Data or interpretations: Impacts of information presentation strategies on diagnostic processes

Romy Müller | Christina Gögel | Rica Bönsel

Faculty of Psychology, Institute for Psychology III, Chair of Engineering Psychology and Applied Cognitive Research, Technische Universität Dresden, Dresden, Germany

Abstract

Industrial fault diagnosis can be supported by assistance systems that infer fault causes from sensor data. The present study asked what information these algorithms should make available to operators. In a computer-based experiment about fault diagnosis in a packaging machine, three information presentation strategies were compared regarding their impacts on information sampling, performance, and knowledge acquisition: Providing only sensor data, sensor data along with three possible interpretations, or only the most likely interpretation. Before submitting a diagnosis, participants could sample process parameters, one of which indicated the fault cause. We hypothesized that providing only sensor data would lead to more parameter checking and slower solutions than interpretations. While providing only one interpretation was expected to enable efficient performance for correct interpretations, it should lead to either of two types of performance costs for incorrect interpretations: Errors if participants refrain from checking parameters, or slowdowns in performance if they keep on checking. The results confirmed that participants with only sensor data performed inefficiently. Participants with only one interpretation thoroughly checked parameters but still were fastest when the interpretation was correct, while when it was incorrect they were three times slower than participants with only sensor data. Participants with three interpretations (one of which was always correct) performed almost as efficiently as those with only one correct interpretation. The results indicate that highly preprocessed information leads to efficient performance when it is correct but prevents learning about fault causes. Overall, providing several possible interpretations seemed to be the best strategy.

Keywords

automation bias, degree of automation, fault diagnosis, information sampling, operator assistance

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

© 2020 The Authors. Human Factors and Ergonomics in Manufacturing & Service Industries published by Wiley Periodicals, Inc.
INTRODUCTION

Fault diagnosis in the processing industry

The processing industry is concerned with the mass production of consumer goods such as food, beverages, and cosmetics. In linear production lines, three-dimensional objects such as cookies or cheese are handled by complex machinery at extremely high speeds. Processing plants are error-prone as a result of complex interactions between the machines and characterstics of the product, packaging material, and environment (Bleisch, Majschak, & Weiß, 2011; Müller & Oechm, 2019; Schult, Beck, & Majschak, 2015; Tsarouhas, 2012). For instance, different types of chocolate react differently to temperature changes (e.g., milk chocolate melts faster than dark chocolate at high temperatures, while dark chocolate breaks faster at low temperatures), and temperature also affects the packaging material (e.g., if the foil is too cold, water condensates on its surface). The consequences of these interactions often cannot be foreseen even by system designers (Hollnagel, 2012; Perrow, 1984), which is why operators need to act as the designer’s extended arm (Rasmussen & Goodstein, 1987). However, fault diagnosis is challenging as one and the same fault can have different symptoms, and one symptom can result from different faults (Cai, Liu, & Xie, 2016). For instance, chocolate bars that are bigger than usual can cause product jams on conveyor belts or increased soiling of machine components. At the same time, soiling can also be caused by high temperatures in the plant or insufficient cooling of the machine. Thus, dealing with faults is a major challenge leading to reduced productivity (Schult et al., 2015): A common problem is that operators quickly remove fault symptoms instead of diagnosing their underlying causes. They can be supported by assistance systems, but it is an open question how such support should be designed. Assistance systems are rare in the processing industry, and the existing ones usually provide instructions for specific actions. During fault diagnosis, this is insufficient as assistance systems cannot always know which action is most suitable. This is because the complex interactions that characterize the domain make it necessary to consider the current context. In light of Industry 4.0 (Lasi, Fettke, Kemper, Feld, & Hoffmann, 2014), this context can be made more transparent as an increased amount of data is available throughout the system and can be used to support operators’ diagnostic actions.

Using sensor data in assistance systems

Fault diagnosis can be supported by assistance systems that use algorithms to infer fault causes from sensor data. Sensors such as light barriers, cameras, and scales provide important information about the current process and potential problems (Mahalik & Nambiar, 2010; Rauber, Barata, & Steiger-Garcia, 1993; Scott, 2008). As sensor data is not readily discernible and inherently ambiguous, it needs to be processed. There are many technological approaches to inferring fault causes from sensor data (Venkatasubramanian, Rengaswamy, & Kavuri, 2003; Venkatasubramanian, Rengaswamy, Kavuri, & Yin, 2003; Venkatasubramanian, Rengaswamy, Yin, & Kavuri, 2003). However, it is unclear how the information provided by these algorithms should be used to support operators’ diagnostic actions.

One way of supporting operators’ diagnostic actions is to provide hypotheses about fault causes based on an algorithm’s interpretation of sensor data. Operators could benefit from this as the generation of hypotheses about fault causes is an important determinant of diagnostic success (Abele, 2017; Adsit & London, 1997; Thomas, Dougherty, Sprenger, & Harbison, 2008). Yet, providing only one automatically generated interpretation might discourage operators from considering alternative hypotheses. Therefore, another strategy is to provide operators with low-level sensor data and leave the interpretation up to them. However, sensor data is hard to interpret for humans (Hoskins, Kaliyur, & Himmelblau, 1991), and thus a third strategy is to provide sensor data along with a list of possible interpretations. In this way, operators have access to several hypotheses and the data underlying them. The present study compares these three strategies, asking how diagnostic processes are affected by a presentation of low-level sensor data or high-level interpretations of this data. To provide a basis for understanding the potential effects of these strategies, the following section summarizes research on the processing of information during diagnostic problem solving and the consequences of automated support.

Diagnostic problem solving and automated support

Diagnostic problem solving can be considered as a process of representing, generating, testing and evaluating hypotheses (Abele, 2017). Information processing during all these stages is subject to several limitations and biases. First, information representation (which relies on information sampling) is often biased, for instance as a consequence of accessibility (Culnan, 1983), cue salience (Platzer & Bröder, 2012), the familiarity of information sources (Fidel & Green, 2004), or people’s habits (Verplanken, Aarts, & Van Knippenberg, 1997). People usually are not aware of these sampling biases but treat their samples as if they were fully representative (Fiedler & Kutzner, 2015). Second, problems can arise during hypotheses generation. Often, hypotheses are generated too quickly (i.e., immediately after identifying an alarm) instead of being preceded by systematic information acquisition (Patrick, Gregov, Halliday, Handley, & O’Reilly, 1999). Moreover, people tend to generate only one or a few hypotheses (Mehle, 1982; Patrick et al., 1999), or generate hypotheses merely based on experience rather than based on a model of the technical system (Abele & von Davier, 2019; Bereiter & Miller, 1989). Third, hypotheses testing and evaluation is often biased. Problem solvers begin by testing the most common problem or focus on hypotheses that are easy to test, and test the same hypotheses repeatedly (Bereiter & Miller, 1989). They are prone to confirmation bias, selectively searching for information in line with their current hypothesis (Klahr & Dunbar, 2000; Morrisson & Upton, 1994;
Nickerson, 1998), react more strongly to information supporting their current hypothesis (Rebitschek, Krems, & Jahn, 2015), and reinterpret available evidence when being convinced that a hypothesis is true or false (Arocha, Patel, & Patel, 1993). Their final diagnosis often reflects their initial hypothesis (Rebitschek et al., 2015). These biases in different stages of the diagnostic process need to be taken into account when designing assistance systems for fault diagnosis: When asking what information these systems should make available to operators, it is important to keep in mind that people tend to neglect particular information sources, focus on a limited set of hypotheses, and do not test all relevant hypotheses sufficiently.

This raises a number of questions about the design of assistance systems for fault diagnosis. For instance, to what extent should algorithms process sensor data before presenting it? Should operators generate hypotheses themselves, while the automated system provides low-level information about sensor data? Or should automated support systems take over hypotheses generation and merely present one or several possible fault causes? In Human Factors research, the extent to which an automated system provides support is referred to as the degree of automation (DOA; Parasuraman, Sheridan, & Wickens, 2000; Wickens, Li, Santamaria, Sebok, & Sarter, 2010). It describes which stages of human information processing are supported and how highly each stage is automated. Four human information processing stages are distinguished: information acquisition, information analysis, decision selection, and action implementation (Parasuraman et al., 2000). A higher DOA is realized through higher automation within each stage and the inclusion of later stages. For instance, when using sensor data to support fault diagnosis, automated information acquisition could mean to directly tell operators what a sensor is doing (e.g., sensor X switches irregularly), while information analysis could mean to further interpret sensor data to infer possible fault causes and only present the results (e.g., the products are positioned incorrectly or the sensor is dirty). The latter strategy has been proposed by Lee and Seong (2007), who designed a support system for fault diagnosis in nuclear power plants that is based on neural networks and provides a list of possible fault causes. Lists of fault causes that reflect plausible hypotheses can improve diagnostic performance (Raaijmakers & Verduyn, 1996): Naval engineers using such lists were able to solve more diagnostic problems than when they had to generate hypotheses by themselves. The DOA increases even more when the system selects or implements appropriate actions. Such support has been proposed by Borlea, Buta, Dusa, and Lustrea (2005): In addition to a fault cause, their system suggests a corrective action to operators of an electrical substation.

Human Factor research has shown that a higher DOA improves performance and decreases workload in routine situation, but can have negative consequences when the automation fails (Onnasch, Wickens, Li, & Manzey, 2014). One of these consequences is automation bias—an uncritical acceptance of automated suggestions (Mosier, Skitka, Heers, & Burdick, 1998). Automation bias develops over time through the interaction with a seemingly reliable system and is associated with inattentive sampling of information that could be used to verify the suggestions of the automation (Parasuraman & Manzey, 2010). While there is a large body of research on automation bias in different tasks, few studies have examined it in the context of fault diagnosis. Manzey, Reichenbach, and Onnasch (2012) examined the information sampling behavior of participants required to detect and fix malfunctions in a simulated space capsule while being supported by an automated system with varying DOA. Under the highest DOA, the system sent repair orders and participants simply had to confirm them, whereas under lower DOAs they had to detect malfunctions and identify appropriate repair actions themselves. They could verify the automated information by checking different process parameters. With a high DOA, participants spent less time cross-checking parameters, especially when it required complex interventions. Using the same task, Bahner, Hüber, and Manzey (2008) found that insufficient information sampling preceded diagnostic errors. In their study, an automated system provided advice on corrective actions for fault management, and sometimes this advice was wrong. Participants who accepted the wrong advice checked fewer relevant parameters than participants who did not accept the advice.

It needs to be noted that such lack of cross-checking has not been observed in all studies. First, Manzey, Gérard, and Wiczorek (2014) found that participants responding to alarms issued by an automated system cross-checked the alarm whenever possible, even when cross-checking was effortful and several parameters had to be checked to completely verify the alarm. Second, Lorenz, Di Nocera, Rötger, and Parasuraman (2002) found that when an automated system implemented corrective actions, participants sampled more information than when they had to implement corrective actions by themselves. The authors argued that this was a consequence of decreased workload, with the highest DOA making it possible to use additional resources to verify the automation. Third, Müller, Ulrich, et al. (2019) found no evidence for automation bias when participants used a case-based assistance system to diagnose faults in a packaging machine: When highly recommended cases were inappropriate in the current situation, participants did not accept these cases but thoroughly checked the available information and thus arrived at correct solutions.

Taken together, fault diagnosis in the processing industry is a challenging task that requires competent operator intervention. Operators’ diagnostic problem solving can be supported by algorithms that use sensor data to infer fault causes. When deciding what information these algorithms should make available, two things need to be considered: First, assistance systems should take into account human biases related to the representation of information as well as the generation, testing, and evaluation of hypotheses. These problems reflect that humans only use limited information sources to generate and test a limited set of hypotheses. Second, the implementation of higher DOA can increase system performance in routine situations but comes with the risk of discouraging operators from cross-checking system parameters, so that automation failures go unnoticed. This creates a dilemma: While the presentation of high-level interpretations of sensor data (i.e., hypotheses about fault
causes) is likely to mitigate human difficulties in hypotheses generation, it could lead to automation bias. Previous studies on presenting sensor data during fault diagnosis did not address these issues as they focused on the technological implementation of effective algorithms (Borlea et al., 2005; Cai et al., 2016; Lee & Seong, 2007), while impacts on human performance have not been studied so far.

1.4 | Present study

The present study investigated how different strategies of presenting sensor data or its interpretations—and thus different degrees of automation—affect human information sampling, performance outcomes, and knowledge acquisition during fault diagnosis. To this end, participants worked with a simulated assistance system to diagnose faults in a chocolate packaging machine. Three groups of participants received different supportive information: In a first group (data-only), they were informed about the abnormal sensor signal (e.g., sensor G1 at the feed conveyor signals that the chocolate bars are too lightweight). This corresponds to a very low DOA, with the assistance system merely acquiring and presenting information about the current machine state. In a second group (data+3interpretations), the same sensor signal was presented along with a list of possible interpretations corresponding to different fault causes (e.g., bars have hollow bottoms, bars have broken corners, bars have too little filling). This format represents an intermediate DOA as it provides both low- and high-level information, allowing operators to decide which interpretations they want to test. In a third group (1interpretation+data), only the interpretation favored by the algorithm was presented (e.g., bars have too little filling). This corresponds to a high DOA as information acquisition and analysis are automated and only the result is made available. Even this DOA can be considered lower than in most previous studies, where a high DOA usually meant that the system automatically selected and implemented repair actions (e.g., Manzey et al., 2012). However, in the context of a task with the ultimate goal of submitting a diagnosis, providing just one interpretation highly restricts operators’ information space and thus serves as a proxy for decision selection as no alternatives are offered. To make sure the task remained solvable even in case of automation failure, the latter group had the additional opportunity to view sensor data on demand. All groups had access to a screen presenting more than 50 process parameters, the values of which could be requested by individual mouse clicks. Exactly one of these parameters deviated from its desired value and thus represented the fault cause. This parameter always matched a cause indicated by one of the three interpretations in the data+3interpretations group. Also, it almost always matched the provided interpretation in the 1interpretation+data group, except for two catch trials towards the end of the experiment in which the interpretation was incorrect and thus referred to a parameter that actually was within its desired range, while another unknown parameter was faulty.

A first set of hypotheses concerned participants’ information sampling behavior. We hypothesized that interpretations would lead to reduced sampling, reflected in a lower number of parameter checks in the two groups receiving interpretations than in the data-only group. However, based on previous findings (Manzey et al., 2014; Müller, Ullrich, et al., 2019) we did not expect participants to refrain from cross-checking in the 1interpretation+data group. Rather, we expected them to focus on exactly one parameter, namely the one indicated by the interpretation, and test only this parameter in standard trials. In catch trials, this test should lead participants to notice the information failure, and thus check as many parameters as the data-only group. A second set of hypotheses concerned diagnostic outcomes. We expected lower solution time with increasing DOA. Thus, the data-only group should take most time to submit a diagnosis, followed by the data+3interpretations group. The 1interpretation+data group should be faster than both other groups, given that in standard trials only a quick check of one parameter was needed to verify the automated interpretation. Conversely, in catch trials we expected solution times in the 1interpretation+data group to rise to the level of the data-only group. Differences in error rates between groups or between standard and catch trials were not expected, given that we assumed that participants in the 1interpretation+data group would notice the automation failure as a result of continued cross-checking. Thus, we assumed automation failures to only reduce diagnostic efficiency by increasing time and effort, but not to impair diagnostic accuracy. A final hypothesis concerns the acquisition of knowledge resulting from the interaction with different types of information presentation: We anticipated the presentation of several interpretations to foster learning about fault causes. Accordingly, when asked about possible causes of sensor data after the experiment, we expected the data+3interpretations group to provide more answers and a higher percentage of correct answers than the two other groups.

2 | METHODS

2.1 | Participants

Participants were recruited from the TU Dresden participant pool (ORSEE, Greiner, 2015), via Facebook, and via announcements on campus. In total, 67 people participated in the study. One did not pass the knowledge test after the instruction seminar and one could not participate in the main experiment due to personal reasons. Of the remaining participants, 26 were male and 39 female. Their age ranged from 19 to 70 years (M = 27.7; SD = 10.5). They were randomly assigned to the experimental groups, resulting in 22 participants for data-only, 21 participants for data+3interpretations, and 22 participants for 1interpretation+data. Participants either received 7€ per hour or partial course credit. All procedures followed the principles of the Declaration of Helsinki.
2.2 | Apparatus and stimuli

2.2.1 | Instruction seminar

Before the experiment, participants went through an instruction seminar based on a Microsoft PowerPoint presentation to ensure that they acquired basic knowledge about the machine and its sensors. The instruction consisted of three parts: An overview of the process of producing and packaging chocolate bars, detailed information about the packaging machine, and the functioning of sensors. The first part explained each step in the process of producing and packaging chocolate bars (molding, transporting, buffering, packaging). The second part presented seven unit operations in the packaging machine (conveying, separating, foil unwinding, foil cutting, injecting, packing, discharging) and explained each of them in detail, including information about the machine parts that contribute to these unit operations. The third part described how sensors are used in the machine, according to which principles they work, and where they are positioned. Thus, the instruction seminar served as a knowledge base about the technical system. Importantly, it did not provide any information concerning fault causes or the ambiguity of sensor data, and it did not instruct participants how to solve the experimental task. To facilitate learning, a multimodal presentation with text, pictures, and videos used several instructional techniques such as animations (Mayer & Moreno, 2003), advance organizers (Mayer, 2008), test questions (Nungester & Duchastel, 1982), and summary slides. After the instruction seminar, participants completed a knowledge test about the seminar contents which consisted of 14 multiple-choice questions with four response options.

2.2.2 | Main experiment

Experiments were conducted in individual sessions. Stimuli were presented on a 19” monitor with a resolution of 1680 × 1050 pixels, and all inputs were made with a standard computer mouse. The experiment was programmed using the Experiment Builder (SR Research, Ontario, Canada) and consisted of 34 fault scenarios. To solve these scenarios, participants could access an assistance screen, a parameter screen, a parameter checklist, a schematic drawing of the machine, and a laptop for entering fault causes.

**Assistance screen**

The assistance screen presented text in white and yellow font as well as blue buttons with white labels on a black background. It consisted of three areas: (a) an alarm message and the corresponding information about data or interpretations, (b) a box containing further information about sensors, and (c) two buttons to check parameters or submit a fault cause. In the upper half of the screen, the alarm message and information about data or interpretations were presented. The alarm message stated that a fault had occurred and the assistance system had analysed it, followed by trial-specific information depending on the experimental group (see Figure 1): For **data-only**, the deviating sensor data was provided (e.g., “Sensor G1 at the feed conveyor signals that the chocolate bars are too lightweight”). For **data+3interpretations**, the deviating sensor data was accompanied by a clickable text saying “The assistance system names the following possible causes (click here)”. Upon clicking this text, three interpretations were presented (e.g., “Bars have hollow bottoms”, “Bars have broken corners”, “Bars have too little filling”). The order of these interpretations was counterbalanced across participants. For **1interpretation+data**, the alarm message said “The assistance system names the following cause” and was followed by only

![FIGURE 1](image-url)
one interpretation (e.g., “Bars have too little filling”), which was randomly selected from the three interpretations of the data interpretations group. It was accompanied by a clickable text saying “Sensor data underlying this fault (click here)”, and clicking this text revealed the deviating sensor data. In the lower left part of the screen, a box contained further information about sensors. It included two clickable questions: “What type of sensor?” and “What does the sensor measure?” Clicking the first question revealed the technical functioning principle (e.g., scale, light barrier), while clicking the second question revealed the data processed by the sensor (e.g., weight of products, reflection of infrared light). In the lower right part of the screen, two buttons could be clicked to access the parameter screen or submit a fault cause.

**Parameter screen**

The parameter screen included 56 clickable parameter buttons and a “back” button that brought participants back to the assistance screen (see Figure 2). Parameter buttons represented features of the packaging process and were color-coded according to seven categories (i.e., sensor states and settings, machine states and settings, interaction of machine and product, product characteristics, interaction of product and packaging material, packaging material characteristics, environment). Clicking a parameter revealed its current value, which appeared as a black text on gray background overlaid on the respective parameter button. Only one parameter could be opened at a time, and clicking a new parameter closed the previous one. Additionally, parameters could be closed by clicking the black areas to the left and right of the parameter screen. Parameter values were fixed within a trial but varied between trials. Deviations from a parameter’s desired value or range indicated a fault cause, and exactly one deviation was present in each trial. Note that each interpretation referred to exactly one of the 56 parameters. For instance, if an interpretation stated “turning wheel is too old”, the corresponding parameter was “turning wheel age”. The same 56 parameters were presented in each trial, but their relevance varied between trials as it depended on the scenario (i.e., fault type). Parameter relevance was defined as to whether a parameter deviation could possibly have caused the respective sensor data. Ratings by two machine experts revealed that 7–14 parameters were potentially relevant in each trial. No information about parameter relevance was made available to participants but this classification was used for data analysis.

**Additional materials**

On participants’ desks, a printed parameter checklist was available that listed all parameters along with their desired values or value ranges, as well as a schematic machine drawing. This drawing included all sensor names, and their respective positions were marked with arrows. Moreover, a laptop was positioned to participants’ left and provided an Excel sheet into which participants had to type the fault causes, using one row for each trial/cause.

**Postexperimental knowledge test**

The postexperimental knowledge test was administered as a paper & pencil questionnaire consisting of four questions. Each question named a sensor signal that had been presented during the experiment and asked participants to write down possible causes. Twelve horizontal lines marked spaces for each cause to indicate that participants should write down as many causes as they could think of.

### 2.3 Procedure

Before the experiment, participants attended an instruction seminar in groups of up to 15 people. Seminars were held by CG (second
author) or a research student, and followed by a written multiple-choice knowledge test about the seminar contents. To ensure that participants had understood the seminar contents, test scores had to exceed 75% to participate in the main experiment. The instruction seminar and knowledge test took about 30 min. Afterwards, individual appointments were made for the main experiment, with the constraint that it had to be within 1 week from the seminar.

In the computer-based main experiment, participants were confronted with faults occurring during the operation of a packaging machine for chocolate bars. Their task was to diagnose fault causes with the help of a simulated assistance system. For each fault scenario, they had to read the fault description and use their group-specific information (i.e., sensor data and/or interpretation/s). Moreover, they could request further information about the sensor, access the parameter screen to identify fault causes, and finally submit a fault cause into an Excel sheet.

The experiment consisted of 35 trials in total: 1 exploration trial, 5 practice trials, and 29 experimental trials. Participants completed the exploration trial with the experimenter guiding them through the interface. During this process, the experimenter explained the contents of the different screens and additional materials. She explained how participants could access this information and told them what it meant, but did not suggest any strategies for how to use the information. For instance, participants were shown which button to click to get to the parameter screen, were told that it represented the current states of all process parameters, and were told that the parameter checklist contained the desired values or value ranges, but they were not instructed to check the parameter screen to reach a diagnosis. After the exploration trial, participants performed five practice trials on their own while being allowed to ask questions. Exploration and practice trials were identical for all participants, and the scenario of the exploration trial was re-used in one of the practice trials. After practice, participants performed 29 experimental trials. Each trial presented a different fault scenario, and thus different sensor data and different interpretations. Scenarios were randomly assigned to trial positions. There were 27 standard trials and two catch trials at positions 23 and 28. Catch trials were only perceivable in the 1|interpretation+data condition as they were defined by the interpretation being incorrect (i.e., not indicating the deviating parameter and thus the actual fault cause). The decision to use only two catch trials and to present them at the end of the experiment was made for the following reasons: First, we wanted to put our hypothesis that people do not refrain from cross-checking (i.e., show no signs of automation bias) during fault diagnosis to a critical test. Previous research has shown that automation bias only emerges after considerable experience with a reliable system (Parasuraman, Molloy, & Singh, 1993) and that once people have experienced an automation failure, they become immune to automation bias and thoroughly cross-check the available information in subsequent trials (Bahner, Elefandt, & Manzey, 2008; Manzey et al., 2012). As this effect of experience is very robust and persists over long periods of time, introducing automation failures early in the experiment would have reduced the potential for automation bias to occur. Second, we wanted to increase the practical relevance of our study. If we had used a less reliable system, an obvious argument from practitioners would have been that the present results only occur with unreliable systems but have no implications for their reliable systems in which errors are the exception.

At the start of each trial, participants received an alarm message along with the group-specific information: In data-only, they saw the sensor data but received no information about possible causes. In data+3interpretations, they also saw the sensor data but could access three possible causes by clicking on the respective text. One of the three interpretations was always correct. Its position was randomized and its identity was counterbalanced across participants. In 1|interpretation+data, participants saw the interpretation favored by the assistance system and could access the sensor data by clicking on the respective text. The interpretation was correct in all but two trials (i.e., catch trials). Irrespective of experimental group, participants could click the questions about the sensor to access further information about its type and measured data. Moreover, they could click the upper blue button to access the parameter screen, and switch back and forth between the assistance and parameter screen.

In the parameter screen, participants could click as many parameters as they wanted (one at a time) to check the current parameter values. They could compare these values to the desired values or ranges provided on their paper checklist to detect parameter deviations and thereby identify the fault cause. Detecting a deviation also informed them that no other parameters would deviate. Note that in the 1|interpretation+data group, the cause could be identified without ever accessing the parameter screen as this cause was reflected in the interpretation. Thus, when participants in this group did check the parameter screen and found the deviation, this corresponded to an indirect feedback that they could have trusted the interpretation right away (up to the first catch trial in trial 23).

Once participants wanted to submit the fault cause, they could click the lower button on the assistance screen. This transferred them to a screen showing a schematic drawing of the machine and text prompting them to enter the cause into their laptop. From this screen, no switching back to the assistance or parameter screen was possible. After submitting the cause, participants had to press the Space bar to start the next trial.

After completing the experiment, participants performed the post-experimental knowledge test. During this test, they were allowed to use the schematic drawing of the machine (indicating all sensor positions) but no other materials. In total, the experiment took between 1.5 and 2 hr.

3 RESULTS

Unless specified otherwise, all data were analysed using 3 (information presentation: data-only, data+3interpretations, 1|interpretation+data) × 2 (trial type: standard trial, catch trial) mixed measures analysis of variance with the within factor trial type and the between factor information presentation. Note that catch trials were only
present in the 1interpretation+data group, but as they occurred at fixed positions (i.e., trial 23 and 28), it was possible to compare these trial positions between groups. All pairwise comparisons were performed with Bonferroni correction. In case of significant interaction effects, only pairwise comparisons for the interaction are reported.

We analysed parameter sampling behavior, solution times, error rates, and knowledge test performance. Error rates were very low (see below), but to make sure that errors did not affect the results, all analyzes were re-calculated while excluding erroneous trials. However, as this had no impact on the results, the reported analyzes include all trials regardless of correctness. Five participants accidently ended trials by clicking the "submit cause" button before finding a cause and explicitly noted this. These trials, seven in total, were excluded from all analyzes.

3.1 Parameter sampling behavior

The number of parameter checks refers to the mean number of different parameters a participant checked during each trial. There were main effects of information presentation, $F(2,62) = 25.029$, $p < .001$, $\eta_p^2 = 0.447$, and trial type, $F(1,62) = 17.173$, $p < .001$, $\eta_p^2 = 0.217$, as well as an interaction of both factors, $F(2,62) = 15.143$, $p < .001$, $\eta_p^2 = 0.282$ (see Figure 3). In standard trials, participants checked more parameters with data-only than with data+3interpretations or 1interpretation+data (12.3 vs. 3.2 and 1.4, respectively), both $ps < .001$. The difference between the two conditions including interpretations did not reach significance, $p = .096$, although numerically the 1interpretation+data group only checked half as many parameters. Conversely, in catch trials participants with 1interpretation+data checked a number of parameters comparable to that with data-only (11.4 and 13.3, respectively), $p > .9$, and both groups checked more parameters than with data+3interpretations (2.7), both $ps < .001$. Note that this analysis does not consider repeated checking of the same parameter, which is not reported as the results were comparable to the ones of checking different parameters. Also, no analysis of the percentage of irrelevant parameter checks is not reported as the results mirrored those of the absolute number of parameter checks.

The previous analysis indicates a strong impact of information presentation on parameter sampling behavior, which is modulated by trial type: Catch trials in the 1interpretation+data group lead to increased sampling (see Figure 4 for a depiction of individual participants’ parameter checks). To better understand this behavior, we analyzed which parameters participants in 1interpretation+data checked and how often they did not check any parameters: In 73.2% of the trials, they checked exactly one parameter. In all but 10 trials across all participants, this was the parameter indicated by the interpretation. There was only one participant who never checked parameters and always accepted the interpretation provided by the assistance system. Three other participants failed to check parameters in one to three trials. Thus, except for one participant there was no indication of automation bias, which is further supported by the finding that half of the participants with 1interpretation+data checked the underlying sensor data in more than 90% of the trials.

3.2 Solution times

Solution time was defined as the interval from the start of a trial until clicking the "submit cause" button. That is, we deliberately excluded typing time as we assumed participants' typing speed to be independent of the experimental manipulations. Participants had to identify the cause on one computer but type it into a different computer. This led to a number of situations in which participants typed in the cause before having clicked the "submit cause" button. As this would bias solution times, we defined a minimum required typing duration of 4 s (i.e., time from pressing the "submit cause" button until starting the next trial). This cut-off was determined by visually inspecting the data: When plotting the solution times for each individual trial per participant, this revealed a clear separation between two distributions: solution times that were very short and solution times that were considerably longer, with only very few data points in between. A value of 4 s best differentiated between these two types of solution times. Thirteen participants were below this cut-off in more than 65% of the trials and thus were excluded from the following analyzes altogether (5 in data-only, 4 in data+3interpretations, 4 in 1interpretation+data). For the remaining participants, only those trials were included in which typing duration exceeded the cut-off, leading to an additional exclusion of 1.9% of the data.

There were main effects of information presentation, $F(2,49) = 7.979$, $p < .001$, $\eta_p^2 = 0.246$, and trial type, $F(1,49) = 13.637$, $p < .001$, $\eta_p^2 = 0.218$, as well as an interaction of both factors, $F(2,49) = 22.586$, $p < .001$, $\eta_p^2 = 0.480$ (see Figure 5). In standard trials, it took longer to identify the fault cause with data-only than with data+3interpretations or 1interpretation+data (7.4 vs. 5.04 and 3.05 s, respectively), both $ps < .003$. Additionally, solution times were longer with data+3interpretations than with 1interpretation+data,
Conversely, in catch trials causes were identified faster with data-only and data+3 interpretations than with 1interpretation+data (56.0 and 45.1 s, respectively), both ps < .001, while solution times did not differ between data-only and data+3 interpretations, p > .9. Thus, while participants with 1interpretation+data were fast in standard trials, their performance broke down in catch trials, where they were slower than both other groups. Numerically, their solution times in catch trials were at the level of the data-only group’s very first trial.

So far, the analyzes suggest that with 1interpretation+data participants solved their task faster than with data+3 interpretations, while the difference in parameter checks did not differ significantly. Does this indicate that 1interpretation+data participants sampled parameters faster, putting less effort into the processing of each individual parameter? To investigate this, we analysed parameter check rates (i.e., parameters per second). There was a main effect of information presentation, \( F(2,49) = 23.535, p < .001, \eta^2_p = 0.490 \), but no effect of trial type and no interaction, both Fs < 1. With data-only, participants checked parameters at a higher rate than with data+3 interpretations and 1interpretation+data (0.165 vs. 0.072 and 0.055 parameters per second, respectively), both ps < .001, while parameter check rates for the two groups with interpretations did not differ, p > .9. Thus, while participants with data-only were fastest in checking parameters, there was no evidence for participants with just one interpretation to check parameters faster than with three interpretations.

### 3.3 | Error rates

Trials were classified as an error if participants submitted a fault cause other than the one indicated by the deviating parameter. As causes could be identified unambiguously by detecting a
parameter deviation, it is not surprising that the percentage of errors was very low overall. In fact, out of 1885 trials in total, only 26 were errors (1.4%). There neither were main effects of information presentation, $F(2,62) = 1.832$, $p = .169$, $\eta^2_p = 0.056$, and trial type, $F(1,62) = 2.729$, $p = .104$, $\eta^2_p = 0.042$, nor an interaction of both factors, $F(2,62) = 1.832$, $p = .169$, $\eta^2_p = 0.056$ (see Figure 6). We specifically looked at catch trials in the interpretation+data group to check for effects of automation bias. Out of 22 participants, only one submitted an incorrect response in both catch trials and two in one of the catch trials.

### 3.4 Knowledge test performance

Knowledge about fault causes was assessed in the postexperimental knowledge test by asking participants to write down as many causes as possible for a given sensor signal. We calculated the mean number of answers per question as well as the mean number of correct answers, and submitted these data to one-way analysis of variance with information presentation as a between factor. An overview of the data is provided in Figure 7. For the total number of answers, there was an effect of information presentation, $F(2,64) = 4.128$, $p = .021$, $\eta^2_p = 0.118$, indicating that participants who had worked with data-only provided more answers with than those who had worked with interpretation+data (3.8 vs. 2.4, respectively), $p = .018$. In contrast, the interpretations+data group (3.2) did not significantly differ from either of the other groups, both $p > .2$. The analysis of correct answers also revealed an effect of information presentation, $F(2,64) = 3.461$, $p = .038$, $\eta^2_p = 0.100$. Similar to the total number of answers, more correct answers were given in the data-only group than in the interpretation+data group (2.7 vs. 1.8, respectively), $p = .039$, while the interpretations+data group (2.4) did not significantly differ from the others, both $p > .2$.

### 4 DISCUSSION

Fault diagnosis in the processing industry can be supported by assistance systems that infer fault causes from sensor data. The present study investigated how information sampling, the speed and accuracy of performance, and the acquisition of knowledge about fault causes were affected by different strategies of information presentation. Participants either had access to only sensor data, sensor data enhanced with three different interpretations, or only the interpretation favored by the assistance system, with sensor data available on demand. In the latter condition, catch trials were used to investigate whether such highly preprocessed information would bring about negative effects of automation bias. Analyzes of information sampling behavior revealed that compared to only sensor data, interpretations led to more efficient checking of fewer parameters. At the same time, sampling did not significantly differ between the two interpretation conditions. This suggests that while interpretations were helpful in general, participants did not uncritically accept them. In fact, only one participant refrained from cross-checking altogether, and only three others did so in a tiny proportion of trials. This thorough cross-checking also prevented errors, which were low overall and did not significantly increase when the one interpretation was incorrect. Conversely, catch trials increased solution times by a factor of six compared to standard trials, indicating that participants with only one interpretation put large amounts of time into parameter sampling: Although they did not sample more parameters than with only sensor data, they did not seem to have learned how to identify relevant parameters quickly. This lack of learning was mirrored in the knowledge test results, reflecting worse understanding of fault causes after having worked with only one interpretation than after having worked with only sensor data.

#### 4.1 Information about fault causes enables efficient performance

Participants without access to high-level interpretations of sensor data needed most time to identify fault causes. Likewise, other research has demonstrated that with no or little support
during fault diagnosis, people need longer to solve their task (Onnasch et al., 2014; Raaijmakers & Verduyn, 1996). Those participants also checked a large number of parameters, with many of them being irrelevant. These observations corroborate findings from earlier studies indicating that people have difficulties identifying and focusing on critical information during fault diagnosis (Abele, 2017; Adsit & London, 1997; Morris & Rouse, 1985).

In contrast, providing high-level information about fault causes yielded the hypothesized performance benefits, expressed in shorter solution times and fewer parameter checks. This positive association between high-level support and performance is in line with previous findings (Bahner, Hüper et al., 2008; Manzey et al., 2012; Raaijmakers & Verduyn, 1996). Interestingly, however, the highest degree of automation (DOA)—providing just one interpretation—did not lead to a minimization of effort. First, participants with one interpretation checked more than just the parameter indicated by the assistance system in about one fourth of the trials, thus acquiring more information than strictly needed. Second, their rate of parameter checking (i.e., parameters per second) was no lower than with three interpretations. This is in contrast with a previous study which has found that participants working with the highest DOA spent less time checking parameters, although the actual number of parameter checks did not differ from the one in groups working with lower DOA (Manzey et al., 2012). The authors suggested an explanation in terms of reduced attention allocation in case of higher DOA (see also Parasuraman & Manzey, 2010). Thus, it will be an interesting question for future research to identify the factors determining the allocation of attention to automated diagnoses.

Generally, it can be concluded that providing hypotheses about fault causes is useful to support diagnostic actions. Moreover, it is noteworthy that providing three interpretations enabled participants to fulfill their task almost as efficiently as when the correct solution was presented right away. Especially in domains such as the processing industry, where one and the same fault symptom can have many different causes (Müller & Oehm, 2019), providing a list of causal hypotheses seems to be a fruitful approach to guide operators through fault diagnosis.

4.2 | No evidence for automation bias

Participants provided with only one interpretation had ample time for learning to trust the assistance system, because before the first catch trial 22 trials confirmed that the interpretation was correct. Still, the vast majority of participants kept on cross-checking this information. This is in line with findings by Manzey et al. (2014) who showed that when participants have the opportunity to cross-check alarms, they will do so. Moreover, half of our participants with only one interpretation checked the sensor data underlying the interpretation in more than 90% of the trials, further strengthening the assumption that they did not simply accept interpretations but actively ensured that they would be able to evaluate them. Thus, it can be concluded that overall, participants did not exhibit signs of automation bias.

While this is in line with our hypotheses and a previous study (Müller, Ullrich, et al., 2019), it contrasts with a large body of literature on automation bias (e.g., Layton, Smith, & Mc Coy, 1994; Manzey et al., 2012; Mosier et al., 1998; Parasuraman & Manzey, 2010; Skitka, Mosier, & Burdick, 1999). Several features of the present study might have contributed to the divergence in findings. First, the fault identification task was rather easy and the effort to verify the automated suggestion was low. Each fault had only one cause indicated by exactly one deviating parameter, and interpretations directly corresponded to individual parameters. For instance, when the interpretation stated that chocolate bars were too wide, the participant had to check the parameter "width of chocolate bars" and when it deviated, they knew this was the cause. Accordingly, incorrect interpretations were easy to detect. In contrast, participants in studies by Manzey et al. (2012) and Bahner et al. (2008) had to check two to four parameters to identify automation failures. Additionally, these studies used two secondary tasks to increase workload and time pressure, which are known to increase the prevalence of insufficient information sampling and automation bias (Molloy & Parasuraman, 1996; Parasuraman et al., 1993). Second, automation bias is usually found with high DOA (Onnasch et al., 2014), for instance when recommending a particular action. In the levels and stages framework of automation (Parasuraman et al., 2000), the assistance system we used supported information analysis (i.e., turning low-level sensor data into high-level interpretations), which corresponds to a relatively low DOA. The system did choose particular interpretations instead of providing all possibilities and thus biased decision selection, but it did not directly recommend actions. In contrast, previous studies on automation bias in fault diagnosis automated action selection (Bahner, Hüper et al., 2008) or action implementation (Manzey et al., 2012). A third reason for the absence of automation bias might relate to the type of task. Fault diagnosis requires information analysis (Abele, 2017), whereas automation bias describes the replacement of thorough information analysis with automated heuristics (Mosier et al., 1998). Accordingly, fault diagnosis might not be as prone to automation bias as other tasks. Instead, diagnosticians are likely to use an automated support system as guidance for where to start searching, instead of blindly following its recommendations (Müller, Ullrich, et al., 2019). In the present study, this assumption was supported by the finding that participants with one interpretation almost always checked the parameter indicated by the interpretation first.

In conclusion, there are several plausible explanations for why automation bias was not found in the present study, and future studies will need to tease apart the impacts of workload, DOA, and task type on the occurrence of automation bias in the context of dealing with sensor data and explanations of algorithms.
4.3 | Impaired learning with only one interpretation

Although no automation bias was found, other performance measures declined substantially in case of incorrect interpretations. This is in line with other research showing strong performance decrements when automation fails (Manzey et al., 2012; Onnasch et al., 2014; Parasuraman & Manzey, 2010). In the present study, catch trials brought parameter sampling effort to a level similar to the group that never had interpretations available. However, to check the same amount of information, participants with one interpretation needed three times longer and were at the level of the former group’s very first trial. This performance drop was much stronger than we had expected. It also suggests that people who always have to do without support can learn how to use the available resources, while people who are used to automated support cannot (Bainbridge, 1983). Additional analyses not reported here indicate that practice effects were comparable across groups. However, apparently they were highly specific to the task that had to be solved in standard trials: While participants with one interpretation learned to quickly find and check the parameter corresponding to their interpretation, they do not seem to have learned how to interpret sensor data. Notably, this was the case although half of the participants did check the sensor data in more than 90% of the trials. Those latter data. Notably, this was the case although half of the participants did not check the sensor data in more than 90% of the trials. Those latter participants’ solution times in catch trials were faster than those of participants who did not check the sensor data as often, but still 2.5 slower than those of participants for whom sensor data was the only information source. Apparently, even when having additional low-level information available, the fact that participants never had to actively use it prevented them from learning how to interpret it quickly. In consequence, when their automated interpretation was incorrect and they had to rely on sensor data, they started at a level of performance where participants with only data had been at the start of the experiment. This is quite striking as it suggests that no practice benefits whatsoever had emerged from already having used the parameter screen in 22 previous trials. Future studies are needed to scrutinize the exact locus of practice effects. For instance, do people without interpretations become more efficient in navigating the parameter screen or in translating data into interpretations? In the present study, the latter option is supported by the results of knowledge test performance.

Knowledge test scores revealed a higher number of answers and correct answers for participants with only sensor data than for participants with one interpretation. Initially, this finding was quite surprising and contrary to our hypotheses. Why did the group that had received no information about fault causes perform best when asked to name possible fault causes? An obvious explanation might be that they had checked more parameters (i.e., three to nine times more than the groups with interpretations). However, this explanation seems unlikely, because all parameters not representing the correct cause in a given trial actually provided negative feedback, and from their values it was not distinguishable whether they could have been an alternative cause. Therefore, a more likely reason is that participants with only sensor data had to generate their own interpretations of sensor data, while participants with automated interpretations did not. This is in line with research on instructional explanations. A review by Wittwer and Renkl (2008) concludes that these explanations should not substitute learners’ own active knowledge construction. Providing instructional explanations can force learners into a passive role, preventing them from generating their own explanations and thus impairing the acquisition of conceptual knowledge (Richey & Nokes-Malach, 2013; Schworm & Renkl, 2006). Instead, prompting learners to generate self-explanations can improve learning and understanding (Chi, De Leeuw, Chiu, & LaVancher, 1994), which might explain the superior fault knowledge of participants who did not receive interpretations in the present study. Taken together, practicing how to interpret sensor data seems to be vital precondition both for managing situations of automation failure and acquiring explicit knowledge about possible fault causes.

4.4 | Practical implications

From a practitioner’s point of view, providing operators with only one interpretation of sensor data might seem like a desirable option: In routine situations, it helps them to quickly identify fault causes, while in case of automation failure they still manage to identify the fault cause correctly, albeit needing more time. Therefore, practitioners might accept the performance costs in failure situations, assuming that they are far outweighed by the performance benefits in routine situations. However, this conclusion seems less valid when taking domain characteristics into account. In complex industrial systems like the processing industry, several simplifications made in the present experiment do not hold, and thus problems may arise when providing just one interpretation. First, faults in processing plants usually result from interactions of multiple influences, and often several causes contribute to a single symptom (Müller & Oehm, 2019). Second, several of these influences cannot be measured at all, not reliably, or not at the right time, and for many of them appropriate process models are lacking (Majschak, 2014). Therefore, algorithms inferring fault causes from sensor data are based on partly incomplete and unreliable information (Cai et al., 2016; Olsson, Funk, & Xiong, 2004). In consequence, in real-world processing plants it can neither be assumed that automated interpretations are exhaustive nor that they are correct in almost all situations.

A second, more human-centered note of caution stems from the finding that providing just one interpretation seems to push operators into a passive role, reducing their competence and obstructing their own knowledge construction activities. Keeping operators in an active role during fault diagnosis is recognized as a major challenge when supporting them with sensor data (Venkatasubramanian, Rengaswamy, Kavuri et al., 2003), and providing only one interpretation seems counterproductive with regard to this goal. Moreover, even when they do cross-check this one interpretation, it is likely to prevent them from looking for alternative fault causes once they have
confirmed a deviation. Such behavior was appropriate in the present experiments where only one deviation was present in each scenario. However, in practice it can be detrimental, especially given that diagnosticists generally have a tendency to generate too few hypotheses too quickly (Mehle, 1982; Patrick et al., 1999), do not consider multiple overlapping faults (Patrick et al., 1999), are prone to confirmation bias (Nickerson, 1998) and tend to terminate their search once they have found a deviation (Berbaum et al., 1990).

Instead, it seems like a fruitful approach to provide a list of possible interpretations of sensor data and thus a diverse set of hypotheses about fault causes. Operators can consider these interpretations in relation to context factors that might not be available to the algorithm, and the mere presence of different interpretations is likely to remind them that multiple fault causes should be considered. Moreover, based on the present results providing different interpretations is not likely to have serious costs: Participants having access to three interpretations identified fault causes almost as efficiently as those with only one interpretation, while their knowledge acquisition did not differ from those who had to solve the task without interpretations.

4.5 | Limitations and questions for future research

Limitations of the present study relate to typical constraints of controlled lab experiments. For the sake of internal validity, we had to remove much of the complexity characteristic for real-world processing plants (Müller & Oehm, 2019). The faults scenarios and tasks were quite simple, assuming singular causes and neglecting parameter interactions. This simplicity also produced ceiling effects in that errors were almost absent. Additionally, external factors such as time pressure or economic constraints is likely to influence operator behavior in real production contexts. Therefore, future research should investigate the interaction between sensor data interpretations, faults of different complexity, and external constraints. Similarly, our participants neither were machine operators nor had any prior experience with packaging machines or the processing industry. Therefore, it is not surprising that high-level information facilitated fault diagnosis significantly, and from the present results we cannot conclude how skilled operators would deal with this kind of information. In the processing industry, operators often acquire their knowledge based on experience (Allais, Perrot, Curt, & Trystram, 2007; Müller & Oehm, 2019). Thus, data-based information provided by an assistance system might actually be a helpful extension of operators’ capabilities during fault diagnosis, enabling them to confirm or disconfirm their spontaneous hypotheses about fault causes. Investigating the influence of expertise on the usage of low- versus high-level assistance provides an interesting challenge for future studies.

A limitation of the experimental design is that we did not manipulate the correctness of interpretations provided in the data +3interpretations group. Thus, it is not possible to draw conclusions about the effects of automation failure on diagnostic performance when several options are provided. Previous work suggests that such situations will not produce automation bias (Müller, Ullrich, et al., 2019), but we do not know whether or to what degree negative performance effects (i.e., increases of solution time, sampling of irrelevant parameters) would emerge in this situation. Alternatively, being made aware of the fact that several causes can underlie the same sensor data might lead to a more sensitive selection of considered information sources, mitigating performance decrements in case of automation failure.

A third limitation concerns the postexperimental knowledge test, which asked about fault causes in a manner that was almost identical to the way sensor data were presented during the experiment. This might have facilitated the transfer of knowledge to the test but does not ensure transfer to more different situations as required in real-world fault diagnosis. For instance, transfer can also mean that people generalize their use of explanation schemes (Engle, 2006). We cannot conclude whether participants have learned to directly memorize possible causes for particular sensor data or have become more proficient in using sensor data to restrict the problem space. This limits the generalization of the present findings as far transfer based on structural similarities rather than surface features is challenging (Gentner, Rattermann, & Forbus, 1993; Holyoak & Koh, 1987).

A final set of limitations relates to the presentation of sensor data and their interpretations. Participants had access to textual interpretations of individual sensor data. This enabled them to identify fault causes efficiently in our simplified setting. However, besides the increased risk of operator passivity resulting from instructional explanation as discussed above, another issue emerges: Simply providing one cause or a list of causes does not support operators’ understanding of interactions between different causes. Knowing and understanding interactions within a system is important for successful diagnosis in complex systems (Abele, 2017; Jonassen & Hung, 2006). Although it was not the goal of the present study to investigate the understanding of system interactions, it raises the question how the results of algorithms can be presented to foster such understanding and active thinking about causal relations.

One approach is to use more elaborate explanations that provide answers about how and why a phenomenon occurs (Roth-Berghofer, 2004). The interpretations used in the present study did not consider instructional principles of explanations (Wittwer & Renkl, 2008) but simply provided one or three possible solutions right away. This limits their usefulness to convey system knowledge. More elaborated explanations could inform operators about how an algorithm derived its interpretation and what information it considered to do so. Explaining the rationale of algorithms is referred to as explainability (Miller, 2019; Ribeiro, Singh, & Guestrin, 2016), which is considered a desirable characteristic of algorithms (Guidotti et al., 2019; Muggleton, Schmid, Zeller, Tamaddoni-Nezhad, & Besold, 2018; Venkatasubramanian, Rengaswamy, Yin et al., 2003). Most research on explainability aims at user compliance and trust (Lim, Dey, & Avrahami, 2009; Ribeiro et al., 2016; Samek, Wiegand, & Müller, 2017) but rarely considers how the design of explanations influences users’ understanding of a system, perhaps because it does
not usually draw on research from the social sciences (Miller, 2019). However, using algorithms to convey system knowledge is likely to improve performance as such knowledge is a major determinant of diagnostic success (Abele, Walker, & Nickolaus, 2014; Nickolaus, Abele, Gschwendtner, Nitzschke, & Greiff, 2012). Besides textual explanations, another approach to making systems understandable is to use visualizations of constraints and relations within a technical system (Bennett, 2017; Borst, Flach, & Ellerbroek, 2015; Vicente & Rasmussen, 1992). These concepts can be transferred to the processing and packaging industry (Jaster & Müller, 2019; Koshman, Blümel, Bönsel, & Müller, 2019). However, this research has mainly focused on the visualization of relations between process parameters in supervisory control interfaces, and it will be interesting to extend it to the communication of outcomes from diagnostic algorithms.

5 | CONCLUSION

Overall, the present study emphasizes the importance of making information available to operators by demonstrating how high-level information about the interpretation of sensor data increased diagnostic performance. Such performance increases pertain to process measures (i.e., information sampling behavior) as well as direct performance measures (i.e., time needed to identify fault causes). However, it also cautions against possible side effects of telling operators what to do, which can impair performance during automation failure and hamper active knowledge construction. Providing several possible interpretations seems like a fruitful approach, and future interdisciplinary research is needed to investigate how such information can be explained and visualized to foster an understanding of causal relations within technical systems.

ACKNOWLEDGMENTS

The authors want to thank Jan Paap for support in data collection, and the KoMMDia team for valuable discussions of the ideas underlying the present study. This work was funded by a BMBF grant for the project KoMMDia (02K16C07).

ORCID

Romy Müller  http://orcid.org/0000-0003-4750-7952
Rica Bönsel  http://orcid.org/0000-0001-6165-3521

REFERENCES

Abele, S. (2017). Diagnostic problem-solving process in professional contexts: Theory and empirical investigation in the context of car mechatronics using computer-generated log-files. Vocations and Learning, 11(1), 133–159. https://doi.org/10.1007/s12186-017-9183-x
Abele, S., & von Davier, M. (2019). Handbook diagnostic classification models. In M. van Davier & S. Abele (Eds.), CDMs in vocational education: Assessment and usage of diagnostic problem-solving strategies in car mechatronics. Cham, Switzerland: Springer.
Abele, S., Walker, F., & Nickolaus, R. (2014). Zeitökonomische und reliable Diagnostik beruflicher Problemlösekompetenzen bei Auszubildenden zum Kfz-Mechatroniker. Zeitschrift für Pädagogische Psychologie, 28, 167–179. https://doi.org/10.1024/1010-0652/a000138
Adsit, D. J., & London, M. (1997). Effects of hypothesis generation on hypothesis testing in rule-discovery tasks. The Journal of General Psychology, 124(1), 19–34. https://doi.org/10.1080/00221309709595505
Allais, I., Perrot, N., Curt, C., & Trystram, G. (2007). Modelling the operator know-how to control sensory quality in traditional processes. Journal of Food Engineering, 83(2), 156–166. https://doi.org/10.1016/j.jfoodeng.2007.02.016
Arocha, J. F., Patel, V. L., & Patel, Y. C. (1993). Hypothesis generation and the coordination of theory and evidence in novice diagnostic reasoning. Medical Decision Making, 13(3), 198–211. https://doi.org/10.1177/0727989X9301300305
Bahner, J. E., Elefpandt, M. F., & Manzey, D. (2008). Misuse of diagnostic aids in process control: The effects of automation misses on complacency and automation bias. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (pp. 1330–1334), Los Angeles, CA: SAGE Publications.
Bahner, J. E., Hüper, A. D., & Manzey, D. (2008). Misuse of automated decision aids: Complacency, automation bias and the impact of training experience. International Journal of Human-Computer Studies, 66(9), 688–699. https://doi.org/10.1016/j.ijhcs.2008.06.001
Bainbridge, L. (1983). Ironies of automation. Automatica, 19(6), 775–779.
Bennett, K. B. (2017). Ecological interface design and system safety: One facet of Rasmussen’s legacy. Applied Ergonomics, 59(Part B), 625–636. https://doi.org/10.1016/j.apergo.2015.08.001
Berbaum, S. K., Franken, E. A., Rooholami, S. A., Dorfman, D. D., Kathol, M. H., Barloon, T. J., ... Montgomery, W. J. (1990). Satisfaction of search in diagnostic radiology. Investigative Radiology, 25(2), 133–140.
Bereiter, S. R., & Miller, S. M. (1989). A field-based study of troubleshooting in computer-controlled manufacturing systems. IEEE Transactions on Systems, Man and Cybernetics, 19(2), 205–219. https://doi.org/10.1109/21.31027
Bleisch, G., Majschak, J.-P., & Weiß, U. (2011). Verpackungstechnische Prozesse: Lebensmittel-, Pharma- und Chemieindustrie (1. Aufl.). Hamburg: Behr.
Borlea, I., Buta, A., Dusa, V., & Lustrea, B. (2005). DIASE - Expert System Fault Diagnosis for Timisoara 220 kV Substation. In Proceedings of the EUROCON 2005 - The International Conference on “Computer as a Tool” (pp. 221–224), Belgrade, Serbia: IEEE.
Borst, C., Flach, J. M., & Ellerbroek, J. (2015). Beyond ecological interface design: Lessons from concerns and misconceptions. IEEE Transactions on Human-Machine Systems, 45(2), 164–175. https://doi.org/10.1109/THMS.2014.2364984
Cai, B., Liu, H., & Xie, M. (2016). A real-time fault diagnosis methodology of complex systems using object-oriented Bayesian networks. Mechanical Systems and Signal Processing, 80, 31–44. https://doi.org/10.1016/j.ymssp.2016.04.019
Chi, M. T. H., De Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. Cognitive science, 18(3), 439–477. https://doi.org/10.1016/0364-0213(94)90016-7
Culnan, M. J. (1983). Environmental scanning: The effects of task complexity and source accessibility on information gathering behavior. Decision Sciences, 14(2), 194–206. https://doi.org/10.1111/j.1540-5915.1983.tb00180.x
Engle, R. A. (2006). Framing interactions to foster generative learning: A situative explanation of transfer in a community of learners classroom. The Journal of the Learning Sciences, 15(4), 451–498. https://doi.org/10.1207/s15327809jls1504_2
Fidel, R., & Green, M. (2004). The many faces of accessibility: Engineers’ perception of information sources. Information Processing & Management, 40(3), 563–581. https://doi.org/10.1016/S0306-4573(03)00003-7
Fiedler, K., & Kutzner, F. (2015). Information sampling and reasoning biases: Implications for research in judgment and decision making. In G. Keren & G. Wu (Eds.), The Wiley Blackwell Handbook of Judgment and Decision Making (pp. 380–403). New York, NY: Wiley.
Parasuraman, R., Molloy, R., & Singh, I. L. (1993). Performance consequences of automation-induced 'complacency'. The International Journal of Aviation Psychology, 3(1), 1–23. https://doi.org/10.1207/s15327108ijap0301_1

Parasuraman, R., Sheridan, T. B., & Wickens, C. D. (2000). A model for types and levels of human interactions with automation. Transactions on Systems, Man, and Cybernetics, 30(3), 286–297. https://doi.org/10.1109/3468.844354

Patrick, J., Gregov, A., Halliday, P., Handley, J., & O'Reilly, S. (1999). Analysing operators' diagnostic reasoning during multiple events. Ergonomics, 42(3), 493–515. https://doi.org/10.1080/001401399185603

Perron, C. (1984). Normal accidents: Living with high-risk technologies. New York, NY: Basic Books.

Platzer, C., & Bröder, A. (2012). Most people do not ignore salient invalid cues in memory-based decisions. Psychonomic Bulletin & Review, 19(4), 654–661. https://doi.org/10.3758/s13423-012-0248-4

Raaijmakers, J. G., & Verduyn, W. W. (1996). Individual differences and the effects of an information aid in performance of a fault diagnosis task. Ergonomics, 39(7), 966–979. https://doi.org/10.1080/00140139608964517

Rasmussen, J., & Goodstein, L. P. (1987). Decision support in supervisory control of high-risk industrial systems. Automatica, 23(5), 663–671. https://doi.org/10.1016/0005-1098(87)90064-1

Rauber, T. W., Barata, M. M., & Steiger-Garcao, A. S. (1993). A toolbox for analysis and visualization of sensor data in supervision. In Proceedings of the TOOLDIAG ’93 – International Conference on Fault Diagnosis (pp. 3–7). Toulouse, France.

Rebischek, F. G., Krems, J. F., & Jahn, G. (2015). Memory activation of multiple hypotheses in sequential diagnostic reasoning. Journal of Cognitive Psychology, 27(6), 780–796. https://doi.org/10.1080/20445911.2015.1026825

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1135–1144), San Francisco, CA: ACM.

Richey, J. E., & Nokes-Malach, T. J. (2013). How much is too much? Learning and motivation effects of adding instructional explanations to worked examples. Learning and Instruction, 25, 104–124. https://doi.org/10.1016/j.learninstruc.2012.11.006

Roth-Berghofer, T. R. (2004). Explanations and case-based reasoning: Foundational issues. In Proceedings of the European Conference on Case-Based Reasoning (pp. 389–403), Berlin, Heidelberg: Springer.

Samek, W., Wiegand, T., & Müller, K.-R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. ArXiv Preprint arXiv, 1708.08296.

Schult, A., Beck, E., & Majschak, J.-P. (2015). Steigerung der Effizienz von Verarbeitungs- und Verpackungsanlagen durch Wirkungsgradanalysen. Pharrma+Food, 18(7), 66–68.

Schworm, S., & Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. Computers & Education, 46(4), 426–445. https://doi.org/10.1016/j.compedu.2004.08.011

Scott, D. M. (2008). Industrial process sensors. Boca Raton, FL: CRC Press.

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? International Journal of Human-Computer Studies, 51(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Thomas, R. P., Dougherty, M. R., Sprenger, A. M., & Harbison, J. I. (2008). Diagnostic hypothesis generation and human judgment. Psychological Review, 115(1), 155–185. https://doi.org/10.1037/0033-295X.115.1.155

Tsarouhas, P. (2012). Reliability, availability and maintainability analysis in food production lines: A review. International Journal of Food Science and Technology, 47(11), 2243–2251. https://doi.org/10.1111/j.1365-2621.2012.03073.x

Venkatasubramanian, V., Rengaswamy, R., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis: Part II: Qualitative models and search strategies. Computers & Chemical Engineering, 27(3), 313–326. https://doi.org/10.1016/S0098-1354(02)00161-8

Venkatasubramanian, V., Rengaswamy, R., Kavuri, S. N., & Yin, K. (2003). A review of process fault detection and diagnosis: Part III: Process history based methods. Computers & Chemical Engineering, 27(3), 327–346. https://doi.org/10.1016/S0098-1354(02)00162-X

Venkatasubramanian, V., Rengaswamy, R., Yin, K., & Kavuri, S. N. (2003). A review of process fault detection and diagnosis: Part I: Quantitative model-based methods. Computers & Chemical Engineering, 27(3), 293–311. https://doi.org/10.1016/S0098-1354(02)00160-6

Verplanken, B., Aarts, H., & Van Kinpenberg, A. (1997). Habit, information acquisition, and the process of making travel mode choices. European Journal of Social Psychology, 27(5), 539–560. https://doi.org/10.1002/ (SICI)1099-0992(199709)27:5<539::AID-EJSP821>3.0.CO;2-A

Vicente, K. J., & Rasmussen, J. (1992). Ecological interface design: Theoretical foundations. IEEE Transactions on Systems, Man and Cybernetics, 22, 1–18.

Wickens, C. D., Li, H., Santamaria, A., Sebok, A., & Sarter, N. B. (2010). Stages and levels of automation: An integrated meta-analysis. In Proceedings of the Human Factors and Ergonomics Society Annual Meeting (pp. 389–393), Los Angeles, CA: Sage Publications.

Wittwer, J., & Renkl, A. (2008). Why instructional explanations often do not work: A framework for understanding the effectiveness of instructional explanations. Educational Psychologist, 43(1), 49–64. https://doi.org/10.1080/00461520701756420

How to cite this article: Müller R, Gögel C, Bönsel R. Data or interpretations: Impacts of information presentation strategies on diagnostic processes. Hum. Factors Man. 2020;30:266–281. https://doi.org/10.1002/hfm.20838