to complete a questionnaire about subjective workload (measured through the NASA Task Load Index; Hart & Staveland, 1988), familiarity with the annotation tool, and some demographic information (e.g., age, gender). We also included the short version of the Rational Experiential Inventory (REI-10; Norris et al., 1998) to measure participants' tendencies to trust

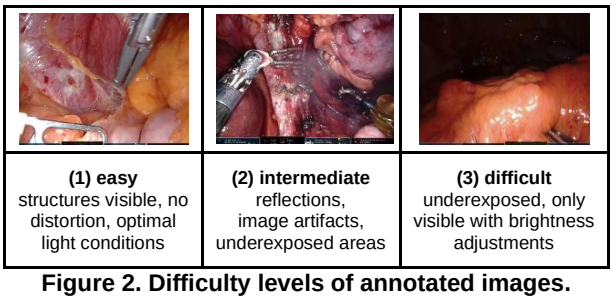


**Figure 2. Difficulty levels of annotated images.**

their intuition. All questions of the questionnaire can be found in online supplement at https://bit.ly/3BeL61T.

Finally, we conducted interviews with our participants. The structure of these interviews was based on the observations we made during the annotation session and was thus highly dynamic. We took inspiration in the list of cognitive probes proposed by Klein et al. (1986) and then tried to adapt these to each individual interviewee to learn more about their judgments and exhibited cues. Frequent questions included "What knowledge helped you in your annotation decision?" or "Why did you stall here?" A detailed description of the prepared interviews questions together with information on the gathered measures and the questionnaire can be found in the online supplement.

### 3.2. Data Analysis

We started the data analysis step by looking at the provided tags. We measured the intra-rater reliability between our experts and time spent on individual images. We then aggregated the questionnaire data and calculated descriptive statistics to learn about our participants' individual workloads, demographics, and traits. To analyze the self-reports and interviews, we transcribed them and conducted open coding (Myers, 2020) to identify text passages that deal with cues used and the knowledge required to assess specific cues during annotation. We found 666 relevant text passages that we compiled into an inventory of eleven cues.

With the cue inventory as our basis, we then performed axial coding (Myers, 2020) to identify text passages describing cognitive sequences. The goal of this step was to heighten our abstraction level beyond individual cues and to identify major patterns occurring during the annotation, which we would then be able to evaluate with respect to the use of intuition. To this end, we focused on subconscious, quick, and certain decisions, which would fulfill attributes of intuition (Hammond, 1996). We paid particularly close attention to our annotators' thought processes in images with low intra-rater agreement (indicating subjective decisions) or large discrepancies in the time taken to annotate the specific image (indicating quick decisions). We converged codes, triangulated data (e.g., annotation time, observations during the annotation) where appropriate, and iteratively discussed our observations within the author team, which led us to a process of cognitive sequences. With frequent indications on intuition, we achieved an improved understanding of experts' judgment process in medical image annotation.

### 3.3 Knowledge Representation

For the final CTA step, *knowledge representation*, we first present a cue inventory with eleven cues (cf. section 4.2). Using this cue inventory, we describe the cognitive sequences of our experts. Additionally, we identified four steps in the annotation process of our expert medical image annotators. We present each step, including frequently used cues and the role of intuition (cf. section 4.3).

## 4. Results

### 4.1. Descriptive Statistics

Our experts provided on average 29.36 tags (i.e., annotated 14.68 images) and took on average 82.91 seconds per image (min: 15, max: 300, SD: 62.99). Three annotators managed to annotate all 20 images in the given time; but only eight images were fully tagged by all annotators. To measure intra-rater reliability for these images, we calculated Fleiss' κ (Fleiss, 1971). For smoke, the annotators showed almost perfect agreement (κ=0.8886), while the agreement was only moderate for blood (κ=0.4849). Regarding participants' tendencies captured by REI-10, they rated their need for cognition on average as 3.53, and their faith in intuition as 3.67. This shows a similar need for cognition and slightly less faith in intuition compared to studies assessing a general population (3.51 and 3.77) (Golley et al., 2015). The annotators rated the task difficulty as medium (mean: 3.5, assessed by NASA-TLX). Annotators with higher experience in medical image annotation (more than 1 year) stated a low mental load (2.3) and a high familiarity with the annotation tool (7.0) while annotators with less experience had medium mental load (4.0) and medium tool familiarity (4.6).

### 4.2. Cues in Expert Medical Image Annotation

Our analysis of cues revealed eleven cues used in expert medical image annotation to come to judgments. We grouped these across four categories based on the required knowledge into: (1) surgical, (2) anatomical, (3) technical, and (4) annotator cues. An overview of all identified cues and the information the annotators drew from them is shown in Table 1.

**Surgical Cues**. The first cues that emerged were those where annotators drew on their knowledge from practical experience in surgery rooms or discussions with surgeons. This knowledge manifested in three cues being the *presence of instruments*, the *surgical field*, and the *perceived danger to the patient*. These cues gave annotators valuable insights into the overall surgical context (e.g., the current surgery phase), which they used for example to judge the plausibility of smoke in a surgery phase: "Here, by the way, I see an electric hook. But since it has barely started to work here in the new step of the surgery, I don't see any smoke." (i01 – Cues:

**Table 1. Principal Cue Inventory for Blood & Smoke Tag Annotation**

| | Cue | Information Generated | Exemplary Quote |
|---|---|---|---|
| Surgical Cues | Presence of Instruments | Identification of surgery phase or presence of smoke (i.e., if instruments can produce smoke) | "There are no instruments that produce smoke, it is rather unlikely that there is smoke on this image." (i02) |
| | Surgical Field | Possible disruptions of surgical process due to blood pools or smoke in the view of surgeon | "Yeah, whether he's […] keep doing the surgery or whether he's really going to take care of the blood now." (i03) |
| | Perceived Danger to Patient | Judgment on urgency of surgical situation (i.e., required intervention) | "Yes, whether you have to intervene or not, or how dangerous the situation seems to be for the patient if you do nothing." (i03) |
| Anatomical Cues | Depth of Blood Pools | Amount of accumulated blood as seen in the image | "Blood pools which are deeper than just on the surface of the tissue, so it would also be clear to me here that it is accumulation." (i06) |
| | Concealed Structures | Potential causes for bleeding occluded by instruments or smoke | "But still overall, if something important is not visible, it's a problem." (i05) |
| | Anatomical Expectations | Possible anatomical deviation (e.g., in color) indicating an injury | "Regarding the blood level here behind the spleen would already be a good collection in such a dark red blood pool." (i03) |
| Technical Cues | Image Quality | Differentiation of smoke to artifacts & benefits of more detailed investigation | "That is not smoke in any way; rather that is simply the image quality." (i06) |
| | Camera Perspective | Assessment of position and size of present structures (e.g., organs, blood pools) | "If it's an overview image of the whole situs and it's sort of this size of blood, it's a lot of blood." (i01) |
| Annotator Cues | Perceived Difficulty of Image | Indication on the need of additional information for the judgment | "Sometimes it is very difficult to tell with Smoke because you don't have the frames from before and after." (i06) |
| | Overall Impression | Shape initial annotation judgment | "You develop a feeling for: 'What is the impression of the image on me?' [...] and that is very beneficial to you." (i06) |
| | Tag Distribution | Shape expectations on amount of specific annotation decisions | "90% of the images, it's either there's a pool of blood and you know that's either [level] 2 or 3." (i02) |

*Presence of Instruments, Surgical Field)*. The cues there (i.e., presence of instruments) are rather objective and should not differ between experts. Our experts also often tried to put themselves in the position of the surgeon to gauge the implications of their annotations with respect to *perceived dangers to patient safety*. This cue heavily relies on subjective judgment. They remarked that at times, this may have led to more cautious annotations to ensure their annotation 'aligned' with ensuring patient safety throughout the surgery: "in an O.R. someone said this doesn't need an intervention, even though it does, that would be a lot more problematic than if you did a little suction" (i05 – *Perceived* Danger to Patient).

**Anatomical Cues**. Annotators also drew cues based on their knowledge about anatomical structures in the images. All three cues in this category (depth of blood pools, concealed structures, and anatomical expectations) influence the experts' annotation of blood in the image by using their knowledge about the human anatomy. For example, indentations in tissues provide information on whether blood could accumulate in that area. While the shape of indentations and therefore the potential topology of blood pools is objectively measurable in theory, this assessment is difficult in many cases: "You can't say [the blood pool is] at least 1 cm deep, of course you can't estimate that, but it has to include a certain depth" (i06 – *Depth of Blood Pools*).

An assessment of the depth of blood pools could also rely on subjective judgment. Especially in dark images, we found that annotators struggled to differentiate between blood pools and other dark regions like the liver or image artifacts like shadows. This shows that anatomical cues do neither solely rely on subjective nor solely on objective assessment but that the assessment strongly depends on the context. For the anatomical cues only medical training and no surgical experience is required to provide useful information, which was beneficial for less experienced annotators.

**Technical Cues.** Annotators also considered technical metadata of the provided image for their judgment. Not only domain-specific context but also relatively objectively measurable meta characteristics like *image quality* or the *camera perspective* provided valuable insights. Annotators often considered the quality of an image to assess whether it contains smoke or is just ill-captured: "That the image here is clouded by smoke, and the fact that it is also very global also suggests that it really is smoke" (i06 – *Image Quality*). Our annotators also took the perspective of the camera into account, particularly to assess possibly obstructed structures: "So you can't see the structures behind [the blood pool] and since you would have to see them, you would have to remove the blood here. Definitely." (i05 – *Camera Perspective, Concealed Structures*). On weird camera angles or zoom, they often sought an anchor point to judge the size of structures: "What you can orient yourself on well here is for example this stapler head. I know about how big it is, so I can imagine how

big this blood pool here should be" (i06 – *Camera Perspective, Presence of Instruments*).

**Annotator Cues.** Lastly, our annotators drew on several cues driven by their own, subjective, annotation experiences like *overall impression* of the image, *tag distribution* and *perceived difficulty of the image*. The tag distribution of all frames is objectively measurable, however, annotators heavily relied on the most recent annotations: "You are fooled by that [situation] because you've already annotated the previous 5 frames as great and it's just still great." (i06 – *Tag Distribution*). The overall impression of an image played a crucial role in its first assessment, as was stated by one interviewee: "Exactly, for smoke now first look at the overall impression of the picture" (i06 – *Overall Impression*).

## 4.3. Role of Intuition in Expert Medical Image Annotation

After synthesizing the cue inventory, we investigated how these cues shape the judgment of experts in medical image annotation. We identified that the annotation judgment process of our annotators consisted of four distinct steps, which we summarize in Figure 3. While we present the judgment in a mostly linear way to ease the reader's understanding, we emphasize that this is a simplification; while this may reflect reality in some cases it is also possible for this to be a more iterative process. For each image, annotators would first (1) make a purely intuitive appraisal of an image, then (2) a quasi-rational appraisal whether it is useful to expend further cognitive resources, and if so (3) do a cognitive appraisal to finally (4) triangulate all previous insights to arrive at a final annotation judgment. At the end of each step, annotators appraised whether they had reached sufficient confidence in their current judgment. If they did, they made their annotation decision and set the corresponding tags; if not, they moved to the next step. Thus, the length and depth of each annotation judgment may vary: "So I have to say [the annotation time] varies a lot. With some images I'm done with it within a second and with some images, I think about it for minutes and then sometimes even go back to it later" (i06). Each step of the process came with its own frequently used cues and indications of intuition (see Figure 3), which we now describe.

**4.3.1. Step 1: Intuitive Appraisal.** Whenever annotators started annotating a new image, they especially relied on their *overall impression* and the *perceived difficulty of the image*. The annotators usually first used their overall impressions of the image as an important cue to identify easy annotation judgments: "Yes, [which tag I start with] depends on what is a clearer decision. If I see that there is no smoke at all then I say smoke is 0 and then I can think about blood." (i05 – *Overall Impression*) In general, our annotators emphasized the value of this first judgment when working with experts: "I think [one's own first judgment] is very beneficial, but I think it is very important that [the annotators] are real experts" (i06). Our annotators also remarked that they often used their first judgment to swiftly deal with easy-to-annotate images: "The thing is simple problems are easy to solve, but then they are over quickly. With complex [problems] you invest your time and in the end, you get something out" (i04 – *Perceived Difficulty of the Image*). This judgment was often accompanied by an overwhelming sense of certainty which is characteristic for intuition. One annotator explained that this certainty can occur when the *quality of the image* is high: "Oh yeah, here I would… there is definitely no smoke here, very clearly" (i05 – *Image Quality*). Since this first step is characterized by quick, subconscious, and certain judgments, this suggests a high use of intuition and we therefore named this first step *Intuitive Appraisal*. If our annotators reached a sense of certainty, they made their annotation decision. If not, they moved to step 2.

**4.3.2. Step 2: Quasi-Rational Appraisal of Cognition's Usefulness.** If annotators did not feel sufficiently confident in their intuitive appraisal, they would next appraise if it was worthwhile to spend additional cognitive resources on an image, mostly based on their perceived *difficulty of the image*. We found this appraisal to be quasi-rational; that is, it involves both intuition and cognition. During the observation, we noticed that the choice of the cues which are relevant for this step happens mostly without evident rational thought. We observed this, when we asked one annotator how they came to their conclusion: "Yeah, for example, that you just can't go to... I too... damn... no idea." (i03). The pattern that emerged was that annotators would provide tags intuitively for images they perceived as easy to preserve their cognitive resources for more difficult images: "When I have to concentrate more on an image, so I have to think more now 'okay what am I going to annotate?' So I want to work through the other problems quickly before I concentrate on that now" (i01 – *Perceived Difficulty of Image*). They needed to assess the image difficulty very quickly to make a choice which cues to use in their judgment. This choice again relies on very subjective appraisal and is usually accompanied by a certain sense of confidence, which indicates the use of intuition.
Meanwhile, the actual appraisal of difficulty happened mostly based on rational thoughts by using cues like occluded structures or the image quality in general. As one annotator said during the self-report: "Yes, I also find it difficult because you don't know what else would

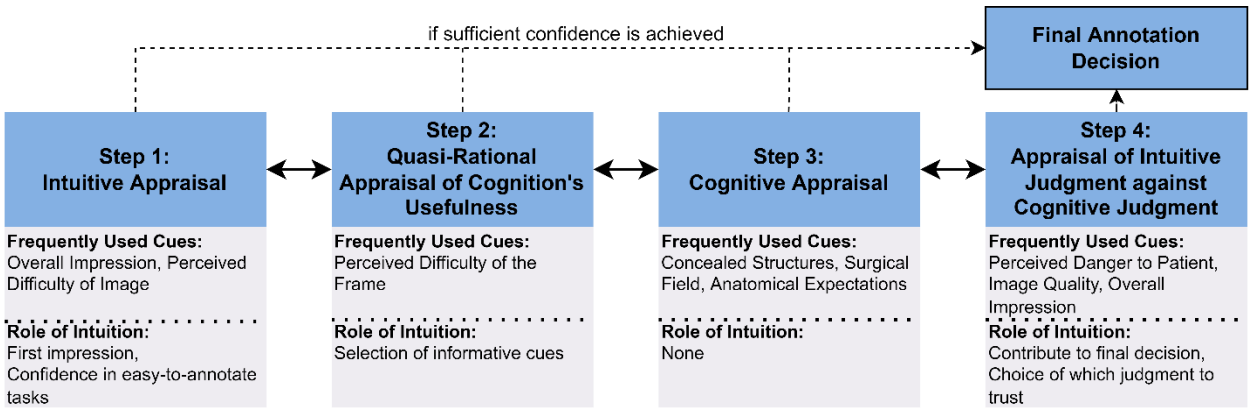


**Figure 3. Overview of Annotation Judgment Process and the Role of Intuition in Each Step**

come behind it if you panned the camera" (i03 – *Camera Perspective*). If annotators appraised an image as too difficult, they also stuck with their subconscious, intuitive judgment, albeit for a different reason. On very difficult images, annotation processes can get very lengthy, and as annotators usually face a high workload, opting to carefully rationalize each image may not be feasible: "In the beginning, of course, I looked into the annotation guidelines for every image, but at some point [...] you have to rely on yourself a bit. It's impossible to look at every picture with the annotation guidelines." (i04). Therefore, one role of intuition in this step is to support in assessing the information gain of cues. If the annotators concluded that their first judgment provided only insufficient confidence and further investigation might be beneficial, they continued to step 3.

**4.3.3. Step 3: Cognitive Appraisal.** In the third step we found that annotators assessed cues often in parallel by taking multiple cues into account, which usually resulted in a lengthy and analytical reasoning process. As our annotators started their assessment in this step, they often drew on the surgical and anatomical cues to orient themselves and help assess other cues. For example, our annotators would often assess their *anatomical expectations* to gain an impression of 'normal' amounts of blood in an image: "This is a phase of the surgery where there always is a lot of blood" (i01 – *Anatomical Expectations, Depth of Blood Pools*). In this step, annotators investigated multiple cues in more detail like *concealed structures*, the *surgical field* or *anatomical expectations*. Therefore, we argue, the judgment in this step is mostly reached by cognition, hence the step is named *Cognitive Appraisal*. Most annotators presented the attitude that a cognition-based decision can achieve higher confidence and therefore, overrules the intuitive judgment. However, cognitive judgment took longer than intuitive judgment. In difficult images, we observed that annotators spent on average 7.44% longer compared to easy images. To this end, our annotators themselves also remarked that it is important to take ample time for the cognitive appraisal in this step: "If you haven't thought about it enough or haven't looked at the image thoroughly enough or haven't looked at the rules thoroughly enough, then mistakes can happen" (i06).

However, even when taking time, in some cases, a cognition-based decision could not surpass the required confidence threshold. This was especially true if the provided decision guidelines were unclear or criteria for multiple levels were met. As one of our interviewees put it: "For me, the problem lies in the definitions; we have defined the different [blood and smoke] levels but sometimes definitions fit multiple classes" (i06). If annotators' cognitive appraisal cannot reach a certainty threshold, they move to the final step of their judgment.

**4.3.4. Step 4: Appraisal of Intuitive Judgment against Cognitive Judgment.** In the final step, annotators triangulated their own intuitive and cognitive appraisals to combine all available information and cues for the most sophisticated outcome: "Then I try to match two things what I am told in my head by facts and by examples and then one that trusts my intuitions" (i04). Our annotators continued in different ways, depending on whether the intuitive and cognitive judgments coincide, contradict, or are both unsatisfactory. If both judgments coincide, the tag was provided without an insight on which judgment was used. In this case, we also observed some remorse in our annotators, as they felt that they could have avoided the lengthy cognitive appraisal just to ultimately come to the same conclusion. As one of the interviewees stated: "Sometimes I get this thought like 'you have to look in the annotation guidelines', and then when I looked 'you were right, you could [have] annotated the way you thought'" (i04).

The same interviewee then continued to justify the comparison with the annotation guidelines: "But there are also moments you need to look into the annotation guidelines, and it is good that you looked into the

annotation guidelines." (i04). This indicates that when intuitive and cognitive judgments are contradictory, annotators may be more confident in their cognitive judgments which might be caused by the longer time annotators spent in the cognitive appraisal. However, in conflicting cases, annotators may also choose to follow their intuitive judgment and trust their *overall impression*: "But here I have blood [...] with a certain depth, that's what the protocol says, and you could argue that way, of course. But for me the overall picture is more decisive, so to speak, in the truest sense, and I then decided in favor of small" (i06 – *Overall Impression*).

Sometimes neither the cognitive nor the intuitive judgment surpassed the certainty threshold, leaving an annotator with two unsatisfying judgments. In this case, annotators sometimes felt pressured, as they ultimately had to decide and set tags. At this point, the final triangulation of the different appraisals was highly subjective. Once this triangulation was concluded, the annotators stated their judgments with certainty, often seemingly arbitrary; for example: "I have no idea. Can you actually already say that there is a very light smoke background here? Let's make a 1 for the smoke here" (i03). This subjective yet very certain appraisal is an indicator for the use of intuition. The bottom line of this step is that annotators use both their intuitive and cognitive resources to produce an annotation decision. If both appraisals contradict, they rely on intuition to decide which judgment they trust more.

# 5. Discussion

## 5.1. Principal Findings

Overall, our study provides insights on the judgment process of experts in medical image annotation and especially the role of intuition therein. Our findings indicate a role of intuition in three out of the four identified steps of the expert medical image annotation process. The results also suggest a potential of integrating intuition into ML pipelines, especially more subjective cues. At the same time, we also saw some skepticism of our experts toward intuition.

First, we identified a four-step-approach that experts follow when assessing the bloodiness and smokiness of medical images in our study. We found that our annotators' approach was mostly guided by the level of confidence in their judgments and majorly influenced by intuition. Depending on where in the judgment process the annotators achieve this confidence, the final assessment relies more on intuition (confidence achieved in step 1 or 2), cognition (confidence in step 3) or a combination of both (step 4). These findings are consistent with works describing decision making as a combination of intuition and cognition (Baylor, 2001) and findings that especially rapid decisions rely on intuition (Chilcote, 2017). Our results indicate that intuition shapes the annotation judgment especially in tasks that are perceived as easy and thus quickly assessed, or in difficult tasks, where neither cognitive nor intuitive judgment affords a certain judgment. This implies a U-shaped relationship between task difficulty and the amount of intuitive thinking an individual engages in at a time, complementing the U-Net theory proposed by Baylor (2001). Specifically, our findings suggest that when dealing with experts (who should be able to apply a lot of intuitive thinking), a 'second U' exists that determines the use of intuition based on task difficulty. To this end, our findings ultimately suggest that intuition comes into play when a task is very easy or very difficult, but less in between.

Second, our results also give insights into those decision cues that are assessed through intuitive thinking. We found that especially 'soft' cues, which required a more subjective judgment, were often assessed through intuition. As such, when trying to integrate expert medical image annotators' knowledge into ML models, these soft cues may pose as interfaces to experts' intuition—they do however differ in how easy they seem to be formalizable. For example, while the cue perceived danger to patient can probably be meaningfully transformed into a numerical value (which could be added as a tag to an image), formalizing cues like overall impression of the image will be more difficult. This highlights that it may ultimately not be expedient to aim for a comprehensive integration of expert medical image annotators' intuition into ML models. Rather, ML model designers might look at cues based on intuition already in the feature selection process of ML pipelines (von Rueden et al., 2021).

Third, we also got some insights into our participants' personal stances toward intuition. In particular, we observed a dilemma in our participants: While they remarked on the general usefulness of intuition in their annotation judgments, they also exhibited a general aversion to using intuition, as they felt it was not rational enough to guide their annotation judgments. With respect to the results from the REI-10 questionnaire, we also observed a smaller faith in intuition (3.67) in our annotators compared to more general populations (e.g., Australians with a score of 3.77; Golley et al., 2015). We think this is interesting, as it highlights the role of intuition as a "troublesome decision tool" (Kahneman & Klein, 2009) across contexts. The medical background of our annotators may have further exacerbated this circumstance. Medicine is an inherently very serious context with the prevailing attitude that topics deserve a highly serious and rational approach (Furnham et al., 2013). However, intuition is all but that, and may thus be dissonant with

the medical context, which may explain our annotators' aversion to intuition. Another explanation could be social desirability bias (i.e., the tendency to answer questions in a manner that is viewed favorably by others; Nederhof, 1985). Our annotators may have had reservations about telling us that they "do not think properly" about their decisions, but rather use intuition.

### 5.2 Implications

For research, our findings first and foremost imply that scholars should be mindful that expert medical image annotators' judgments may be substantially shaped by intuition. Hence, scholars should move beyond the thought of medical image annotation as a highly analytic process based on only facts and well-defined rational thoughts. Our findings also imply that there may not only be differences in how intuition is used by experts and non-experts (as proposed by U-Net theory; Baylor, 2001), but also differences in how individual experts use their intuition based on task difficulty. To this end, we find that experts will use their intuition mostly on either very easy, or very difficult tasks. During tasks with medium difficulty, the experts often reached sufficient confidence during their cognitive appraisal, indicating more rational and less intuitive decisions. With respect to integrating expert knowledge into ML models, our findings imply some cues may serve as interfaces to tap into expert intuition, indicating ways to incorporate experts' intuition into ML models, if research focuses on the more subjective cues. Our study provides some first steps on this arduous journey of integrating expert intuition into ML models.

Our findings also affirm that CTA-based methods are useful to investigate intuition. We particularly exemplify benefits of triangulating data from multiple sources to study intuitive judgment processes. For example, we especially found data which was not explicitly verbalized by the experts (e.g., hesitations during annotations or gazes away from the screen) beneficial, as this gave us insights on subconscious processes, which are difficult to capture with interviews.

For practice, our findings imply that even when providing expert medical image annotators with detailed, rational guidelines on how to shape their annotation judgments, they may still do so by drawing on their intuition. To this end, practitioners should be mindful of the aforementioned U-shaped relationship between task difficulty and the use of intuition. Hence, group discussions should reflect the range of difficulty across images. These discussions of difficult images foster the exchange of rational thoughts although the decisions were often made intuitively. Our study can contribute to improved group discussions by raising awareness on intuition in medical image annotation and providing a list of cues in decision making of experts.

### 5.3. Limitations and Future Research

W e acknowledge several limitations of our study that also pave ways for future research. First, given that we investigated intuition as a particular tacit phenomenon only in one study setting, we encourage the need to treat our results with some caution. While we tried to apply CTA-based methods to increase the rigor of our approach, we acknowledge that intuition is a phenomenon that is inherently fuzzy and thus greatly benefits from investigations from multiple angles. As such, future research may find it beneficial to tune parameters of our CTA approach and for example focus on particularly challenging tasks (Crandall et al., 2006). While out of scope for this study, we want to highlight the potential benefits of physiological measurements such as eye-tracking technologies to capture implicit cues that signalize intuitive thinking processes.

Second, our derived cues are strongly bound to our research context, the blood and smoke annotation of intra-surgical images from robotically assisted esophagectomies. While we think that the resulting insights on intuition are abstract enough to be generalizable to some degree, we acknowledge that future research is necessary to test this assumption. Research may thus find it useful to investigate whether our findings on intuition hold in other expert medical image annotation tasks (e.g., segmentation tasks).

Lastly, while the results help us understand intuition in expert medical image annotation, we acknowledge that our results are rather abstract and thus difficult for ML model designers to turn into action. A next step would be to formalize cues relying on intuition in a way that it is applicable to ML pipelines. This could result for example in transforming the surgical field into attention maps. How this formalization of each cue could be conducted needs further research. Researchers may also want to think about using our results as a basis to formulate actionable prescriptions on how to shape annotation environments so that intuition can thrive.

## 6. Conclusion

In this study, we investigated how experts rely on their intuition in medical image annotation tasks. Following a CTA-based research approach, we observed and interviewed six expert medical image annotators to identify a set of eleven cues, based on which we investigated the role of intuition in the annotation process. Our findings suggest that expert medical image annotators start with an intuitive appraisal, followed by a cognitive appraisal, which may

then ultimately be converged to achieve a sufficiently confident annotation judgment. We found intuition to play a major role in large parts of this process. Overall, we encourage researchers and practitioners to embrace intuition as a core part of expert annotators' knowledge, which may augment expert medical image annotation processes or ultimately be fed into ML models.

## References


Akinci, C., & Sadler-Smith, E. (2012). Intuition in Management Research: A Historical Review. *International Journal of Management Reviews, 14*(1), 104–122.

Bastick, T. (1982). *Intuition: How we think and act*. Wiley.

Baylor, A. L. (2001). A U-shaped model for the development of intuition by level of expertise. *New Ideas in Psychology, 19(*3), 237–244.

Campbell, L., & Angeli, E. (2019). Embodied Healthcare Intuition: A Taxonomy of Sensory Cues Used by Healthcare Providers. *Rhetoric of Health and Medicine*, 2(4), 353–383.

Chang, C.-M., Yang, X., & Igarashi, T. (2022). An Empirical Study on the Effect of Quick and Careful Labeling Styles in Image Annotation. *Graphics Interface, Montreal, Canada*

Chilcote, D. R. (2017). Intuition: A Concept Analysis. *Nursing Forum*, 52(1), 62–67.

Crandall, B., Klein, G. A., & Hoffman, R. R. (2006). *Working minds: A practitioner's guide to cognitive task analysis*. MIT Press.

Dane, E., & Pratt, M. G. (2009). Conceptualizing and Measuring Intuition: A Review of Recent Trends. In: *International Review of Industrial and Organizational Psychology* (1)1–40, Wiley.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382.

Freeman, B., Hammel, N., Phene, S.,… & Sayres, R. (2021). Iterative Quality Control Strategies for Expert Medical Image Labeling. *AAAI Conference on Human Computation and Crowdsourcing*, Nov 14-18, 2021, online

Furnham, A., Chan, P. S., & Wilson, E. (2013). What to wear? The influence of attire on the perceived professionalism of dentists and lawyers. *Journal of Applied Social Psychology, 43*(9), 1838–1850.

Golley, S., Corsini, N., Topping, D., Morell, M., & Mohr, P. (2015). Motivations for avoiding wheat consumption in Australia: Results from a population survey. *Public Health Nutrition, 18*(3), 490–499.

Hammond, K. R. (1996). Human Judgement and Social Policy: Irreducible Uncertainty, Inevitable Error, Unavoidable Injustice. *Oxford University Press USA*.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. In *Advances in Psychology* (52), 139–183. Elsevier.

Hensman, A., & Sadler-Smith, E. (2011). Intuitive decision making in banking and finance. *European Management Journal*, 29(1), 51–66.

Kahneman, D. (2011). *Thinking, fast and slow* (First edition). Penguin Books Ltd.

Kahneman, D., & Klein, G. (2009). Conditions for intuitive expertise: A failure to disagree. *American Psychologist, 64*(6), 515–526.

Klein, G. (2008). Naturalistic decision making. *Human factors*, 50(3), 456–460.

Klein, G. A., Calderwood, R., & Clinton-Cirocco, A. (1986). Rapid Decision Making on the Fire Ground. *Proceedings of the Human Factors Society Annual Meeting,30*(6), 576–580.

Litjens, G., Kooi, T., Bejnordi, B. E., …, & Sánchez, C. I. (2017). A survey on deep learning in medical image analysis. *Medical Image Analysis, 42*, 60–88.

Merriam-Webster. (n.d.). Intuition. In Merriam-Webster.com dictionary. Retrieved June 13, 2022, from https://www.merriam-webster.com/dictionary/intuition

Myers, M. D. (2020). *Qualitative research in business and management* (Third edition). SAGE.

Nederhof, A. J. (1985). Methods of coping with social desirability bias: A review. *European Journal of Social Psychology, 15*(3), 263–280.

Norris, P., Pacini, R., & Epstein, S. (1998). The rational-experiential inventory, short form. *Unpublished inventory. University of Massachusetts at Amherst.*

Okoli, J. O., Watt, J., & Weller, G. (2022). A naturalistic decision-making approach to managing non-routine fire incidents: Evidence from expert firefighters. *Journal of Risk Research, 25*(2), 198–217.

Ørting, S. N., Doyle, A., Van Hilten, A., … & Cheplygina, V. (2020). A Survey of Crowdsourcing in Medical Image Analysis. *Human Computation, 7* (1), 1–26.

Pandl, K. D., Feiland, F., Thiebes, S., & Sunyaev, A. (2021). Trustworthy machine learning for health care. *Proceedings of the Conference on Health, Inference, and Learning.*

Pustejovsky, J., & Stubbs, A. (2013). Natural language annotation for machine learning. *O'Reilly Media.*

Rädsch, T., Eckhardt, S., Leiser, … & Sunyaev, A. (2021). What Your Radiologist Might be Missing: Using Machine Learning to Identify Mislabeled Instances of X-ray Images, *54th Hawaii International Conference on System Sciences*, Jan 5-8, 2021, virtual

Simon, H. A. (1992). What is an "Explanation" of Behavior? *Psychological Science*, 3(3), 150–161.

von Rueden, L., Mayer, S., Beckh, K., … & Schuecker, J. (2021). Informed Machine Learning—A Taxonomy and Survey of Integrating Prior Knowledge into Learning Systems. *IEEE Transactions on Knowledge and Data Engineering*, early access

Wagner, M., Bihlmaier, A., Kenngott, H. G., … & Müller-Stich, B. P. (2021). A learning robot for cognitive camera control in minimally invasive surgery. *Surgical Endoscopy, 35*(9), 5365–5374.

Wang, S., Yu, L., Li, K., ..., & Heng, P.-A. (2020). DoFE: Domain-Oriented Feature Embedding for Generalizable Fundus Image Segmentation on Unseen Datasets. *IEEE Transactions on Medical Imaging, 39*(12), 4237–4248.

Ward, T. M., Fer, D. M., Ban, Y., …, & Hashimoto, D. A. (2021). Challenges in surgical video annotation. *Computer Assisted Surgery, 26*(1), 58–68.

Warsinsky, S., Schmidt-Kraepelin, M., Thiebes, S., Wagner, M., & Sunyaev, A. (2022). Gamified Expert Annotation Systems: Meta-Requirements and Tentative Design. In*: 17th International Conference on Design Science Research in Information Systems and Technology*, Jun 1-3, St. Petersburg, FL, USA