Human-AI teaming for Conformity Assessment of Welded Joints: A Human Factors Perspective

Marleen J. Hof
Faculty of Behavioural, Management and Social Sciences, University of Twente
Master Thesis

First supervisor: Prof. W.B. Verwey
Second supervisor: Dr S. Borsci
External Supervisor DEKRA: M. Robers
External Supervisor Intergo: M.P. Zeilstra

Contents

Abstract

AI systems to assist weld inspectors are on the rise. However, research on how to integrate the AI and create successful human-AI teams is lacking. In this research, a theoretical framework for human-AI teams was created to determine factors that play a role in human-AI interaction. Subsequently, a study was conducted to test the effect of two independent variables (Task division and the AI's accuracy) and one predictor variable (Propensity to trust) on four outcome variables: Accuracy, efficiency, fear-induced stress, and the feeling of having a meaningful job (FoMJ).

In the study, 18 weld inspectors of the company DEKRA were asked to interpret x-ray images with the AI. The study had a within-subject design. An extension to the study was conducted with 6 different weld inspectors to determine a consensus. This consensus was used to determine the expected correct answers for each x-ray. This was used to determine the accuracy in the main study.

The parallel task division led to more accurate results. The sequential task division was more efficient, although large individual differences were present. There was no effect of the type of task division on fear-induced stress or FoMJ. Propensity to trust positively influenced FoMJ with history-based trust as a mediator.

Since high accuracy is extremely important in weld inspection, the parallel task division should be favoured. The large individual differences in efficiency and the effect of propensity to trust on FoMJ illustrate that individual characteristics are important for successful implementation of an AI.

Keywords: Human-AI, weld assessment, conformity assessment, artificial intelligence, human performance, AI performance

Executive Summary

Artificial intelligence (AI) to assist weld inspectors has been brought to the market. The company developing the AI promises high task quality and increased task efficiency when the AI is integrated. Unfortunately, these claims are not as easy as it seems. Good AI's have the *potential* to be successful, but other variables play a role as well: How are you going to integrate the AI? How are your employees going to react to it? Will they trust the AI or will they reject and ignore the output? In this research a human factors perspective was taken to answer such questions. The focus was on four outcome criteria to ensure sustainable performance: efficiency, accuracy, and human well-being (split into fear-induced stress and the feeling of having a meaningful job).

A theoretical framework for human-AI teams was developed. It is a generic framework that can be used for other human-AI teams outside weld inspection as well. The framework includes factors that impact the outcome criteria. It includes factors that are directly related to the task and human-AI interaction (e.g., trust, mental workload). More distant factors are integrated as well. For example: how is the technology introduced by the company?

A study was conducted to test the effect of three variables of the framework: the task division between the human and AI, the AI's performance in indicating relevant regions of interest, and the *propensity to trust* (what is someone's general attitude towards AI systems?). Possible mediators between the input and output variables were explored as well.

The task division influenced both efficiency and accuracy. In the parallel task division, the human first sees an x-ray without the AI's interpretation. After interpreting the x-ray himself, the human receives the AI's interpretation, compares both interpretations, and makes a final decision. In the sequential task division the human immediately sees the x-ray with the AI's interpretation and makes a final interpretation with this x-ray. The parallel task division led to more accurate results. The sequential task division was more efficient, although large individual differences in efficiency were found. Due to the importance of accuracy in weld inspection, the parallel task division is favoured with the current AI. However, if the AI is improved it might be necessary to again determine which of the two task divisions is favoured.

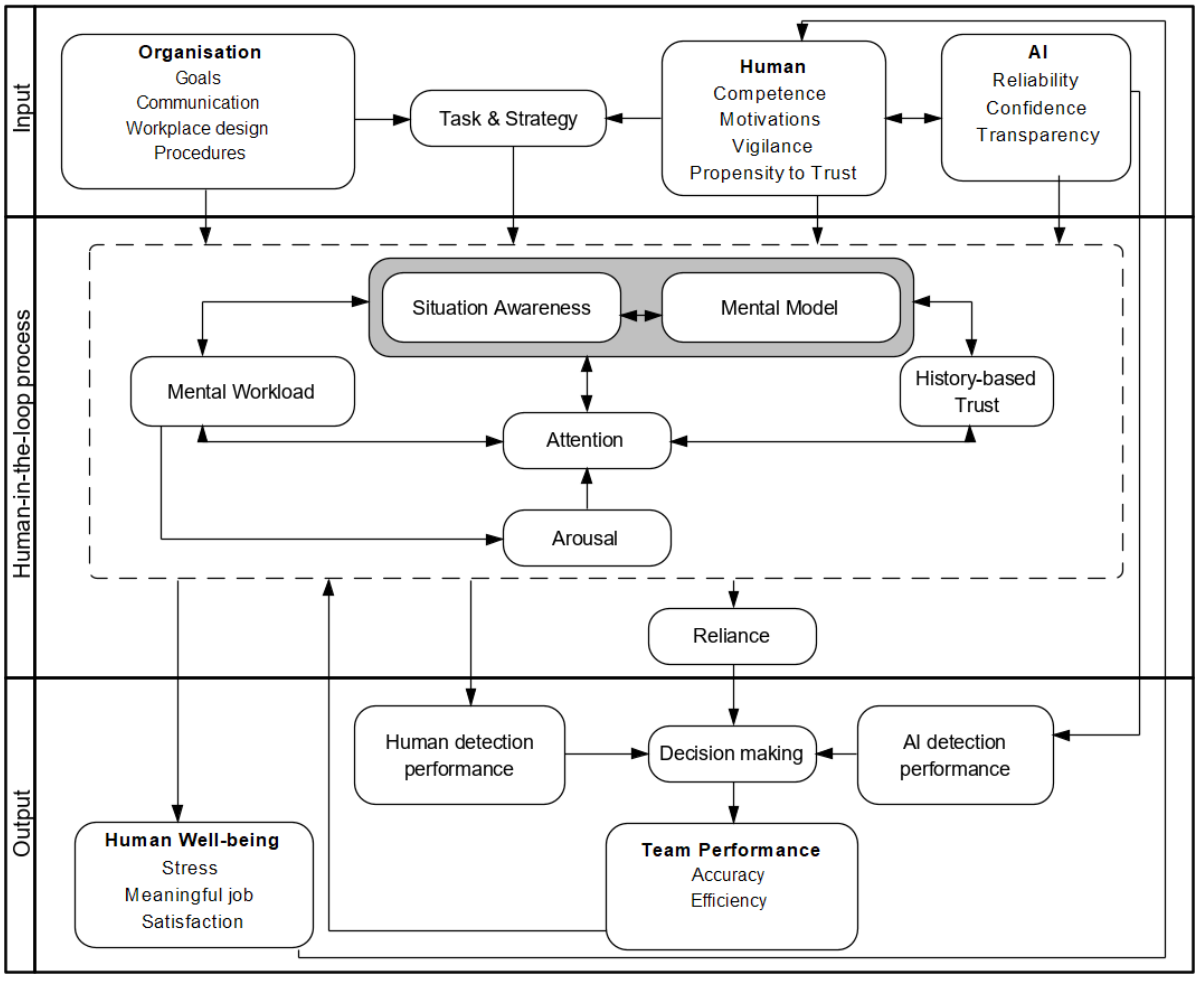
The type of task division did not have an effect on any of the two well-being criteria. However, the individual characteristic 'propensity to trust' influenced whether the participants felt if they had a meaningful job when the AI was integrated. If the inspectors do not feel like they are contributing to the team results, their motivation for the task might decrease. This might result in less input from the human inspector. As the human inspector is still of value and has different strengths as the AI, this will result in lower team performance.

Another insight of this research is the importance of individual characteristics and individual differences. This is illustrated both by the effect of propensity to trust and the individual differences in efficiency. These differences in efficiency are likely not only due to the AI. In the debrief participants mentioned that they in general worked faster than some of their colleagues. If the goal is to increase efficiency, this will not be met by just implementing an AI system. The elements underneath 'individual characteristics' in the theoretical framework might provide an explanation for these differences and could be investigated in future research.

This research was a starting point and focused on a few aspects of the theoretical framework only. Nonetheless, it illustrates that it is important to consider how one will integrate an AI system and how it interacts with and influences the human team member. To conclude, only focusing on the technical aspects is not sufficient. Even the best performing AI can become a failure if insufficient attention is devoted to human factors.

Figure 2 [copy from page 10 of the report]

Generic Framework for Human-AI Interaction to Determine Well-Being and Team Performance



1. Introduction

People encounter many situations in which they rely on structures that contain welds, including the bridges they drive on, or the railways of the train they travel with. To ensure safety in these structures, conformity assessments are conducted. In these assessments, it is checked whether there are discontinuities in the welds that could lead to safety hazards (Bertovic, 2016).

Currently, conformity assessments are executed by human inspectors. However, artificial intelligence (AI) has been developed to assist the human inspectors (TrueFlaw, 2020a, 2020b). Since the human and AI need to work together, it is important to consider the factors that determine successful teamwork. For a successful implementation it is important to achieve high team performance and to guarantee good well-being of the human inspector.

The main goal of this thesis is to explore the human factors issues that play a role within human-AI teams and how these influence performance and well-being in the context of weld inspection. An empirical study was conducted to investigate the impact of the AI's accuracy, different task divisions between the human and AI, and propensity to trust on four outcome variables. These outcome variables were efficiency, accuracy, fear-induced stress, and the feeling of a meaningful job. Moreover, the role of trust, mental workload, and reliance as mediators and interaction variables was explored. The AI developed by TrueFlaw was used in this study (TrueFlaw, 2020a, 2020b).

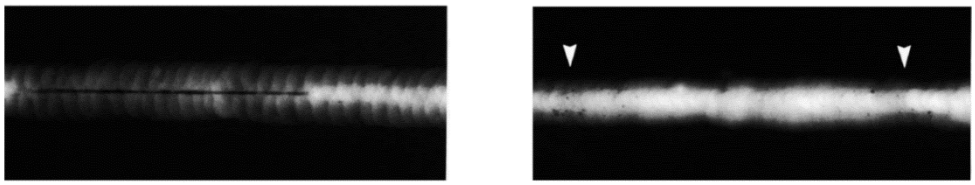
In this section (1. Introduction), the current state of weld inspection and the rise of AI within this context will be discussed, a human-AI team framework will be proposed and two possible task divisions between the AI and human are described and discussed. The section ends with the hypotheses for the empirical study.

1.1. Non-Destructive Testing

An often-used technique for weld inspection is radiographic testing. X-ray images of welds are created and these images are inspected for possible flaws (DEKRA, n.d.-a, Prakash, 2012). Radiographic testing is a widely used technique and one of the few techniques for which AI has been developed (TrueFlaw, 2020a). In Figure 1, example x-ray images and possible flaws are shown. The light grey areas show the welds and the darker spots on these welds indicate possible flaws.

Figure 1

Examples of X-ray Images of Welds



Note. The light grey areas are the welds and the dark spots within the welds are indications of flaws (GE Inspection Technologies, 2006). Both images contain different types of flaws: the left image has lack of penetration; the right has scattered porosity.

1.1.1. Explanation of Procedures

There are two types of radiographic testing: conventional and digital testing (DEKRA, n.d.-a, n.d.-b). With conventional testing the x-ray images are presented as photographic films whereas for the digital testing the images are shown digitally. As the AI is developed to interpret digital x-ray images, the focus in this thesis will be on digital testing.

Weld inspection consists of multiple steps (Bertovic, 2016; DEKRA, 2016). First, the inspector needs to prepare by checking the quality of the x-ray image. Second, the inspector inspects the weld in terms of discontinuity by determining the type of flaw and determining whether the discontinuity is critical. A discontinuity is critical when it is detrimental to the purpose of the (Ali et al., 2012). The inspector uses regulations, instructions, his expertise about welds, and the context to determine whether a flaw is critical. Third, the inspector needs to document their findings. This thesis will focus on the second step: the inspector has to detect critical flaws together with an AI.

1.1.2. Challenges of the Task

Weld inspection is a difficult task for multiple reasons. First, deciding whether a flaw is critical requires expert knowledge and the application of many regulations (Pereira & De Melo, 2020). Second, weld inspection is a monitoring task requiring high vigilance. Many images need to be assessed each day and many of these do not contain critical flaws (Hou et al., 2020; Warm et al., 1996). Unfortunately, vigilance decreases fast for humans, resulting in a decreased performance (Teichner, 1974; Woods & Hollnagel, 2006). Third, the task induces high pressure for the inspector to find all the flaws. Missing critical flaws might have catastrophic effects because it decreases the strength of the overall structure (Bertovic, 2016). These challenges should be considered when integrating AI into this context.

1.2. Artificial Intelligence in Weld Inspection

The most general definition for AI is, a system that makes intelligent decisions (Asan et al., 2020). AI systems can learn and create their own criteria to make a decision. The AI system developed by TrueFlaw (2020a) does this by using machine learning. The AI learns to distinguish welding flaws by looking at example x-rays on which flaws are labelled. Based on these input images, the AI learns how to recognise flaws.

Although the developers claim that adding AI will lead to more accurate and efficient weld inspection, research for this is lacking (TrueFlaw, 2020a, 2020b). A false assumption that is often made is that adding AI to a task will solely lead to benefits and release the burden on the human (Bansal et al., 2019a; Bansal et al., 2019b; Woods & Hollnagel, 2006; Zhou & Chen, 2019). However, AI is likely to come with new challenges and tasks which may result in an increased burden for the human and impact team performance. For example, a new challenge is that the human needs to know when to rely on the AI and when not. A new task is that the human has to perceive, interpret, and decide on the AI's advice.

To conclude, merely good AI performance and human performance on the primary task does not necessarily result in good team performance. Many other factors should be considered to ensure a positive outcome.

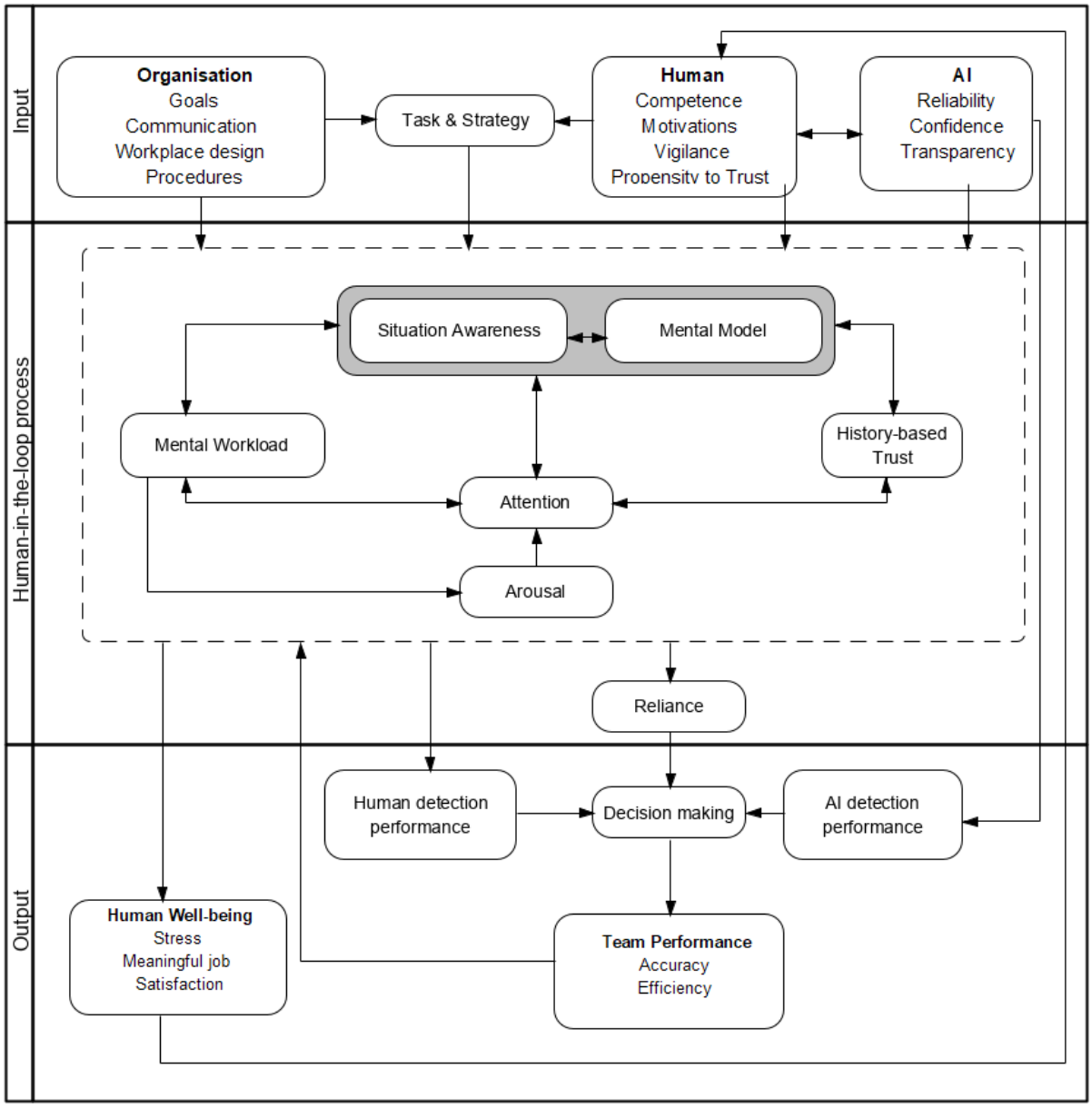
1.3. Theoretical Framework for Human-AI teams

A theoretical framework was developed that shows the important factors to create successful human-AI teams (Figure 2). The framework is inspired by a heuristic human-automation framework developed by Zeilstra (2020). In this thesis a literature review was done to confirm and adjust the framework to human-AI teams. The framework is divided into three parts: input, human-in-the-loop, and output factors. This section highlights the framework's elements that are useful for predicting a preferable task division between human and AI. The focus is on four output factors: accuracy, efficiency, fear-induced stress, and the feeling of having a meaningful job (FoMJ) (for an in-depth description of the framework see Appendix A).

Team performance. Team performance can be divided into efficiency and accuracy. When AI guides the human's attention towards potential flaws, efficiency could increase. A requirement for this interaction is that the human needs to rely on the AI. If the human does not rely on the AI, they would still inspect the whole weld on their own.

Figure 2

Generic Framework for Human-AI Interaction to Determine Well-Being and Team Performance



Accuracy is the proportion of cases that is correctly classified. Accuracy is based on the hits, misses, false alarms, and correct rejections of the team (Wickens, 2002). These four categories can be defined as follows: A flaw that is present is detected as such (*hit*), a flaw that is present is not detected (*miss*), a weld without a flaw is detected as such (*correct rejection*), and a weld without a critical flaw is detected as having a flaw (*false alarm*). The higher the hits and correct rejections, and the lower the false alarms and the misses, the higher the accuracy. Two terms related to accuracy are sensitivity and specificity (Wickens, 2002). Sensitivity indicates how many of the flaws that are present, are classified as such. The higher the hits and

the lower the misses, the higher the sensitivity. Specificity indicates how many of the cases in which there are no critical flaws, are correctly interpreted as such. The higher the correct rejections and the lower the false alarms, the higher the specificity.

The team's accuracy is determined by the complementarity of the AI's and human's individual performances. The human and AI have different strengths and limitations and thus can complement one another (Woods & Hollnagel, 2006). Whereas the AI's accuracy is solely determined by the AI's performance, the human's performance is determined by many more factors besides their skill to interpret x-rays (e.g., mental workload, attention).

Reliance. Even though the team performance depends on the complementarity of the individual performances, the team performance will not exceed individual performance levels if there is no appropriate reliance on the AI system (Bansal et al., 2019b). If the reliance is too high, the human might *misuse* the system by accepting all the AI's indications without properly assessing the x-ray themselves (Bansal et al., 2019b; Lee & See, 2004). In contrast, if reliance is too low the human might *disuse* the system and ignore the AI's suggestions. When reliance is optimal the human takes the AI's advice in scenarios where the AI's capabilities exceed the human's capabilities and follows their own advice when their own capabilities exceed the AI's capabilities.

Trust. Trust is a factor that has a large influence on reliance (Lee & See, 2004). Because of this, trust levels should be appropriate. Too high trust will lead to misuse whereas too low trust will lead to disuse. Trust can be divided into two types: history-based trust and propensity to trust. Propensity to trust is a trait-like tendency to trust or not trust AI systems (Merritt & Ilgen, 2008). History-based trust is the trust one has in a specific AI system after gaining experience with this specific system (Merritt & Ilgen, 2008). History-based trust is known to influence the reliance on the AI (Hoff & Bashir, 2015; Lee & See, 2004). In turn, history-based trust is expected to be influenced by propensity to trust, depending on the experience one has with the system (Jessup et al., 2019; Lee & See, 2004). After using the AI for a while, the impact of experience on history-based trust increases whereas the influence of propensity to trust decreases (Lee & See, 2004). However, the effect of propensity to trust on history-based trust remains a point of discussion (Jessup et al., 2019).

To sum up, to translate the individual performance of both the AI and the human it is important to have appropriate reliance and appropriate trust levels within the AI. If this is not the case the individual performance of the human and AI might not translate to high team performance.

Stress levels. In weld inspection it is important to find all critical flaws to ensure safety. The thought of missing flaws is expected to promote fear-induced stress (e.g., Hayashi et al., 2009). If the human inspector has the impression that the AI can be trusted and is a compatible teammate, the addition of AI can reduce fear-induced stress. However, the AI can also increase stress when trust in the AI is low and the AI is high in autonomy and lacks transparency (Endsley, 1995; Morris et al., 2017; Vereschak et al., 2021). In sum, one needs to consider compatibility, trust, the AI's autonomy, and the AI's transparency to ensure low levels of fear-induced stress.

The feeling of having a meaningful job. Implementing the AI brings a very important other risk: the fear of being replaced or being redundant (Evans et al., 2020). This feeling of redundancy can be the result of different factors. If relative trust in the AI is high (Lee & See, 2004), the feeling of redundancy might occur. This implies that the human trusts the AI's capabilities more than their own. Additionally, it is expected that the feeling of redundancy is larger when the human and AI share the exact same task.

1.4. Task Divisions Between Human and AI

As stated above, AI can be beneficial for weld inspection if implemented correctly. In this study, two task divisions proposed by TrueFlaw (2020a, 2020b) will be compared: the sequential and parallel task division. Whereas the sequential task division is designed for efficiency, the parallel task division is designed for accuracy. The question is to what extent the focus on accuracy/efficiency sacrifices the other performance characteristic. Furthermore, the effect of the two task divisions on fear-induced stress and the FoMJ is considered.

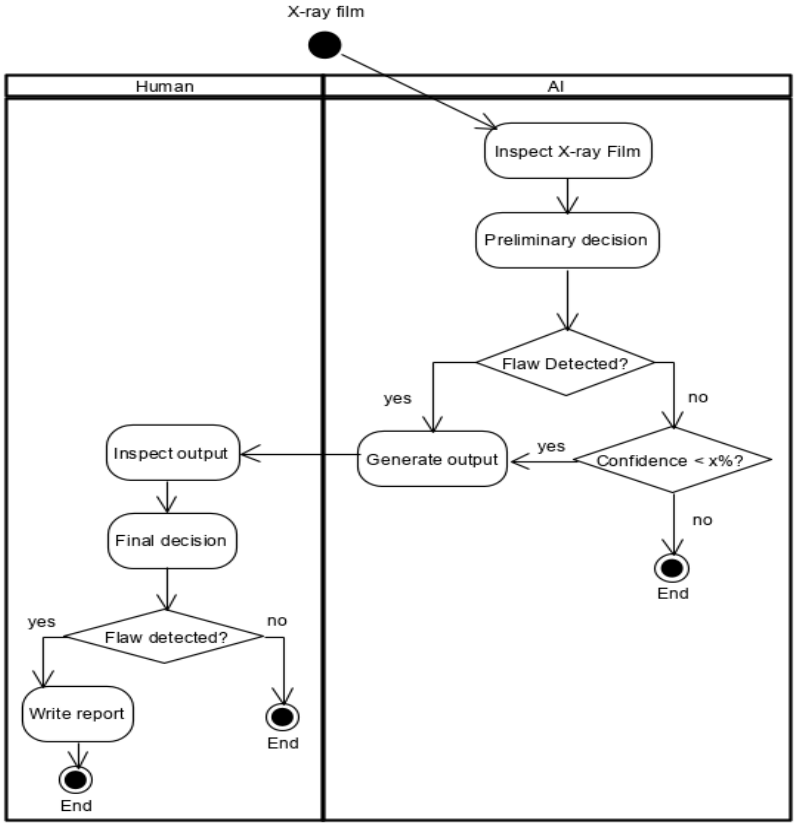
Below, a short description of each task description is given, followed by the expected strength and weaknesses of the task divisions.

1.4.1. Sequential Task Division

In the sequential task division, the AI examines the x-ray images first (Figure 3). If a flaw is found, the AI marks the flaws and the x-ray image is sent to the human inspector for further inspection. The human will make the final decision in this scenario. This task division also has a so-called *filter function:* if the AI detects no flaws in an x-ray image, the image is filtered out and not sent to the human. This filter function increases the AI's autonomy.

Figure 3

Sequential Task Division



Note. First, the AI filters out welds it is highly confident about being non-critical. For the remaining images, the AI marks the flaws and send them to the human inspector.

Strengths. TrueFlaw (2020a, 2020b) favours the sequential task division due to its high efficiency. Another benefit of this task division could be the different tasks the human and AI have. The AI selects potential interesting x-rays with indications, the inspector then decides whether this 'interesting' x-ray contains a critical flaw. This difference could result in high FoMJ.

Weaknesses. Besides these potential strengths, this approach has its own risks as well. The complementarity between the human and AI is not optimal. This can be detrimental for the team's accuracy. The reason for the low complementarity is twofold. The human does not have the opportunity to correct the AI's misses that are filtered out. Moreover, the AI's decisions might bias the human's decision by guiding their attention only to the areas highlighted by the AI (Endsley, 1995; Lee & See, 2004). The human might pay most attention to the areas

highlighted by the AI, resulting in the misses of the AI not being corrected. This high reliance is especially likely to occur when history-based trust in the AI is high (Lee & See, 2004).

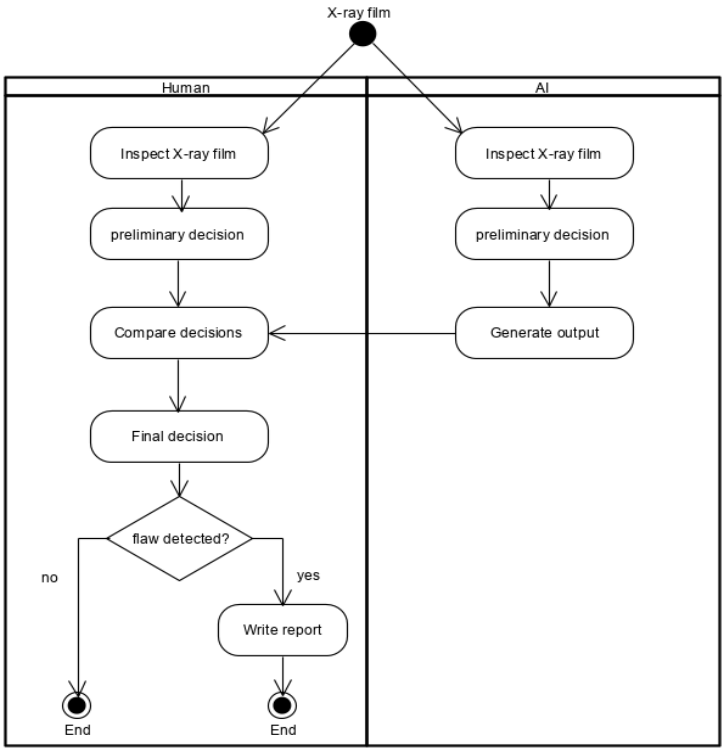
Other disadvantages of this task division are the AI's lack of transparency and high autonomy. Especially if trust in the AI is low, it is likely that the human will fear missing critical flaws, resulting in fear-induced stress (Hayashi et al., 2009). This fear-induced stress can only be avoided if the AI's accuracy is high and the trust in the AI is appropriate.

1.4.2. Parallel Task Division

For the parallel task division, the human and AI inspect the x-ray simultaneously but independent from each other (Figure 4). Both make an individual preliminary assessment. Only when both the human and the AI have made a preliminary assessment, the human inspector will be able to see the AI's assessment. To communicate the AI's assessment to the human, the AI generates an output that highlights areas where the AI found a potential flaw, including non-critical flaws. The human inspector will compare his own and the AI's assessment and make the final decision on whether or where critical flaws are present.

Figure 4

Parallel Task Division



Note. First, the human assesses the x-ray film independently. Subsequently, the human compares his and the AI's assessment to make the final decision.

Strengths. The strengths include complementarity, unbiased decision making, and low fear-induced stress. As the human and AI interpret all the x-ray images individually, the human and AI's capabilities can be combined, resulting in high complementarity. Moreover, the human makes a first unbiased decision without being guided by the AI's markings. Lastly, fear-induced stress is likely to be low as the human has the final say for all the x-rays.

Weaknesses. This approach has its downsides as well. The task of the human and AI are partly overlapping, risking low FoMJ and reduced effort from the human inspector. If the human inspector reduces effort the complementarity decreases (de Waard, 1996). Both the risk of low FoMJ and reduced effort are expected to be higher if trust in the AI is high (Lee & See, 2004). Another drawback from this task division is the low efficiency because the human needs to inspect each x-ray twice.

1.5. The Present Research

The goal of the current research was to test the impact of task division (parallel vs sequential), AI accuracy (high vs low), and propensity to trust on the outcome variables. The outcome variables were accuracy, efficiency, FoMJ, and fear-induced stress.

An empirical study was conducted to test the relation between the predictor and outcome variables. Expert weld inspectors worked together with an AI to identify critical flaws in welds. The welds were presented using x-ray images. It was also investigated how the relations between the input and output variables could be explained by human-in-the-loop factors (history-based trust, mental workload, and reliance).

Additional data collection was conducted to determine the correct interpretations for the x-ray images that were presented. Within weld inspection it is difficult to establish the correct interpretation, because there exists a large variety among inspectors' interpretations. For this reason, a consensus was created to develop an estimated truth. Different inspectors than those that worked with the AI had to inspect x-rays. They inspected the same x-ray images as the participants that worked with the AI, only without being assisted by the AI. The answers of these inspectors were used to determine a consensus: if the majority of them indicated an area as a flaw, it was classified as a flaw. This consensus could be used to determine the accuracy of the interpretation of those that had worked with the AI. As the consensus is an estimation and not the absolute truth, terms such as accuracy, hits, misses, etc. will be proceeded by '*subjective*' when discussing the results.

1.5.1. Hypotheses

For each of the four outcome variables (accuracy, efficiency, fear-induced stress, and the FoMJ), separate hypotheses were set up. These hypotheses were tested separately for the

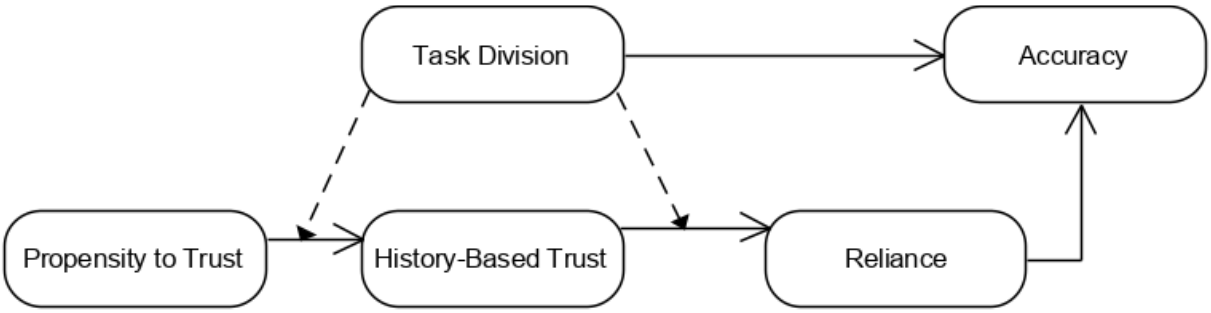
AI with high and low accuracy. This was done to explore whether the AI's accuracy influenced the expected relations.

Accuracy. As the parallel task division is expected to have a higher complementarity than the sequential task division, it was hypothesised that *team accuracy would be higher for the parallel task division than for the sequential task division* (H1). Accuracy was measured by comparing the team decision to the consensus.

The relationship between task division and team accuracy was expected to be partly mediated by reliance (Figure 5). Reliance on the AI was expected to impact complementarity, and thus team accuracy. As explained above, the reliance was expected to be higher in the sequential task division because the human's decision is more likely to be influenced by the AI's decision (Endsley, 1995). This bias is especially likely when the trust is high (Lee & See, 2004). In other words, it was expected that *the final decision the human makes relies more heavily on the AI's decision in the sequential task division than in the parallel task division, especially when history-based trust in the AI is high* (H1.1). Reliance was measured by using an agreement percentage: the higher the agreement between the AI and the human, the higher the reliance (Zhang et al., 2020). History-based trust was measured with a questionnaire.

Figure 5

Expected Model for the Team's Accuracy



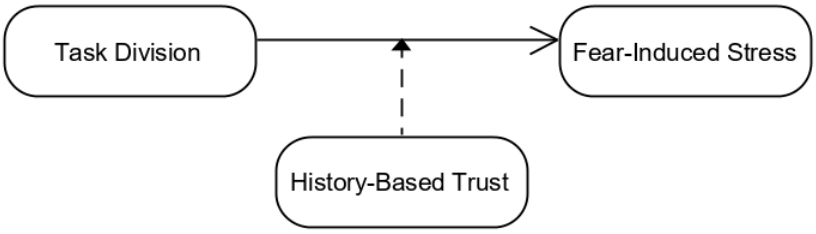
Note. The arrow with a dotted line indicates an interaction. The open arrow heads indicate a direct effect.

A last addition to Hypothesis 1 is the influence of the Propensity to Trust on History-Based Trust (Jessup et al., 2019; Lee & See, 2004). History-Based Trust is expected to be influenced by an interaction between Task Division and Propensity to Trust. Since calibrating trust in the AI might be more difficult for the sequential task division, propensity to trust may have a larger impact (Lee & See, 2004). Thus, it was expected that *history-based trust is influenced by propensity to trust more in the sequential task division than in the parallel task division* (H1.2). Propensity to trust was measured using a questionnaire.

Fear-Induced Stress. An interaction between history-based trust and task division on fear-induced stress was expected. Whereas the human sees all x-rays in the parallel task division, they do not have the possibility to check all the scenarios in the sequential task division. For this reason, history-based trust is expected to play a more important role in the sequential task division. It was expected that *a decrease in history-based trust increases fear-induced stress more in the sequential than in the parallel task division* (H2, see Figure 6). Fear-induced stress was measured with the physiological measure of heart rate.

Figure 6

Hypothesized Interaction Between Task Division and History-Based Trust on Stress



Note. The arrow with a dotted line and the closed arrowhead indicates an interaction. The open arrowhead indicates a direct effect.

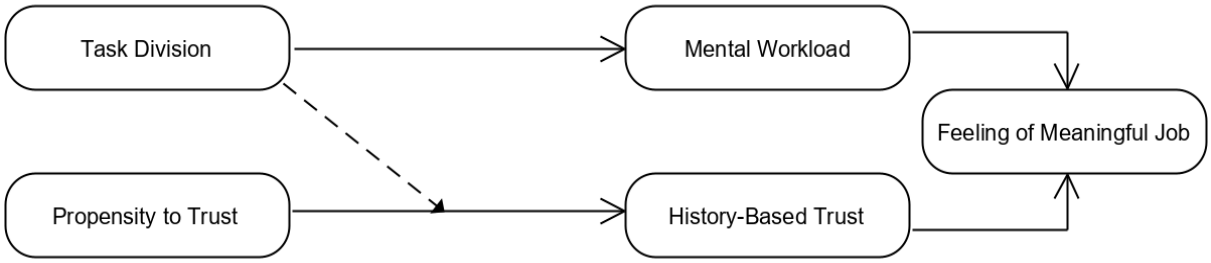
Feeling of Meaningful Job. As the human's and AI's task overlap more in the parallel task division than in the sequential task division, it was expected that *FoMJ is lower in the parallel task division than in the sequential task division* (H3). It was expected that the relationship between the type of task division and FoMJ is mediated by mental workload. Underload is a risk for weld inspection, because it is a monotonous task (Young & Stanton, 2002). If underload occurs, this can induce feelings of redundancy on the human inspector. In the parallel task division, the human might rely on the AI to conduct the part of the task that is the same for the human and the AI. Based on the above, it was expected that *the relationship between task division and FoMJ is mediated by mental workload in such a way that the parallel task division leads to underload, which decreases the FoMJ* (H3.1.). Mental workload was measured with heart rate variability (HRV) and a questionnaire. FoMJ was measured with a self-made questionnaire.

History-based trust is expected to play a role in determining FoMJ as well. Too high levels of history-based trust might induce the feeling of redundancy (Evans et al., 2020). Thus, it was expected that *high history-based trust causes a lower FoMJ* (H3.2). The interaction

between task division and propensity to trust on history-based trust was also taken into account for this framework (Figure 7).

Figure 7

Factors Influencing the Feeling of Having a Meaningful Job



Note. The arrow with a dotted line indicates an interaction. The open arrow heads indicate a direct effect.

Efficiency. As the human inspector needs to look at each x-ray twice for the parallel task division, *a higher efficiency was expected for the sequential than for the parallel task division* (H4). Efficiency was measured by the number of welds assessed within 30 minutes.

2. Methods

Two separate methods are described in this section. The first is for the main study. The goal of this study was to gather data to test the hypotheses. The second method section is for an extension called the consensus study. The goal of this study was to gather data to derive a consensus that could be used to determine the accuracy of the participants in the main study. The participants in the main study differed from those that took part in the extension study.

2.1. Main Study

2.1.1 Participants

18 weld inspectors from DEKRA took part in the study. The participants' age ranged from 31 to 66 years (M = 43.33, SD = 11.77). The mean experience within the field was 18 years (SD = 10.42). The study had an online and on-location version. Eight weld inspectors from Germany participated in the on-location version and ten weld inspectors from South-Africa participated in the online version. Although the two versions are slightly different, the two studies are treated as one. In the analyses of the data, the version was taken into account when possible. The research was approved by the BMS ethics committee and all participants gave written consent prior participating.

2.1.2. Design

The study involved a 2 (task division: parallel vs sequential) x 2 (AI accuracy: high vs low) within-subject design. This resulted in four possible combinations and thus four

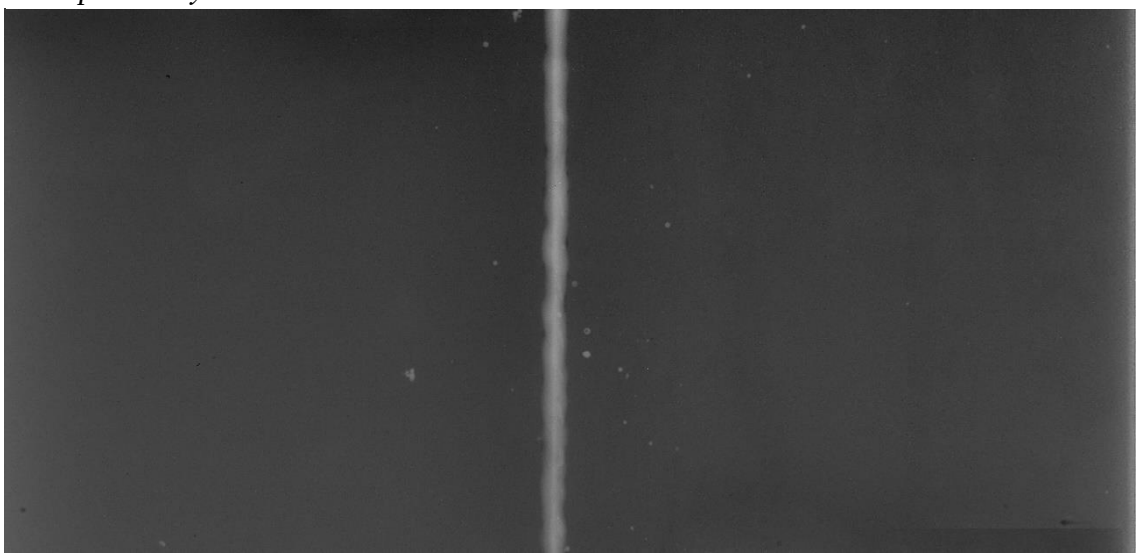
conditions: parallel-low, parallel-high, sequential-low, and sequential-high. For each of the conditions, the participants completed one run. The order of the runs was counterbalanced for AI accuracy. The participants first completed two rounds with one of the AI's accuracy levels and then two rounds with the other AI's accuracy level. The order of the task divisions was randomised. Due to the addition of propensity to trust as an additional predictor variable, the study was a quasi-experiment.

2.1.3. Materials

A sample of 900 x-rays displaying welds were used. The x-rays were made by DEKRA for their client Meyer Shipyards Turku (https://www.meyerturku.fi/). Meyer Shipyards Turku gave permission to use the x-rays for this thesis. TrueFlaw (Trueflaw.com) processed the x-rays. They created the images with the AI's indications and made four folders with x-ray images: 150 for sequential-low, 150 for sequential-high, 300 for parallel-low, and 300 for parallel-high. The same x-ray was never shown in more than one of the conditions. The images in the sets parallel-low and parallel-high consisted of 150 x-rays without the AI's assessment (Figure 8) followed by the same x-rays with the AI's indications (Figure 9). The AI's assessment consisted of white rounded rectangles to indicate possible flaws. For sequential-low and sequential-high, 150 x-rays with the AI's assessment were present per condition (Figure 9).

Figure 8

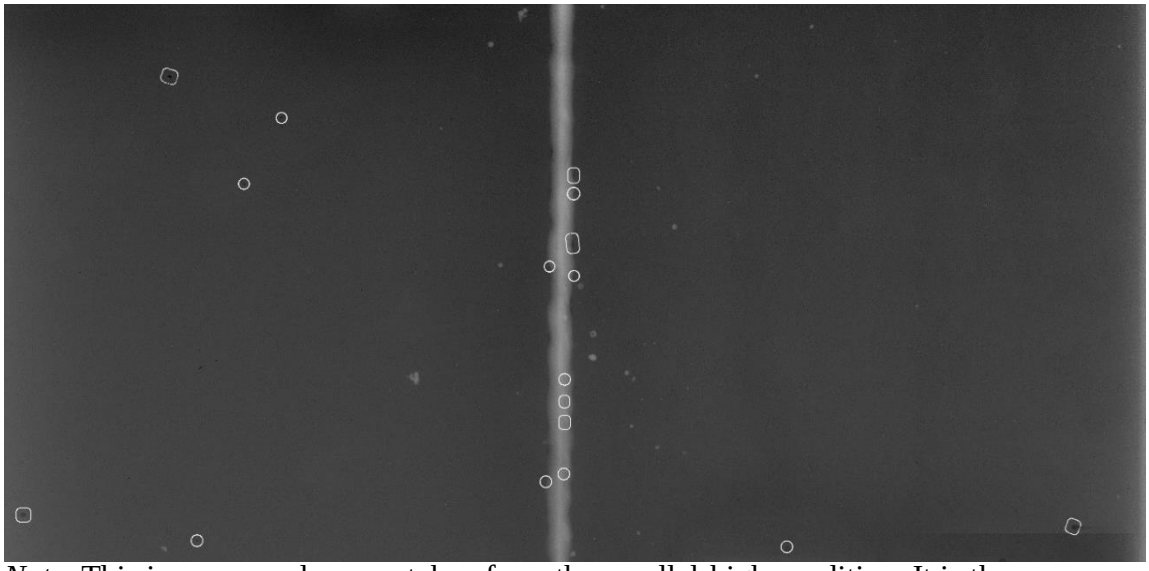
Example X-Ray Without the AI's Indications



Note. This is an image of a digital x-ray as normally encountered by an inspector. It is an image used in the parallel-high condition. The light grey area in the middle is the weld that needs to be assessed.

Figure 9

Example X-Ray With the AI's Indications



Note. This is an example x-ray taken from the parallel-high condition. It is the same x-ray as in Figure 8, but now includes the AI's markings. The AI's markings can be recognised as white outlinings around – what the AI interprets as – interesting areas. Interesting areas include all the flaws, also the non-critical ones.

Trueflaw used different methods to create the AI's indications for high and low accuracy. For high accuracy, the default output of TrueFlaw's AI was used. To create images for the two conditions with the low AI accuracy, the default output of the AI was compromised by taking out some of the AI's indications (removing hits, increasing misses) and adding random indications (increasing false alarms).

The Image Annotator Tool was used by the participants to assess the welds and mark the critical flaws (https://www.robots.ox.ac.uk/~vgg/software/via/via-1.0.6.html). With this tool, the participants made rectangular markings on the x-ray images to highlight critical flaws. Their interpretations were saved as coordinates for the highlighted areas. In the on-location version, the x-rays were presented on a 21-inch HP monitor. The screen size in the online version was variable, as the participants used their own equipment.

The web-based software Qualtrics was used to provide participants with detailed information, ask for consent, and to conduct questionnaires. The information included a general information sheet about the aim and topic of the study, a detailed explanation about the different types of task divisions, and a short explanation of the task division to be used in the upcoming run.

A semi-structured interview set-up was used to conduct a debrief. It included the topics trust, mental workload, stress, and FoMJ. The audio of the debrief was recorded using the ASR voice recording application for Android.

Additional on-location equipment for Germany. Additional equipment that was used in the on-location version were the Pupil Invisible eye-tracking glasses (https://pupil-labs.com/products/invisible/) with the corresponding Invisible Companion app and the H10 heart rate sensor Polar and the HRV application.

Additional online equipment. Certain materials were used for the online version only. There was a Google Drive folder with all the x-ray images in four different folders representing the four conditions. Microsoft Teams was used to communicate with one another and to exchange necessary links (Qualtrics and the Image Annotator Tool) and additional documents (the participants' assessed x-ray images).

2.1.4. Measures

Accuracy. Subjective accuracy was measured by comparing the participants' final decisions with the consensus. This consensus was used to determine the subjective accuracy in terms of subjective hits, misses, false alarms, and correct rejection. The subjective accuracy was analysed on pixel level, meaning that each pixel of each x-ray was labelled as a subjective hit, miss, false alarm, or correct rejection (Wickens, 2002). How the consensus was derived is explained in more detail in 2.2. Study Extension: Consensus Study. Besides the subjective accuracy, the subjective sensitivity and specificity of the team's final decision were determined as well.

Reliance. Reliance was measured with the so-called agreement percentage. This is the percentage of trials in which the participant's final decision is in line with the AI's decision (Zhang et al., 2020). A higher agreement percentage should indicate a higher reliance. The underlying assumption is that in all conditions the human and AI should have the same agreement percentage given that the AI remains stable and the x-rays are randomized. If the agreement percentage is higher in one of the conditions than in the other conditions, this indicates that the human tends to rely more on the AI's decision in this condition compared to the other condition. Similar to accuracy, the agreement percentage was analysed on pixel level. For each pixel it was determined whether the human's and AI's assessments were similar.

To see whether the higher agreement percentage was caused by visually relying on the AI's indications an eye-tracker was used. Unfortunately, data from the eye-tracker could not be used due to malfunctions with the equipment.

Heart rate measures: Mental workload and stress. The H10 heart rate sensor Polar and the HRV application were used for heart rate (variability) measures. Stress was measured with mean heart rate. The higher the mean heart rate, the higher the experienced stress. Additionally, measures of heart rate variability – the pNN50 and RMSSD – were used to measure mental workload. The pNN50 is "the percentage of successive RR intervals [time between beats in ms] that differ by more than 50 ms" (Shaffer & Ginsberg, p. 2). The RMSSD is "the root mean square of successive RR interval differences" (Shaffer & Ginsberg, p. 2). The higher the heart rate variability measures, the lower the mental workload.

Questionnaires: Meaningful job, trust, and mental workload. FoMJ was measured with a self-made questionnaire (Appendix B). The questionnaire consisted of 11 statements that were rated on a 5-point Likert-scale. Items were created for three different categories: focused on personal development (e.g. 'I learn new things when working with the system'), meeting the employer's demands (e.g. 'As a team, we can meet the demands of the employer more efficient than either of us alone'), and the larger society (e.g. 'The reliability of welds will increase when the system and I work together').

The propensity to trust scale by Merritt (2011) was used to measure the propensity to trust. In this 6-item questionnaire the respondent must indicate to what extent they agree with a statement on a 5-point Likert scale (Merritt et al., 2013). An example item is 'I usually trust machines until there is a reason not to'.

For history-based trust the TOAST questionnaire was used (Wojton et al., 2020). The TOAST measures two factors related to trust: system understanding and system performance. An example item for system understanding is 'I understand what the system should do'. An example for system performance is 'the system performs the way it should'. The scale consists of 9 items which need to be answered with a 7-point Likert scale ranging from strongly disagree to strongly agree.

The rating scale mental effort (RSME) was used to measure workload (Zijlstra & Van Doorn, 1985). The RSME has only one question: how much mental effort did it take you to complete the task? The answer sheet consists of one line with nine anchor points. The anchor points are accompanied by descriptive labels. It ranges from 0 (no effort) to 150 (highest effort possible).

All the questionnaires were translated to German and were available in German and English for the participants.

2.1.5. Procedure

Procedure on-location version. The participants were welcomed at the start. A German-English interpreter was present and introduced to them. The participants were given the option to talk in German to the translator, who in turn would translate their speech into English to the researcher. Moreover, the translator could translate the researcher's English speech to German if requested. After establishing whether the participant wanted to communicate via the translator, the participants had to read and fill in the first part of Qualtrics. This consisted of the information sheet, informed consent, demographic questions, and the propensity to trust scale by Merritt (2011). The participants were told that if they had any questions about the information sheet or informed consent, they could ask the researcher.

Subsequently, the participants read the second part on Qualtrics. This contained the detailed explanation of the different task divisions (parallel and sequential) and AI accuracy levels (high and low). Participants were told that it was extremely important that they understood the differences and that they should ask questions when parts were unclear.

After the detailed explanation, the researcher demonstrated how to use the Image Annotator tool. Moreover, to ensure the participants understood the two task divisions, the researcher demonstrated in the Image Annotator tool how to perform the task in the parallel and sequential task division. Hereafter, the participants were shown how to put on the Polar band and the eye-tracker off-set was conducted.

Subsequently, the four experimental runs of 30 minutes were conducted, one for each condition. The order of the experimental runs was counterbalanced between participants. Each run started with a short explanation in Qualtrics about the task division and AI accuracy level for the upcoming run. Hereafter, the researcher started the heart rate monitor and eye-tracker and the participant started the weld inspection. For each of the x-ray images the participant identified whether they thought a critical flaw was present or not. For the conditions with the sequential task division, they directly saw the x-ray with the AI's indications. They used this x-ray to come to their one and final decision. For the conditions with the parallel task division, they had to do two interpretations: they first did their own interpretation independent from the AI and after this first interpretation they had to interpret the same x-ray again with markings from the AI. This second decision was the final team decision in the parallel task division. After each round of 30 minutes, the researcher stopped the participant and wrote the number of interpreted images down. Subsequently, the participants filled in the TOAST, RSME, and self-made FoMJ questionnaire. After each round, the participants had a break of 5 minutes before

the next round started. After all four rounds, a short debrief of approximately 15 minutes was conducted. The total study duration was approximately four hours.

Procedure online version. The online version was similar to the on-location version, although a few aspects were different. The participants could and did not wear the eye-tracker and heart rate equipment. The explanation of the Image Annotated Tool was more detailed as the participants now needed to upload and save the x-ray images themselves. During the assessment, uploading and saving of the x-rays the participant had to share their screen, although most participants chose to share their screen for the whole session. A link to the Google Drive folders with the input x-rays was distributed to the participants via an email in which they were requested to download these folders prior the start of the study. Lastly, there was no interpreter available for the online study because the participants were native English speakers.

2.2. Study Extension: Consensus Study

2.2.1. Participants

Six participants took part in the consensus study. The participants were weld inspectors from DEKRA Duisburg that did not participate in the main study described above. The participants' age ranged from 38 to 60 years old (M = 51.50, SD = 8.02). The mean experience within weld inspection was 20.17 years (SD = 12.24). The study was on location in Germany. The research was approved by the BMS ethics committee and all participants gave written consent prior to participating.

2.2.2. Materials

The sample of x-ray images of the main study for the sequential-low and sequential-high conditions were used. However, this time the x-rays of the welds were shown without the AI's indications (see Figure 8 for an example). The Image Annotator Tool was used to assess and indicate the critical flaws (https://www.robots.ox.ac.uk/~vgg/software/via/via-1.0.6.html). The x-rays were presented on a 21-inch HP monitor.

2.2.3. Procedure

The participants were welcomed at the start. Three participants had to perform the weld inspection task simultaneously. The German-English translator was introduced to them, and they were given the option to talk via the translator. Subsequently, they read the information sheet and signed the informed consent. They filled in a sheet with demographic questions. Some verbal explanations were provided to make sure that the participants understood their task. After this, the researcher demonstrated the use of the Image Annotator tool.

Two runs of 30 minutes were conducted during which the participants had to perform the weld inspection task. For each of the x-ray images they identified whether they thought a critical flaw was present or not. The participants were stopped after 30 minutes by the researcher and the researcher wrote the number of interpreted images down. Between the two rounds, the participants had a break of 10 minutes before starting the second round. The total duration of the study was around 1.5 hours.

3. Results

First the consensus will be described. The choices that are made to derive at the consensus are given. An example visualisation of the consensus for one x-ray is given as well. Hereafter, the results of the main study in relation to the hypothesis will be provided. This is split up in quantitative analysis, qualitative analysis, and triangulation.

3.1 Deriving the Consensus

The participants' interpretations from the extension study were used to derive a consensus for the sequential-low and sequential-high condition. Data from the main study was used to derive a consensus for the parallel-low and parallel-high conditions. More specifically, in the parallel task division, the participants first interpreted the x-ray without assistance from the AI. These first interpretations in the main study were used to determine a consensus for the conditions with the parallel task division.

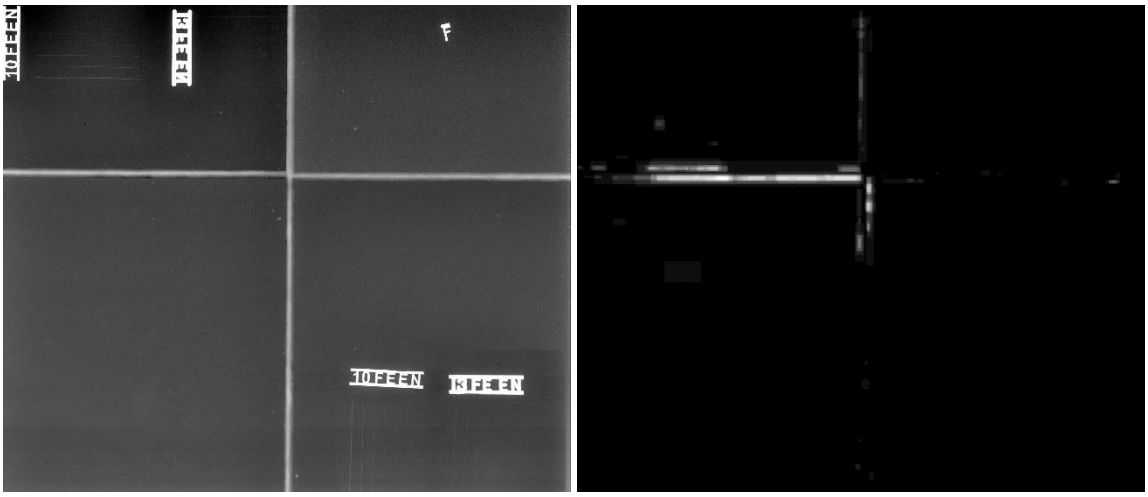
For an image to be used in the consensus, at least four of the participants should have assessed the image. All images before the last picture assessed by the participant were considered as assessed images[1]. An area was indicated as a critical flaw if at least 50 percent of those that interpreted the image marked the area.

The consensus was generated as a pixel matrix and for each pixel it was described how often this was marked by a human inspector. These matrices could be used to compare the consensus to the final team decision, the AI's decision, or the human's first interpretation in the parallel task division. An example of a visualization of the consensus can be seen in Figure 10. The lighter an area, the larger the number of inspectors that marked the pixel. Hence, black areas are not marked, grey areas are marked by some, and white areas are marked by the largest number of inspectors.

[1] Some of the x-rays were of too low quality to be interpreted by the human eye and participants were told they could skip these x-rays. This should be considered when interpreting the accuracy data.

Figure 10

Example Image of the Consensus for One X-Ray



Note. The image on the left is an example of an x-ray that was presented to the participants. The image on the right displays the consensus. The brighter the area, the higher the number of weld inspectors that marked it. For black areas no weld inspector indicated a critical flaw. White indicates that all weld inspectors indicated a critical flaw in that area.

3.2. Quantitative Analyses

Quantitative analyses were conducted to test the hypotheses for accuracy, efficiency, fear-induced stress, and the feeling of meaningful job. The samples for the analyses were determined for each outcome variable separately. To illustrate, if heart rate measures were missing for one participant, but the data for accuracy was complete, the participant's data was included for accuracy but not for fear-induced stress. Taking out all participants that missed data in one or some of the measures would result in an unusable small sample size.

Moreover, all the analyses were conducted both with and without outliers. Taking out the outliers did not give different results with the exception of one analysis. The results discussed below are including outliers. If the sample without outliers gave different results, this is noted. Lastly, all analyses were conducted separately for the conditions with high and low AI accuracy. These analyses were done separately as this part of the research is exploratory to investigate whether the AI's accuracy influences the expected relations.

3.2.1. Subjective Accuracy

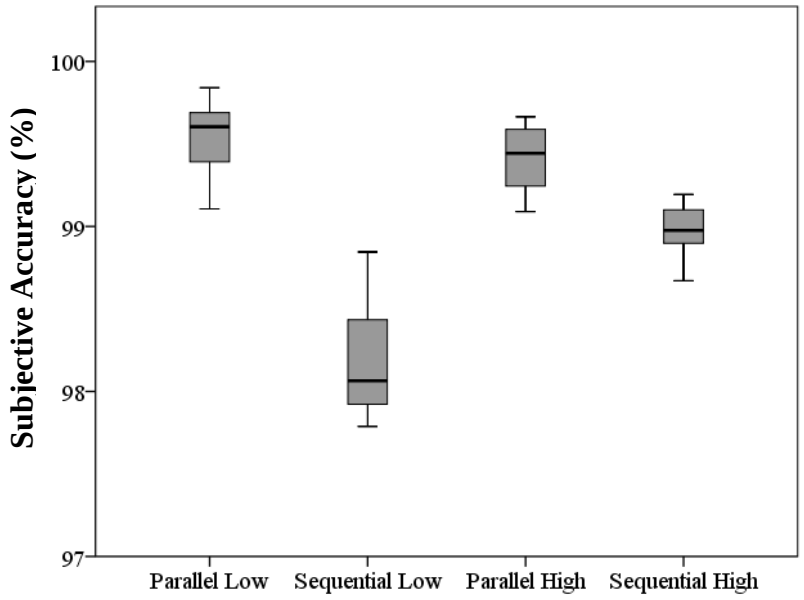
Data from the main study's online and on-location version were used for the analysis. A sample size of 16 was used, because data was missing for two participants. It was not possible to test the whole model as described in Figure 5. Using a linear mixed model gave validity issues that could not be solved. Generalized linear models could not be used as the assumptions

for these models were not met. See Appendix C for a more in-depth explanation of these issues. For these reasons, only the effect of task division on subjective accuracy was tested using the non-parametric Friedman's test.

The Friedman's test's results indicated that the subjective accuracy was higher in the parallel task division than in the sequential task division, both for high AI accuracy ($\chi^2(1, n = 32) = 12.25, p = .001$) and low AI accuracy ($\chi^2(1, n = 32) = 16.00, p < .001$). For an overview of these results see Figure 11.

Figure 11

The Subjective Accuracy (%) for Each of the Conditions



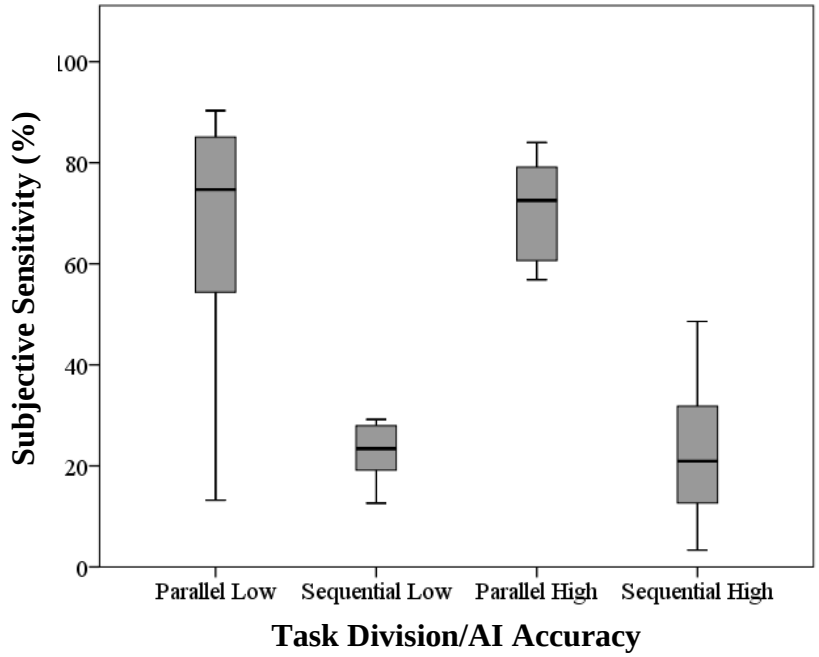
Note. A higher subjective accuracy is found for the parallel task division than for the sequential task division, regardless of the AI's Accuracy.

As accuracy is determined by sensitivity and specificity, the effect of task division for those separate constructs was tested as well. In general, the higher the sensitivity and/or specificity, the higher the accuracy. A higher subjective sensitivity in the parallel task division than in the sequential task division was found (Figure 12). This relation was found for high AI accuracy ($\chi^2(1, n = 32) = 12.25, p = .001$) and low AI accuracy ($\chi^2(1, n = 32) = 16.00, p = .004$). For specificity, a higher subjective specificity was found in the sequential task division than in the parallel task division for the high AI accuracy ($\chi^2(1, n = 32) = 16.00, p < .001$) and low AI accuracy ($\chi^2(1, n = 32) = 6.25, p = .021$). Although, the difference is significant, the absolute difference in specificity between the parallel and sequential is less than 0.5 percent, for both low and high AI accuracy. For the sample without the outliers ($n = 13$), the effect of task

division on subjective specificity was found only for the high AI accuracy ($\chi^2(1, n = 26) = 13.00, p < .001$).

Figure 12

The Subjective Sensitivity (%) for Each of the Conditions



Note. The parallel task division resulted in a higher subjective sensitivity than the sequential task division, both for low and high AI accuracy. Outliers were taken out of the image to increase readability of the figure.

To conclude, the parallel task division resulted in more subjective accurate assessments than the sequential task division. This is in line with hypothesis 1. It was found that the higher subjective accuracy occurred due to a higher subjective sensitivity. Whether the relation between task division and accuracy is mediated by history-based trust and reliance could not be studied because no fitting statistical model was found to test this.

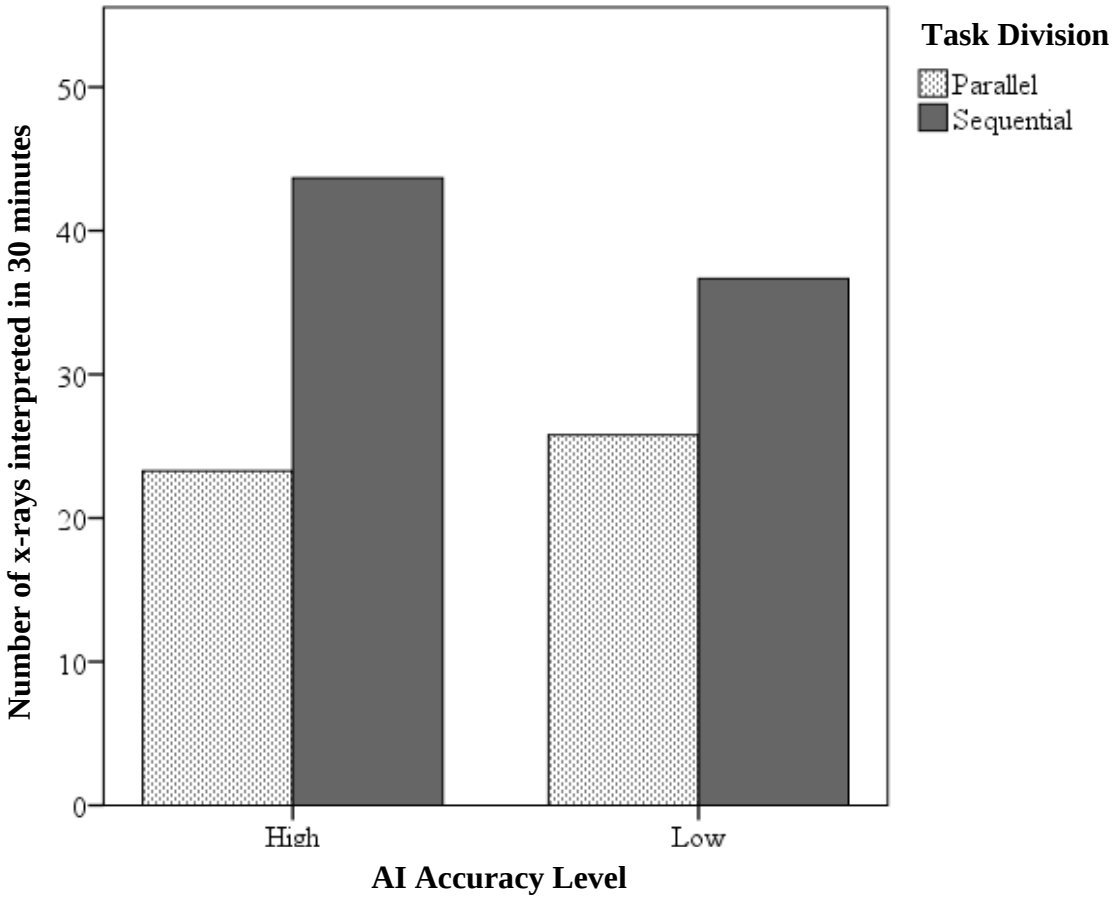
3.2.2. Efficiency

Data of all 18 participants was available for the analysis. Friedman's non-parametric test was used due to the violation of normality. The Friedman's non-parametric test showed that in the sequential task division more welds were interpreted in 30 minutes than in the parallel task division. This effect was found both for high AI accuracy ($F_r = 18.00$, $p < .001$) and low AI accuracy ($F_r = 8.00$, $p = .008$). For high AI accuracy, the mean rank for the parallel task division was exactly 1 and the mean rank for sequential task division was exactly 2. This indicates that every participant interpreted more images in the sequential task division than in

the parallel task division. For low AI accuracy, one participant interpreted more x-rays in the parallel task division than in the sequential task division. For an overview of the results see Figure 13.

Figure 13

Number of Images Interpreted in 30 Minutes for the Different Task Divisions, Categorised by AI Accuracy

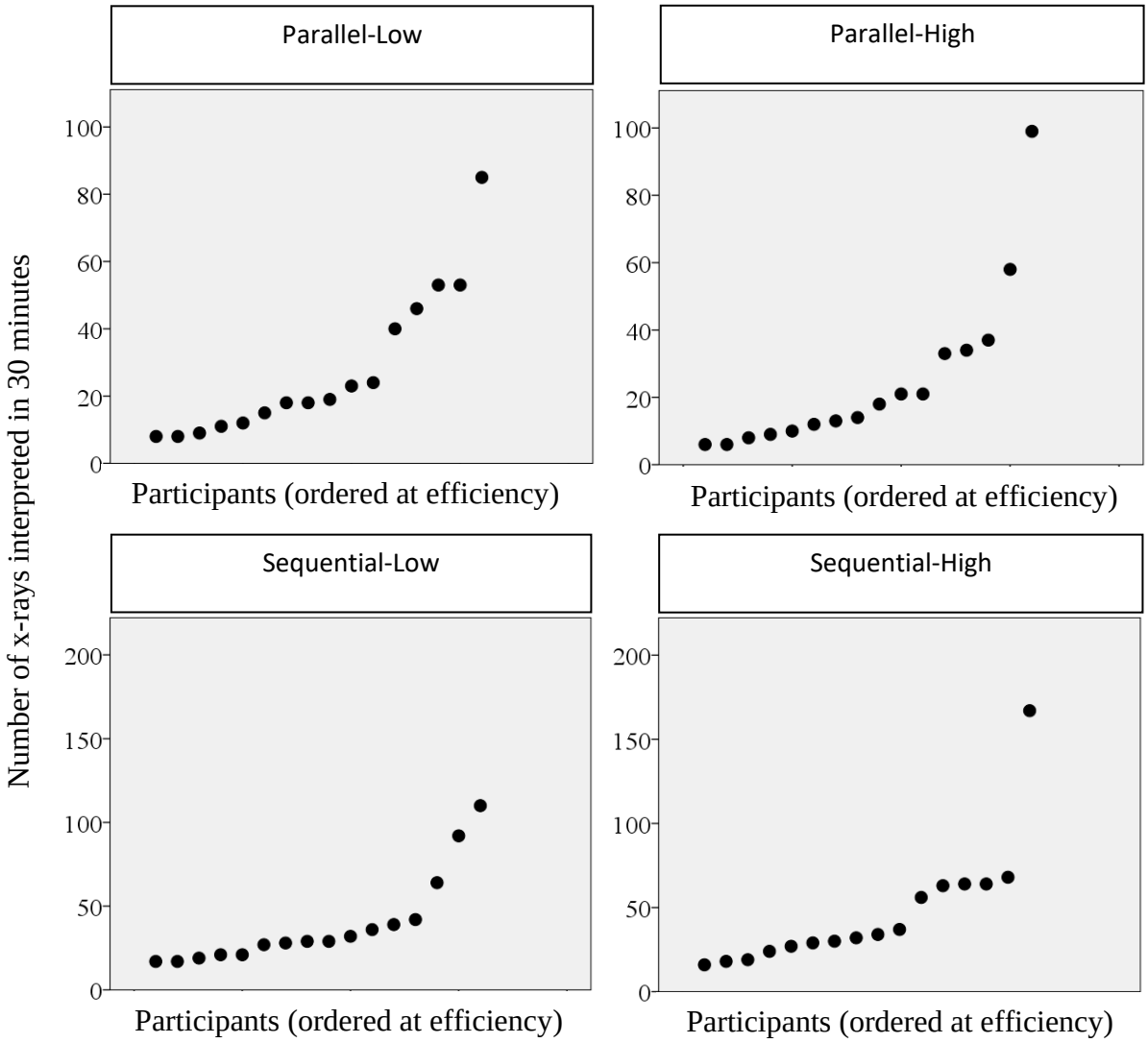


Note. The sequential task division was more efficient than the parallel task division.

Besides a difference in efficiency for the different task divisions, large individual differences were found within the conditions (Figure 14). Part of the participant group interpreted twice as many images as the other part. This large difference is of importance, as one of the main premises of the AI according to TrueFlaw (2020a, 2020b) was increased efficiency.

Figure 14

Individual Differences in Efficiency Within the Four Conditions

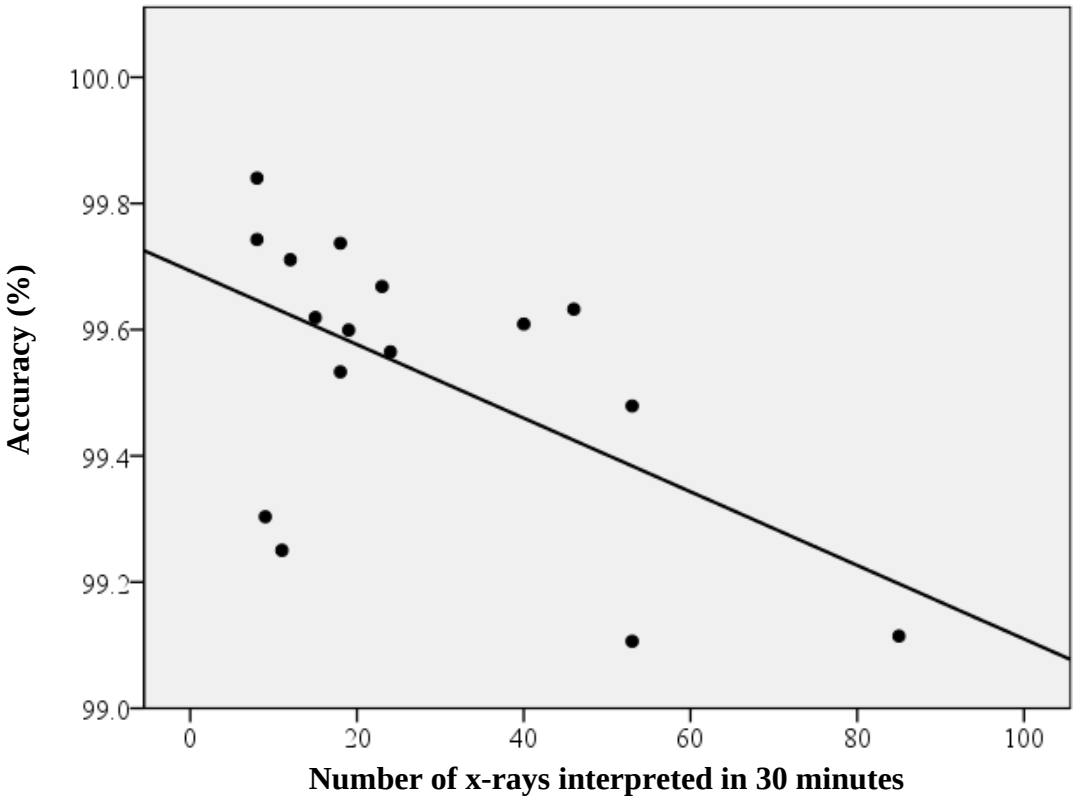


Note. Large individual differences in efficiency within each condition can be found. 'High' and 'Low' indicate the AI's accuracy. Note that the y-axis for the tables in the sequential task division go up till 200 whereas the tables displaying the efficiency for the parallel task division go up till 100.

The remaining question is why some participants worked more efficient than others. Spearman's rank correlation was computed to assess whether a relationship between accuracy and efficiency exists. In the parallel-low condition, there was a negative correlation between the two variables, $r(14) = -.51$, $p = .046$ (Figure 15). No correlations were present in the other three conditions. This indicates that other variables, such as individual factors, might play a role. These individual characteristics should be studied and addressed as well to increase efficiency.

Figure 15

The Negative Correlation Between Efficiency and Accuracy for the Parallel-Low Condition



Note. The slower the participant, the more accurate their interpretation. This correlation was found in the parallel-low condition only. For the other three conditions no correlation between accuracy and efficiency was found.

To conclude, the results indicate that the sequential task division is more efficient than the parallel task division, supporting hypothesis 4. However, the results also show that there are large individual differences regarding efficiency. For the parallel-low condition a negative correlation between efficiency and accuracy was found. For the other conditions this correlation was not found and it seems that other influences related to the individual might play a larger role in the efficiency with which the task is performed.

3.2.3. Effect of Task Division x History-Based Trust on Stress

An analysis was conducted to test whether the interaction Task Division x History-Based Trust on fear-induced stress was present. Only data from the on-location study could be used (n=8) in which data was missing for two of the participants. This resulted in a sample size of six.

A linear mixed model was tested with a Task Division x TOAST design on heart rate. The results showed no interaction between task division and TOAST on heart rate. Neither for

high AI accuracy ($t(3.95) = .606$, $p > .05$) or for low AI accuracy ($t(3.40) = 2.14$, $p > .05$). These results reject hypothesis 2 which states that a decrease in history-based trust increases fear-induced stress more in the sequential task division than in the parallel task division.

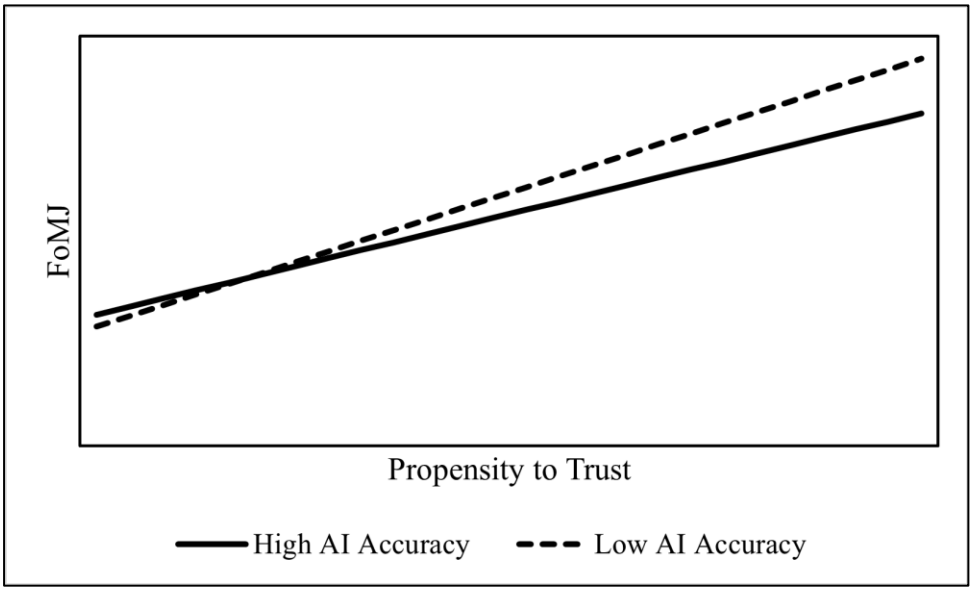
3.2.4. Feeling of Meaningful Job

Two analyses were conducted to test the effects of task division and propensity to trust on FoMJ. Data of all 18 participants was used for the analysis. A linear mixed model was conducted with propensity to trust, task division and Propensity to Trust x Task Division as predictors for FoMJ. Task division had no significant effect on FoMJ (high AI accuracy ($t(16.00) = .31$, $p > .05$), low AI accuracy ($t(16.00) = .05$, $p > .05$)). No interaction was found between task division and propensity to trust on FoMJ, neither for high AI accuracy ($t(16.00) = -.28$, $p > .05$), or low AI accuracy ($t(16) = -.13$, $p > .05$). As the expected relations between the predictors and FoMJ were not found, the expected mediators (H3.1 and H3.2) were not further investigated.

Interestingly, propensity to trust did have an effect on FoMJ (Figure 16), both for high AI accuracy ($b = 1.20$, $SE = .42$, $t(17.40) = 2.9$, $p = .010$) and low AI accuracy ($b = 1.26$, $SE = .33$, $t(22.36) = 3.87$, $p = .001$). The results indicate that a higher propensity to trust resulted in a higher FoMJ. This correlation has the opposite direction as expected.

Figure 16

The Effect of Propensity to Trust on FoMJ



Note. A positive effect of propensity to trust on FoMJ was found for both low and high AI accuracy. The direction of the effect is the opposite as was expected.

The possibility of history-based trust as mediator was analysed. As expected, a significant positive effect of propensity to trust on history-based trust was found (high AI accuracy: b =1.23, SE = .41, t(17.86) =2.94, p = .009; low AI accuracy: b = 1.09, SE = .48, t(20.27) = 2.25, p = .036). A significant negative correlation between history-based trust and FoMJ was found (High AI accuracy: b = .33, SE = .14, t(32.94) = 2.29, p = .029; Low AI accuracy: b = .30, SE = .11, t(26.63) = 2.72, p = .011). This correlation was the opposite of the expected direction. Additionally, the effect of propensity to trust on FoMJ remained significant when including history-based trust and propensity to trust as predictors (high AI accuracy: b = .84, SE = .38, t (16.53) =2.23, p = .040; low AI accuracy: b = .94, SE = .27, t (16.52) = 3.49, p = 0.003). Altogether, it indicates that history-based trust could be a partial mediator for the effect of propensity to trust on FoMJ. Propensity to trust positively influences history-based trust which in turn positively influences FoMJ.

To conclude, task division did not affect FoMJ, rejecting hypothesis 3. Interestingly, a direct effect was found of propensity to trust on FoMJ with history-based trust as a mediator. Propensity to trust positively influenced history-based trust, as expected. History-based trust in turn positively influenced FoMJ. This latter relationship between history-based trust and FoMJ is in the opposite direction as expected in hypothesis 3.2.

3.3. Qualitative Debrief Results

3.3.1. Task Divisions: Experienced Advantages and Disadvantages

The participants' preference in task division was inconsistent. 8 participants preferred the parallel task division, 8 preferred the sequential task division, and 2 participants were undecided (for an overview of the preferences and reasons for these preferences see Table 1).

The most often-mentioned advantage for the sequential task division was its efficiency. Other advantages included that the AI guides attention and that it requires less effort. One participant said it was advantageous that the sequential task division made them look harder for more flaws. One participant indicated that it was an advantage that they agreed more often with the system in the sequential task division. Disadvantages mentioned for this task division were that the AI's decision biased their inspection and that the AI's markings cover part of the weld and hinder inspection due to this.

For the parallel task division, the most often mentioned advantage was its similarity to the current task, because the AI functioned as a second opinion. Other mentioned advantages were increased team performance and a confidence boost when their own and AI's decision matched. The most often mentioned disadvantage was that it was time-consuming as each x-ray needed to be inspected twice. Lastly, two participants mentioned that the AI could create

uncertainties about their own capabilities if the initial interpretation of them and the AI did not match.

Table 1

Task Division: Preferences and Reasons for These Preferences

Sequential		Parallel	
Pros	Cons	Pros	cons
• Guides attention (4x) • Time-wise efficient (8x) • It makes you look harder for more flaws (1x) • High agreement with the system (1x)	• Markings cover part of the weld (3x) • Biased by AI's decision (2x)	• Double amount of work (4x) • High similarity to current practice (3x) • Higher accuracy (2x) • Confidence boost on own performance(1x)	• Double amount of work (4x) • Uncertainty about own capabilities (2x)

Note. The number between brackets indicates how many participants noted the reason

3.3.2. General Experience in Working With the AI

Comments on the general experience in working with the AI were divided into five categories: the AI's transparency, the AI's accuracy, trust in the AI, effort, and team accuracy.

Five participants commented on the AI's lack of transparency. They desired more information on the AI's workings and algorithms. The participants stated that the AI sometimes did not understand the AI's decision. One example is that the AI marked flaws on x-rays with poor picture quality. The participants wondered how the AI could interpret images that the human eye could not inspect. The participants noted that more transparency and information on the AI's workings could help them to understand how the AI comes to its decisions.

Most participants agreed that the AI's current level of accuracy was not sufficient yet. 14 participants thought the AI could still be improved further. This was because the AI was too sensitive (mentioned by seven), missed too many flaws (mentioned by seven), and/or was inconsistent with its markings (mentioned by three). Only one of the participants explicitly mentioned that the AI's current accuracy was sufficient. The participant noted that the AI consistently identified flaws, although it was quite sensitive in its evaluation. In conclusion, most participants agreed that the AI has great potential, but that it needs further development for the human to rely on it and for it to become a more valuable teammate.

Despite the low perceived accuracy and lack of transparency, most participants did state they trusted the AI as a teammate. Eleven participants indicated they did trust the AI and five

participants indicated that they did not trust the AI. Two participants were undecided on the trust in the AI due to the limited experience with the AI.

Although the majority stated they trusted the AI as a teammate, most participants did not trust the AI's filter function. Those with low trust in the AI (five participants) all indicated they would not trust the AI's filtering function. Those that were undecided on the trust in the AI due to the lack of experience with the AI (two participants) did not trust the filtering function. For the majority of the participants that said they trusted the AI, this trust had a prerequisite: nine participants mentioned they would trust the AI as a teammate, if they themselves could make the final decision. One of them said they trusted the AI to filter out images if he could check a sample of those x-rays that are filtered out. The last participant did not provide a clear comment on whether he trusted the filter function. Hence, in total 16 participants explicitly mentioned they wanted the final call and thus did not trust the AI to filter out images.

Half of the participants did not mention effort as a relevant factor or did not notice a difference in effort (9 participants). Merely one of the participants noted that adding AI to the work increased effort. The remaining six participants noted that the AI decreased the effort because it guided them to interesting areas. Interestingly, three of these participants mentioned that the effort decreased only for the sequential task division.

Most of the participants stated that working with the AI resulted in a better performance than working alone (11 participants). They said the AI assists them in correcting their misses when they are tired and that in general 'four-eyes' find more than two. Three participants thought the AI would not be beneficial for team performance. The remaining four participants were in doubt whether the AI would increase team performance. When comparing the two task divisions, most of the participants that mentioned this, said there was no difference in accuracy between the two task divisions (mentioned by seven). Two participants thought the parallel led to more accurate results. None of the participants thought the sequential task division led to more accurate results. Regarding efficiency, 6 participants explicitly mentioned this would be higher for the sequential task division. No participant experienced the parallel as more efficient. Lastly, participants mentioned in the debrief that there exist large individual differences in efficiency when doing weld inspection without the AI.

To conclude, the AI's transparency and accuracy must be improved for the participants to fully accept the AI. At the moment, most participants do trust the AI enough to be a teammate that provides suggestions and think that this will increase the overall performance for the weld inspection task. More transparency and a more accurate AI are needed for the participants to trust the AI's filter function as well.

3.4. Triangulation: Qualitative and Quantitative Data Combined

In this section the quantitative and qualitative results for the four outcome variables are compared: accuracy, efficiency, stress, FoMJ. For an overview see Table 2.

Similar results for efficiency were found for the qualitative and quantitative results. The most often mentioned advantage for the sequential task division was that it was more efficient than the parallel task division (6 times). The higher efficiency for the sequential task division was found in the quantitative analyses as well.

The quantitative and qualitative results differed in terms of accuracy. The debrief results indicate that there were no large differences in terms of accuracy when comparing the task divisions. However, the quantitative analyses indicated that the parallel task division led to more accurate results than the sequential task division (with the exception for the sample without outliers and for low AI accuracy).

Regarding the well-being measures, both FoMJ and stress were not affected by task division according to the quantitative analyses. Both aspects were not mentioned as interesting variables in the debrief. When asked about it, none of the participants noted a difference between the conditions. Thus, the quantitative and qualitative results were similar.

Table 2

Comparison Quantitative and Qualitative Results

	Debrief	Statistical analyses
Accuracy	-Higher for parallel (2*) -Higher for sequential (0) -No difference (7)	-Higher for parallel (exception sample (n = 13), low AI accuracy)
Efficiency	-Higher for sequential (6) -Higher for parallel (0)	-Sequential more efficient
Stress	-Not mentioned as an interesting variable	-No links found
Feeling of Meaningful Job	-Not mentioned as an interesting variable.	-No effect of task divisions -Mediation effect found from propensity to trust on history-based trust on Feeling of Meaningful Job

Note. The number indicates the number of participants that mentioned it.

4. Discussion

The purpose of this study was to gain a better understanding of human-AI teamwork in weld inspection. A quasi-experiment was conducted to compare the effect of two task divisions, two levels of AI accuracy, and propensity to trust on team performance (accuracy and efficiency) and well-being (fear-induced stress and the FoMJ). Although the parallel task division led to more accurate results, the sequential task division was more efficient. These results are in line with hypotheses 1 and 4. Task division did not affect fear-induced stress or FoMJ (rejecting H2 and H3 partly). Propensity to trust positively affected FoMJ.

4.1. Performance: Accuracy and Efficiency

4.1.1. Accuracy

The higher subjective accuracy in the parallel task division than in the sequential task division was in line with hypothesis 1. A higher accuracy in the parallel task division was expected because of a higher complementarity between the human and AI in this task division. With high complementarity the different strengths of the two team members can be combined (Woods & Hollnagel, 2006). This complementarity was confirmed in the debrief. The participants stated that the AI made misses and false alarms, but also found flaws they themselves initially missed.

In the sequential task division, complementarity was expected to be impaired for two reasons: due to a biased decision-making by the human and the filtering function of the AI. It was expected that the AI's decision biased the human's decision, if history-based trust is too high (Endsley, 1995; Lee & See, 2004). More specifically, it was expected that the human mostly focused on the AI's highlighted areas in the sequential task division. In this scenario, the misses of the AI would not be corrected. In line with these expectations, the debrief results showed that some participants thought their inspection was biased by the AI's results. Unfortunately, due to malfunctions with the eye-tracker it was not possible to confirm the participants' statements empirically.

The filter function of the sequential task division was expected to impair the complementarity of the human and AI's qualities. With the filter function, the human does not have the possibility to correct the AI's misses for the filtered-out x-rays. However, when TrueFlaw (trueflaw.com) provided the AI's data, it became clear that the AI was more sensitive than expected beforehand. The AI was oversensitive and highlighted at least one area on each x-ray, resulting in none of the x-rays being filtered out even though there were many x-rays on which no critical flaws were present. If the AI is further developed it might be possible that the AI filters out images that contain flaws. Future research is needed to study the effects this has.

The higher accuracy could be explained by a higher subjective sensitivity in the parallel task division than in the sequential task division. The higher sensitivity indicates that there were relatively more hits and less misses in the parallel task division than in the sequential task division (Wickens, 2002). This difference between the task divisions is important, because one of the main challenges of weld inspection is finding all critical flaws because missing critical flaws can have catastrophic consequences (Bertovic, 2016).

4.1.2. Efficiency

Efficiency was higher for the sequential task division than for the parallel task division. This is in line with the premise of TrueFlaw (2020a, 2020b) and hypothesis 4. As the human needs to look at each image twice in the parallel task division and only once in the sequential task division, this result is not surprising. If the filter function of the sequential task division will be added, the difference in efficiency between the two task divisions will increase further.

Although the results showed that the sequential task division was more efficient, the large individual differences in efficiency indicate that other factors besides task division play a role. Within the same condition, part of the participants interpreted twice as many images as other participants. For the parallel-low condition, efficiency was negatively correlated with accuracy. For the other conditions this correlation was not found. A possible explanation for this correlation in the parallel-low condition might be high reliance on the AI. This can explain why this difference is found in the parallel low- but not in the parallel high-condition: The team accuracy will decrease further when relying on the AI with low accuracy than on the AI with high accuracy. Future research is needed to investigate the possible role of reliance.

The correlation between accuracy and efficiency does not provide a full explanation as it was only found for the parallel-low condition. Plausible additional explanations might be differences in the individual's motivations and competence. For example, the assessors' efficiency in weld assessment in general or computer literacy might play a role. Differences in competence were mentioned in the debrief by a few participants. Future research is needed to investigate the role individual characteristics play for efficiency

4.1.3. Trade-of: Quality Versus Quantity

Whereas the parallel task division has higher accuracy and sensitivity, the sequential task division is more efficient. Accuracy, and especially sensitivity, goes above efficiency in weld assessment, because missing flaws can have catastrophic consequences (Bertovic, 2016). Pressure to become more efficient should never jeopardize the accuracy. Taking this into account, the parallel task division is favoured over the sequential task division in terms of performance. In addition, studying the individual human characteristics that possibly influence

efficiency, the parallel task division might become more efficient without jeopardizing accuracy.

4.2. Well-being: FoMJ and Fear-Induced Stress

4.2.1. Fear-Induced Stress

The results did not support the interaction between task division and history-based trust on fear induced-stress. It was expected that the filter function in the sequential task division would increase fear-induced stress, if history-based trust was low (Hayashi et al., 2009; Lee & See, 2004). The debrief results suggest that trust in the filter function was low, thus this cannot explain the absence of the relation between task division and fear-induced stress.

There are two plausible reasons that there was no difference in fear-induced stress for the two task divisions: lack of emphasis on the filter function and of real-life consequences. First, the filter function of the AI was not clearly visible for the participants. It was only explained shortly prior to the two runs with the sequential task division. During the run, it was not communicated to them if and how many images were filtered out. The lack of visibility of the filter function possibly impacted the participants' understanding of the filter function. During the debrief participants seemed confused by the meaning of the filter function. If the user is not aware of the filter function, it will not influence their stress levels. Second, it was clear to the participants that this was an experiment and thus there were no real consequences of missing critical flaws. This can also explain the absence of fear-induced stress.

4.2.2. Feeling of Meaningful Job

No effect of task division on FoMJ was found, but an effect of propensity to trust on FoMJ was found. The absence of a relation between task division and FoMJ might be because of the AI's performance. The reasoning for hypothesis 3 was that the human might feel redundant and feared to be replaced by the AI due to its high performance. However, during the empirical study it became clear that the AI's performance was lower than expected. Participants noted the AI made many misses and false alarms and said that the AI should be further improved. This lack of AI performance might have influenced the FoMJ. The human did not see the AI as a threat. The feeling of redundancy is not likely, because the participants perceived their own skills as higher than those of the AI (Evans et al., 2020; Lee & See, 2004). In sum, the low AI performance (for both levels of AI accuracy) might explain the absence of a relationship between task division and FoMJ.

Propensity to trust did positively influence FoMJ. This effect was mediated by history-based trust: Propensity to trust positively influenced history-based trust, which in turn positively influenced FoMJ.

The first part of this mediation – propensity to trust influencing history-based trust, adds to the discussion on whether propensity to trust is related to history-based trust (Jessup et al., 2019). Whereas some studies do find a relationship between the two types of trust, other studies do not (Jessup et al., 2019). The difference between the current study and the studies discussed by Jessup et al. is that the current study used an AI whereas the other studies used automation. Experience with the system (AI or automation) is needed to establish history-based trust (Lee & See, 2004). If the user does not have enough experience with the system, the trust in this particular system has to be based on something else than experience and understanding: the propensity to trust AI/automation systems in general (Lee & See, 2004). As AI systems are complex and difficult to understand (Zednik, 2019), more experience might be needed to understand these systems. More experience might be needed to decrease the effect of propensity to trust on history-based trust for AI systems than for automation (Asan et al., 2020; Aziz & Dowling, 2019).

The second part of the mediation – history-based trust positively influencing FoMJ – was in the opposite direction as expected in H3.2. As said before, too high levels of history-based trust and high AI accuracy might induce the feeling of being redundant and decrease the FoMJ (Evans et al., 2020). However, the participants experienced the AI's accuracy level as relatively low compared to their own. A possibility is that, instead of a linear relationship, there is an inverted U-shape curve between history-based trust and FoMJ. If history-based trust is too high, the human might feel redundant, decreasing the FoMJ (Evans et al., 2020). However, if history-based trust is too low, the human might experience the AI as a clumsy aid that does not help them to perform better or to personally grow from, decreasing the FoMJ. As the AI's performance was perceived as low in this study, this latter experience (AI as a clumsy aid) is more likely.

4.2.3. Well-Being: Implications for the Integration of AI

The study does provide any worrying results regarding stress levels or the FoMJ. For both aspects of well-being, no effect of the task division was found and this study does not provide a preference for one of the two task divisions. However, it was found that propensity to trust influences the FoMJ. This highlights the importance of considering individual variables when integrating an AI system. Moreover, the research adds to the discussion whether propensity to trust influences history-based trust. In this study it is argued that with complex systems such as AIs, propensity to trust might still have an influence on history-based trust if the experience is too limited to properly understand the specific AI system at work.

4.3. Pitfalls and Opportunities of the AI

Some general pitfalls and opportunities for using an AI system within weld inspection are of importance. The first pitfall is the AI's current performance. The AI's performance was perceived as too low by the participants and needs to be improved. TrueFlaw (trueflaw.com) has been working on improving the AI and a newer and better version than the one used in this study has already been released. This has complications for the results of these studies. For example, trust in the AI is likely to be different if the AI is more reliable (Lee & See, 2004). It might also influence the relation between history-based trust and FoMJ if the inverted U-shape graph is a correct assumption. It might be needed to test the effect of the improved AI performance on the hypotheses.

A second pitfall is the lack of transparency and understanding. The results show the need of transparency of the AI and understanding in the AI by the inspectors. If the inspector understands the AI system better, automation surprises are less likely to occur (Woods & Hollnagel, 2006). Understanding can be increased during the introduction of the system. For example, during the training someone with understanding of how the AI makes decisions, is trained, or is set-up should be present. This allows the participants to immediately ask questions while practicing with the AI system.

An opportunity of the AI is that it might decrease the monitoring aspect of the weld inspection task. It was mentioned by the participants that the AI helped them to stay attentive for a longer period, or that they noticed that the AI started to correct their misses as their attention decreased over time. These results are promising because the required vigilance and sustained attention are a challenge in weld inspection (Teichner, 1974; Woods & Hollnagel, 2006). However, these claims by the participants should still be studied empirically.

4.4. Limitations of the Current Study and Future Research

Four considerable limitations of this study are the sample size, the malfunctions of the eye-tracker, the influence of presentation order, and the use of the consensus.

The small sample sizes caused some issues. The hypothesized model for accuracy could not be tested because either problems with the assumptions arose or validity issues occurred. Moreover, the sample size for fear-induced stress (n = 6) made it difficult to see whether the normality assumption was met. This makes the results for fear-induced stress less powerful. With a larger sample, these issues might be solved.

Due to the malfunctions of the eye-tracker it could not be studied whether the relationship between task division and accuracy was mediated by reliance. Future research

should incorporate the eye-tracker to answer the questions about reliance that remain unanswered in this study.

The order in which the conditions were presented to the participants might have influenced the results. It was counterbalanced whether the participants first conducted two runs with the AI with low accuracy or with high accuracy to consider the possibility of a learning effect. However, this change in order might have affected the results. It is possible that first experiencing an AI with high accuracy and then one with low accuracy might effect history-based trust in the AI differently than first experiencing the AI with low accuracy and then with high accuracy.

The consensus to determine accuracy, sensitivity, and specificity has some flaws. The classification for hits, misses, etc. was done on pixel level. Due to this, the proportion of correct rejections was considerably high as the welds and the flaws within the welds were a small proportion of the whole image. This high proportion of correct rejections influenced the accuracy and specificity estimates. A more advanced program that can determine the flaws instead of the pixels for the consensus might increase the consensus' reliability. Moreover, this consensus remains an approximation of the truth. For the field of weld inspection, it is important to come to an agreement and get a solution on how to determine accuracy. This is important for training the AI and checking the performance of the AI, human inspector, and human-AI team.

Besides tackling the abovementioned limitations in the future, other future research should be carried out as well. First, future research should look more into the effect of individual characteristics and investigate how these impact the outcome variables of performance and well-being. If the influence of these characteristics is known, companies can use this knowledge to take this into account when integrating an AI and avoid unexpected difficulties. Second, future research should be done to further validate the proposed human-AI framework. This study focused on the task & strategy (task division), the individual characteristic propensity to trust, and history-based trust. The effect of other factors should be studied as well. Two considerable interesting factors are vigilance and sustained attention, as high vigilance demands are a challenge in weld inspection (Hou et al., 2020; Teichner, 1974; Woods & Hollnagel, 2006). Future empirical research should confirm or deny the claims of the participants in this study regarding AI supporting the human inspector's sustained attention. If these claims can be confirmed, this will be a large advantage of AI systems in weld inspection.

4.5. Conclusion

The present study provides new insights for the integration of AI systems in weld inspection. Furthermore, it provides a generic theoretical framework for human-AI teams. For

weld inspection, the research results show that, with the current level of AI accuracy, the parallel task division is the best option of the two, because it provides higher accuracy in weld inspection than the sequential task division. Even though the sequential task division is more efficient, it should not be used with the current AI as it jeopardizes the assessment's accuracy. Moreover, the relation between propensity to trust and FoMJ shows that an individual's characteristics should also be considered. Although not all elements of the theoretical framework are addressed in this study, the framework can serve as an aid to see which factors might play a role when integrating AI. Moreover, the research addresses the importance of looking at differences within experimental groups. The individual differences in efficiency show that it is important to look at the differences within the experimental groups.

To conclude, this research provides advice for the integration of AI specific for weld inspection, but the research and especially the theoretical framework can also be used for other human-AI teams to gain a better understanding of variables that play a role here.

5. References

Ali, A. H., Balint, D., Temple, A., & Leevers, P. (2012). The reliability of defect sentencing in manual ultrasonic inspection. *NDT & E International, 51*, 101-110. https://doi.org/10.1016/j.ndteint.2012.04.003

Asan, O., Bayrak, A. E., & Choudhury, A. (2020). Artificial intelligence and human trust in healthcare: focus on clinicians. *Journal of medical Internet Research, 22*(6), 1-7. https://doi.org/10.2196/15154

Aziz, S., Dowling, M. (2019). Machine Learning and AI for Risk Management. In Lynn T., Mooney J., Rosati P., Cummins M. (Eds.), *Disrupting Finances* (33-50). Palgrave Pivot. https://doi.org/10.1007/978-3-030-02330-0_3

Bansal, G., Nushi, B., Kamar, E., Lasecki, W. S., Weld, D. S., & Horvitz, E. (2019a). Beyond accuracy: The role of mental models in human-AI team performance. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7*(1). 2-11. https://ojs.aaai.org/index.php/HCOMP/article/view/5285

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W. S., & Horvitz, E. (2019b). Updates in human-ai teams: Understanding and addressing the performance/compatibility tradeoff. *Proceedings of the AAAI Conference on Artificial Intelligence 33*(1), 2429-2437. https://doi.org/10.1609/aaai.v33i01.33012429

Bertovic, M. (2016). *Human factors in non-destructive testing (NDT): risks and challenges of mechanised NDT* [doctoral dissertation, Technische Universitaet Berlin].

https://verlag.tu-berlin.de/dissdb/

Davies, D. R., & Parasuraman, R. (1982). *The psychology of vigilance*. Academic Press.

DEKRA (n.d.-a). *Radiographic testing.* https://www.dekra.com/en/radiographic-testing/

DEKRA (n.d.-b). *Advanced radiographic testing.* https://www.dekra.com/en/advanced-radiographic-testing/

DEKRA (2016). *Work instruction radiographic testing: Radiographic inspection of welded components on power generation plants.* Unpublished internal company document.

Endsley, M. (1995). Towards a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32-64. https://doi.org/10.1518/001872095779049543

Evans, I., Porter, C., Micallef, M., & Harty, J. (2020). Stuck in limbo with magical solutions: The testers' lived experiences of tools and automation. *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications, 2*. 195-202. https://doi.org/10.5220/0009091801950202

GE Inspection Technologies. (2006). *Industrial radiography: Image forming techniques* (GEIT-30158EN). https://www.jwjndt.com/resource/industrial-radiography-image-forming-techniques/

Hayashi, N., Someya, N., Maruyama, T., Hirooka, Y., Endo, M. Y., & Fukuba, Y. (2009). Vascular responses to fear-induced stress in humans. *Physiology & Behavior, 98*(4), 441-446. https://doi.org/10.1016/j.physbeh.2009.07.008

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors, 57*(3), 407-434. https://doi.org/10.1177/0018720814547570

Hou, W., Zhang, D., Wei, Y., Guo, J., & Zhang, X. (2020). Review on Computer Aided Weld Defect Detection from Radiography Images. *Applied Sciences, 10*(5), 1-16. https://doi.org/10.3390/app10051878

Jessup, S. A., Schneider, T. R., Alarcon, G. M., Ryan, T. J., & Capiola, A. (2019). *Measurement of the propensity to trust automation* [Master thesis, Wright State University]. Core Scholar Wright State University.

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50-80. https://doi.org/10.1518/hfes.46.1.50_30392

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Human Factors, 53*(4), 356-370. https://doi.org/10.1177/0018720811411912

Merritt, S. M., Heimbaugh, H., LaChapell, J., & Lee, D. (2013). I trust it, but I don't know why: Effects of implicit attitudes toward automation on trust in an automated

system. *Human Factors, 55*(3), 520-534. https://doi.org/10.1177/0018720812465081

Merritt, S. M., & Ilgen, D. R. (2008). Not all trust is created equal: Dispositional and history-based trust in human-automation interactions. *Human Factors, 50*(2), 194-210. https://doi.org/10.1518/001872008X288574

Mogford, R. H. (1997). Mental models and situation awareness in air traffic control. *The International Journal of Aviation Psychology, 7*(4), 331-341. doi: https://doi.org/10.1207/s15327108ijap0704_5

Morris, D. M., Erno, J. M., & Pilcher, J. J. (2017). Electrodermal response and automation trust during simulated self-driving car use. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 61*(1). 1759-1762. https://doi.org/10.1177/1541931213601921

Pereira, A. B., & de Melo, F. J. (2020). Quality Assessment and Process Management of Welded Joints in Metal Construction—A Review. *Metals*, *10*(1), 1-18. https://doi.org/10.3390/met10010115

Shaffer, F., & Ginsberg, J. P. (2017). An overview of heart rate variability metrics and norms. *Frontiers in Public Health, 5*. https://doi.org/10.3389/fpubh.2017.00258

Smith, E. E., & Kosslyn, S. M. (2016). Attention. *Brain and Cognition*. Pearson.

Teichner, W. H. (1974). The detection of a simple visual signal as a function of time on watch. *Human Factors, 16*(4), 339–352. https://doi.org/10.1177/001872087401600402

TrueFlaw. (2020a). *Webinar: Machine learning for digital x-ray* [video]. https://trueflaw.com/ml/ml4xray

TrueFlaw. (2020b). *Webinar: Machine learning for UT* [video]. https://trueflaw.com/ml/ml4ut

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction, 5*, 1-39. https://doi.org/10.1145/3476068

De Waard, D. (1996). *The measurement of drivers' mental workload* [doctoral dissertation, rijksuniversiteit Groningen]. https://research.rug.nl/en/publications/

Warm, J. S., Dember, W. N., & Hancock, P.A. (1996). Vigilance and work-load in automated systems. In R. Parasuraman & M. Mouloua (Eds.), *Automation and human performance: Theory and applications* (pp. 183-200). Erlbaum.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human Factors, 50*(3), 433-441. https://doi.org/10.1518/001872008X312152

Wickens, T. D. (2002). *Elementary Signal Detection Theory.* Oxford University Press.

Wilson, J. R., & Rutherford, A. (1989). Mental models: Theory and application in human factors. *Human Factors, 31*(6), 617-634. https://doi.org/10.1177/001872088903100601

Wojton, H. M., Porter, D., T. Lane, S., Bieber, C., & Madhavan, P. (2020). Initial validation of the trust of automated systems test (TOAST). *The Journal of Social Psychology, 160*(6), 735-750. https://doi.org/10.1080/00224545.2020.1749020

Woods, D.D., & Hollnagel, E. (2006). *Joint Cognitive Systems: Patterns in Cognitive Systems Engineering.* Taylor & Francis.

Yerkes, R. M., Dodson, J. D. (1908). The relation of strength of stimulus to rapidity of habit-formation. *Journal of Comparative Neurology and Psychology, 18*(5). 459-482. https://doi.org/10.1002/cne.920180503

Young, M. S., & Stanton, N. A. (2002). Malleable attentional resources theory: a new explanation for the effects of mental underload on performance. *Human factors, 44*(3), 365-375. https://doi.org/10.1518/0018720024497709

Young, M. S., and N. A. Stanton. (2005). Mental workload. In N. A. Stanton, A. Hedge, K. Brookhuis, E. Salas, & H. W. Hendrick (eds.). *Handbook of Human Factors and Ergonomics Methods* (pp. 390-400). London, UK: Taylor & Francis.

Zednik, C. (2021). Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology, 34*(2), 265-288. https://doi.org/10.1007/s13347-019-00382-7

Zeilstra, M. (2020, February 14). *Werkbelasting 3-D: de mens en hoge automatiserings-graad* [PowerPoint slides]. Intergo.

Zijlstra, F. R. H., & Van Doorn, L. (1985). The construction of a scale to measure perceived effort. Technical Report. Delft University of Technology.

Zhou, J., & Chen, F. (2019). Towards trustworthy human-AI teaming under uncertainty. In *IJCAI 2019 Workshop on Explainable AI (XAI)*. https://opus.lib.uts.edu.au/handle/10453/136189

6. Appendices

Appendix A: Explanation of the Theoretical Human-AI Framework

Input Variables		
Relevance: Influence the human-in-the-loop process. The development of an understanding about the AI and trust within the AI starts with communication and how and by whom the AI is introduced (Hoff & Bashir, 2015)		
Variable	**Aspects**	**Additional notes/definitions**
Organisational Characteristics	Goals	• For example, is a company focused more on quantity or quality?
	Communication	• This includes *how* and *by whom.* Communication plays an important role in the eventual acceptance by the user and trust in the system (Lee & See, 2004).
	Workplace design	• Workplace design includes the direct environment.
	Procedures	• For example, are digital x-rays used or conventional x-ray films. According to which standards are the welds inspected?
Individual characteristics	Competence	• How skilled is someone at the task? • There will be differences in skills between individuals
	Motivations	• Motivations can be about the task itself or towards working with an AI
	Vigilance	• This is the ability to maintain attention over a prolonged period of time (Davies & Parasurman, 1982). • Vigilance decrement mostly sets in after 15 minutes for a monitoring task such as weld inspection (Teichner, 1974).

	Propensity to trust	• Trait-like trust, the general tendency to trust automation and/or AI systems.
AI	Reliability	• How well and consistent does the AI perform? The better and more consistent the performance, the higher the trust within the AI (Lee & See, 2004; Zhou & Chen, 2019).
	Confidence	• This is the AI's assessment of its own performance: how certain is the AI that the decision it made is correct?
	Transparency	• How well does the system communicate what it does, why it does it, and what it will do next • AI systems often are not transparent due to their highly complex algorithms they use to make decisions. This lack of transparency makes it hard for a user to understand why an AI makes certain decisions and can impede trust in the AI.
Human-in-the-loop process		
These factors are a result of the input variables. They explain how the input variables impact the output variables and thus why a highly qualitative AI might not work in a certain context or for certain individuals.		
Variable	**Definition**	**Notes**
Reliance	A behaviour: to what extend does the human use the AI? (Lee & See, 2004).	• Appropriate reliance is necessary, both relying too much (*misuse*) and relying too little (*disuse*) will be detrimental for the team performance (Lee & See, 2004) • Although reliance is influenced by many of the other human-in-the-loop

		variables, it is most closely intertwined with trust (Lee & See).
Trust (history-based)	An attitude towards a *specific* AI system (or automation) (Lee & See, 2004).	• Trust impacts reliance. But it varies how much (Lee & See, 2004): ⇨ if the construct to be trusted is too difficult to fully comprehend, trust plays a larger role. ⇨ if it is easy to understand mental models play a larger role. ⇨ As AI systems are fairly difficult, trust is expected to play an important role. • *Calibrated trust* (=not too little and not too much trust) is of importance to ensure appropriate reliance.
Situation Awareness	Awareness of what the AI system is doing at a certain point in time (Endsley, 1995).	• It is not merely about *knowing* what the AI does, it is also important that one *understands* it.
Mental Model	A construct that enables individuals to understand, explain, and predict phenomena in order to decide what action to take (Wilson & Rutherford, 1989).	• Mental models are fairly static. • Situation awareness and mental models are closely intertwined. On the one hand, mental models are used to create situation awareness. Having adequate mental models is a prerequisite for achieving situation awareness. On the other hand, the contents of situation awareness are used to build and modify the mental model (Mogford, 1997).
Mental Workload	The resources that are necessary to meet the task demands (Young & Stanton, 2005).	• Both too low and too high mental workload can be detrimental for the human's performance (de Waard, 1996).

		• As weld inspection is a monotonous task, too low mental workload seems to be a higher risk in this situation. • Adding elements – such as AI – to the environment can both decrease or increase mental workload.

Attention	The ability to select some information for further processing and inhibit other information from receiving further processing (Smith & Kosslyn, 2016).	• Influences on attention include:
		  ○ **Trust:** when trust becomes too high, the human is likely to pay less attention to the task.
		  ○ **Workload:** if there is underload, the inspector might get distracted from the primary task, decreasing attention (Warm et al., 2008).
		  ○ **Mental models:** influence what is attended to (Smith & Kosslyn, 2016).
		  ○ **Vigilance:** the higher the vigilance, the longer one can stay attentive.
		• Influences of attention:
		  ○ **Situation awareness:** low attention is linked to low situation awareness (Endsley, 1995).
		  ○ **Mental workload:** paying attention is a process that requires cognitive resources and thus increases mental workload (Warm et al., 2008).

Arousal	The state necessary to sustain the demanded attention (Yerkes & Dodson, 1908).	• The optimal level of arousal has an inverted U-shape (Yerkes & Dodson, 1908).

		⇨ When arousal is too low, one does not get enough stimulation and attention decreases ⇨ When arousal is too high, one has too much stimulations.

Output Variables

These are the variables that are of interest for the company, such as performance and well-being of the employee. Although they are called output variables, it is important to be aware that they also feed back to the human-in-the-loop variables and/or the input variables.

Output Variable	**Notes**
AI detection performance	• The AI performance is determined solely by the AI's performance.
Human detection performance	• The human performance reflects how well the human performs the task on its own. • It is influenced by many of the input and human-in-the-loop variables, such as skills, attention, and mental workload.
Decision making	• The human needs to decide whether to go with the AI's decision or their own decision if the two decisions do not match. • This variable decides whether the human or AI's performance has more influence on the final team performance (accuracy) • This decision making can be seen as a form of *reliance.*
Team performance (accuracy)	• Accuracy is how well the team has classified all the x-rays and flaws. • It is of importance to find all flaws to avoid possible catastrophically consequences. • However, it also is important that the team does not indicate many parts that actually aren't flaws (*False Alarms*).
Team performance (efficiency)	• Efficiency is how fast the team can inspect all the welds. • Efficiency and accuracy are often traded of against each other and one should make sure that they are well-balanced.

Human well-being	• The well-being of the employee should also be taken into account. This includes factors such as stress, satisfaction, and the feeling of having a meaningful job.

Appendix B: Meaningful Job Questionnaire

Employee
• The system and I complement each other • The system motivates me to do my work better • I enjoy my work more when using the system • I perceive my job as useless when I work with the system (R*) • I learn new things when working with the system • I think the system takes over my job (R)
Company/employer
• As a team, a collaboration between human and artificial intelligence, we can meet the demands of the employer better as either of us alone • As a team we can meet the demands of the employer more efficient as either of us alone • The system can meet the demands of the employer on its own (R)
Society
• The reliability of welds will increase when the system and I work together • The system alone can guarantee that the safety norms are met (R)

Note. The items for the Feeling of Meaningful Job interview were divided into three categories. Items related to the employee, employer/company, and items related to society.
*These items are reversed.

Appendix C: Syntax Statistical Model for Accuracy

Different statistical models were considered to test the proposed relationships to determine accuracy (Figure 5). Unfortunately, different problems arose.

First a linear mixed model was created with SPSS to test the proposed relationships. The syntax was as follows:

```
MIXED Accuracy WITH PropensityToTrust BY TaskDivision Country
/FIXED=Country PropensityToTrust TaskDivision PropensityToTrust*TaskDivision
/PRINT=DESCRIPTIVES SOLUTION
/RANDOM=intercept | SUBJECT(id) COVTYPE(VC).
```

This syntax gave the following error in SPSS (both for the dataset for high and low AI accuracy): *The final Hessian matrix is not positive definite although all convergence criteria are satisfied. The MIXED procedure continues despite this warning. Validity of subsequent results cannot be ascertained.*

There are three common causes of this error:

1. The number of fishing scoring steps is too low. Increasing the number of step-halvings can solve the problem.
2. The covariance structure is too complex. A simpler structure specification might solve the problem.
3. A subject variable on the RANDOM comment is neglected. Adding a subject variable can solve the problem.

First it was tested whether increasing the number of fishing score steps would work. This possible solution was tested first as there was no need to change the variables in the model for this. The MXSTEP was increased to a 1000. Unfortunately, the error remained.

The syntax was as follows:

```
MIXED Accuracy WITH PropensityToTrust BY TaskDivision Country id
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(1000) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0,
    ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED=Country performance_order PropensityToTrust TaskDivision
PropensityToTrust*TaskDivision | SSTYPE(3)
  /METHOD=REML
  /PRINT=SOLUTION
  /RANDOM=id | COVTYPE(VC).
```

Second the statistical model itself was changed. It was tried to take Country out as covariate. This could however increase noise, because the version of the experiment (on-location in Germany and online in South-Africa) or the nationality might play a role. The following model was tested, but the error remained:

```
MIXED Accuracy WITH PropensityToTrust BY TaskDivision id
  /CRITERIA=CIN(95) MXITER(100) MXSTEP(1000) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0,
    ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
  /FIXED= PropensityToTrust TaskDivision PropensityToTrust*TaskDivision | SSTYPE(3)
  /METHOD=REML
  /PRINT=SOLUTION
```

```
 /RANDOM=id | COVTYPE(VC).
```

As this did not work, the decision was made to take out propensity to trust and only test the effect of Task Division on Accuracy. However, the error still remained. The syntax for this model was as follows:

```
MIXED Accuracy BY TaskDivision id
 /CRITERIA=CIN(95) MXITER(100) MXSTEP(1000) SCORING(1)
SINGULAR(0.000000000001) HCONVERGE(0,
   ABSOLUTE) LCONVERGE(0, ABSOLUTE) PCONVERGE(0.000001, ABSOLUTE)
 /FIXED= TaskDivision | SSTYPE(3)
 /METHOD=REML
 /PRINT=SOLUTION
 /RANDOM=id | COVTYPE(VC).
```

It was not tested whether adding a subject variable on the RANDOM subcommand would work, because to the knowledge of the researcher there was no additional variable besides ID that would be logical to add here.

Unfortunately, this means that the problem could not be solved. As the last sentence of the error states ‘validity of subsequent results cannot be ascertained’. For this reason, an alternative model had to be tested.

I tried to use general linear models. However, the data did not meet the assumptions for any of the general linear models. For this reason, it was chosen to alter the model to be tested and use the non-parametric Friedman’s test.