ECAI 2023
K. Gal et al. (Eds.)

© 2023 The Authors.
This article is published online with Open Access by IOS Press and distributed under the terms
of the Creative Commons Attribution Non-Commercial License 4.0 (CC BY-NC 4.0).

doi:10.3233/FAIA230644

Creating Explainable Dynamic Checklists via Machine Learning to Ensure Decent Working Environment for All: A Field Study with Labour Inspections

Eirik Lund Flogard [a;*], **Ole Jakob Mengshoel**[b], **Ole Magnus Theisen**[a] **and Kerstin Bach**[b]

[a]Norwegian Labour Inspection Authority
[b]Norwegian University of Science and Technology

Abstract. To address poor working conditions and promote United Nations' sustainable development goal 8.8, "protect labour rights and promote safe working environments for all workers [...]", government agencies around the world conduct labour inspections. To carry out these inspections, inspectors traditionally use paper-based checklists as a means to survey individual organisations for working environment violations. Currently, these checklists are created by domain experts, but recent research indicates that machine learning (ML) could be used to generate dynamic checklists to increase inspection efficiency. A drawback with the dynamic checklists is that they are complex and could be difficult to understand for inspectors. They have also never been field-tested. In this paper, we therefore propose user-oriented explanation methods for Context-aware Bayesian Case-Based Reasoning (CBCBR), which is the current state-of-art ML method for generating dynamic checklists. We also introduce a prototype of CBCBR and present a field study where we test it in real-world labour inspections. The results from the study indicate that using the explainable dynamic checklists increases the efficiency of the labour inspections, and inspectors also report that they find the checklists useful. The results also suggest that current ML evaluation methods, where model prediction performance is evaluated on existing data, may not fully reflect the real-world field performance of checklists.

1 Introduction

Labour inspections are conducted by government agencies around the world to address poor working conditions and promote United Nations' sustainable development goal (SDG) 8.8 "to protect labour rights and promote safe working environments for all workers". The inspections are carried out in workplaces (organisations) on a large scale to enforce national and international labour laws and standards, pursuant to the International Labour Organization's (ILO) Labour Inspection Convention (1947). Despite the inspection efforts, there are still 1.9 million registered deaths worldwide annually attributable to occupational health and safety risks [29]. Increasing the efficiency of inspections is therefore important. To carry out these inspections, inspectors often use checklists to survey individual organisations for non-compliance to health, safety, and environment (HSE) regulations [8, 19]. A labour inspection checklist consists of a non-fixed-sized subset of K out of N items, where each item has a binary response (non-compliant/compliant) and corresponds to a specific health and safety regulation. Each inspected organisation may be subjected to hundreds of different regulations [13], but a checklist ideally contains between 5 and 30 items. Creating such checklists manually is difficult since checklists are situation-dependent and labour inspections can be executed in many different ways depending on the context [17, 5]. This means that the contents of the checklists need to vary between individual inspected organisations. Traditional static, paper-based checklists are still often too long, making it difficult to differentiate between critical and less important tasks for their users [21].

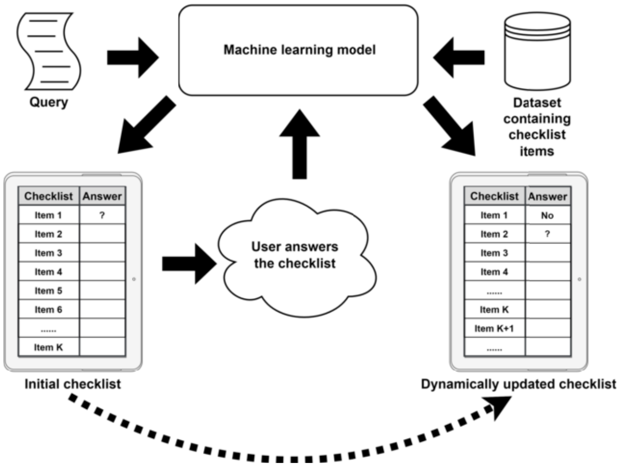


Figure 1. An illustration of how dynamic checklists work. First, the initial checklist to the left is created via a query specified by the user. After the user answers the checklist, it is then dynamically updated by the ML model. The updated checklist is shown on the right side.

Instead of relying on traditional checklists, it is possible to use ML to generate dynamic checklists. Figure 1 shows an overview of a dynamic checklist, where an ML model is given an inspection target (workplace) as input. The model then creates an initial checklist containing a set of K out of N possible items (small K, large N). Based on how the user answers the checklist, it is dynamically updated with additional items to make it more contextually relevant to the situation it is being used in. To our knowledge, the state-of-the-art method for generating dynamic checklists is Context-aware Bayesian Case-Based Reasoning (CBCBR) [15]. CBCBR aims to create checklists

[*] Corresponding Author. Email: eirik.flogard@arbeidstilsynet.no

that maximize the number of violations found during inspections, by generating and dynamically updating checklists specifically for each inspection target. It is assumed that the dynamic checklists can increase the detection and rectification of working environment violations in the inspections compared to traditional paper-based checklists, thereby increasing efficiency [15]. The main purpose of this paper is to test this assumption by conducting an empirical field study, to determine if ML-based checklists are indeed superior for real-world inspections.

Motivation. It is hard to tell how effective current ML-generated checklists are, since they have never been tested in real-world environments [32, 23]. Flogard et al. propose a cross-validation approach that shows promising results for the dynamic checklists in labour inspections. However, lacking ground truth cases, the approach is essentially a simulation that is mostly based on labels that are generated from existing data [14, 15]. Since the approach relies on existing data, it also does not account for real-world factors that could impact labour inspection performance, such as intervention effects from replacing the current domain expert-designed checklists [31]. This is potentially problematic as inspections are complex tasks, and the success of using checklists could depend on many factors, such as implementation details or how users interact with them [5, 36]. Another problem is that CBCBR currently lacks explanation methods. Dynamic checklists are complex constructs, rendering it difficult for inspectors to understand the dynamic changes to their checklists during inspections. If the dynamic checklists are not understood or justified, they may be difficult to use, undermining any advantages they may have on task performance [5]. Moreover, forthcoming EU regulations will require a certain level of explanation in ML and related technologies [7].

Contributions. To address the problems mentioned above, our scientific contributions in this paper are as follows:

(1) Technical: We propose methods for explaining the content of dynamic checklists to their end-users (inspectors), focusing on justification and transparency as explanation goals [35]. We also developed a prototype based on the state-of-the-art method for generating checklists (CBCBR), implementing the explanation methods.[1]

(2) Social: As far as we know, ML-based checklists remain untested in the field, let alone implemented or adopted. Collaborating with the Norwegian Labour Inspection Authority (NLIA), we conducted a field study, testing the prototype in dynamic real-world environments with seven inspectors carrying out 69 valid inspections across various industries. Both qualitative and quantitative results are presented and compared to inspections conducted by the same inspectors, using ordinary static checklists created by domain experts. The results show that dynamic checklists increase the efficiency and number of violations being addressed in labour inspections. This insight is essential in order to determine whether the adoption of the dynamic checklists is worth the investment, as most labour inspection authorities have limited resources [37].

(3) Analytical: Our analyses of the results from the study show significant discrepancies between field performance and existing cross-validation performance estimates of dynamic checklists. Current ML evaluation practices may therefore be insufficient for estimating field performance of checklists. This insight could have implications for research in other domains where ML or AI is used to create checklists, such as medicine [40, 22, 18].

Social Impact. Due to the results from our field study, the Norwegian Labour Inspection Authority (NLIA) plans to adopt the dynamic checklists into their inspections. We believe that our work could inspire labour inspection authorities in other countries to do the same. A widespread adoption of dynamic checklists could increase levels of compliance with working environment laws and labour rights in society and reduce both long and short-term injuries (SDG indicators 8.8.1 and 8.8.2), as inspections become more efficient in addressing and reducing violations in workplaces.

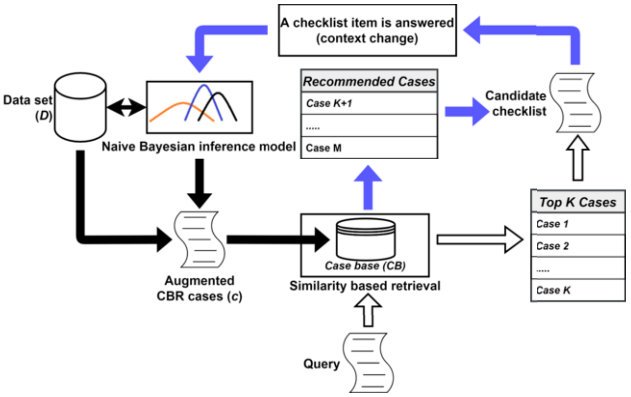


Figure 2. The figure [15] shows an overview of CBCBR. The black arrows show how the case base is created or updated. The white arrows show the creation of a candidate checklist. The checklist is dynamically updated via the blue (and black) arrows, starting from the candidate checklist on the right-hand side.

2 Related Work

Dynamic Checklists. Digital dynamic checklists have also been proposed to deal with context changes in medical applications, such as emergency care or surgeries. However, these checklists are currently created via rule [11] or process-based models [6] that are not based on ML, requiring manual construction and maintenance that limits the models' complexity and nuance [15, 21]. Nevertheless, some of these have been successfully tested in medical trials. De Bie et al. show that dynamic checklists improved user compliance, compared to traditional paper checklists for intensive care units in a trial [12]. Kulp et al. propose a dynamic digital checklist for trauma resuscitation modeled as an iterative process via user interviews, and test it in a trial [21]. The results are promising, but the checklists' impact on task execution and performance in many medical situations turned out differently than expected, underscoring the complexity of checklists and the importance of analyzing and testing them in the field before adoption. This is also a motivation for our field study, especially considering that ML-based checklists have never been field-tested before.

Machine Learning Methods for Creating Checklists. Research on ML for creating checklists is currently limited. Besides CBCBR and BCBR proposed by Flogard et al. [14, 15], there are other methods for checklist creation [40, 22, 4]. These are unsuitable for creating dynamic checklists or checklists for labour inspections [15]. As far as we know, CBCBR is currently the only readily available ML method for creating dynamic checklists for labour inspections. CBCBR is a hybrid method that uses a Naive Bayesian inference (NBI) model to construct features for cases used in Case-Based Reasoning (CBR) [1]. CBCBR creates a dynamic checklist for a given organisation targeted for an inspection by retrieving and reusing past cases with checklist items used in similar organisations, that also

[1] The source code for the prototype can be found at: https://github.com/ntnu-ai-lab/cbcbr-prototype.

have high estimated probabilities for non-compliance. CBCBR, see Figure 2, operates in two phases. (1) In the first phase, an initial checklist is created. A naive Bayesian inference (NBI) model is used to generate probability estimates for non-compliance (θ^{be}), based on empirical distributions from the dataset $\mathcal{D}$. The probability estimates are added to dataset instances $\mathbf{d}_j \in \mathcal{D}$ as new features, to create new augmented CBR cases $\mathbf{c}_j$ for a case base $\mathcal{CB}$. Similarity based retrieval is then used to create an initial candidate checklist by retrieving K CBR cases with unique items, using a query $\mathbf{q} = (x^{cnd}, \theta, \kappa)$ that contains feature values for the target organisation (x^{cnd}), a fixed target value (θ) for the probability estimate embedded in each of the CBR cases and a target value (κ) for the number of observed instances that are used to calculate the probability estimate. (2) The second phase consists of dynamic updates to the candidate checklist via the case-base. After the user answers a checklist item, the NBI model updates the CBR cases $\mathbf{c}_j \in \mathcal{CB}$ with new posterior probability estimates for non-compliance. CBCBR then retrieves any additional cases that have sufficiently increased estimates, which are appended to the checklist as a dynamic update. Depending on the setup, this phase is repeated after a certain number of checklist items are answered. A complete formal definition of dynamic checklists and a more detailed description and analysis of the CBCBR framework is given by Flogard et al. [15].

Explanations for Dynamic Checklists. As far as we know, no one has proposed any method that offers user-oriented explanations of dynamic checklists. However, CBCBR is a good starting point for new explanation approaches since it is based on two transparent methods: CBR and parameter estimates from empirical distributions (NBI) [1, 10]. There are many examples of CBR systems being used to provide explanations, often as post-hoc or in twin configurations with black box systems [20, 9]. Many methods based on model agnostic approaches also exist, such as LIME or SHAP [34, 27]. However, most of the current explanation methods address other explanation goals than ours and are not good starting points for explaining the content of dynamic checklists, and are therefore not considered within the scope of our work. Explanations should generally be goal-oriented and serve a specific target audience [27, 35]. Thus, we propose approaches for providing user-oriented explanations with justification and transparency goals in mind, both for initially created checklists (before inspection starts) and for any dynamic updates to the checklists that are made during inspections.

3 Explanation Methods for Dynamic Checklists

To reach the explanation goals mentioned earlier, we propose two approaches. Both approaches are based on showing traces of model logic to the users [35].

Showing Estimated Probabilities for Non-Compliance. The first approach is to show CBCBR's estimated probability of finding non-compliance on each checklist item to the users. The probability estimates are calculated via the NBI model that was proposed by Flogard et al. [15], but these have not been used for explanations in previous work. As each item on a checklist has its own estimate that depends on the inspection target, the purpose of the explanations is to provide prediction transparency and justify the use of the items. Since the estimates are based on sufficient statistics using empirical distributions [10], they should in theory reflect the probabilities observed in the real world if unbiased data is used and all prior parameters are known. This property should ensure that the estimates are as consistent as possible with the real world, within the limitation of the dataset being used, which is necessary to promote long-term trust [30]. Probability estimates are also an intuitive way to communicate uncertainty [10].

Showing the Most Important Answer for a Dynamic Update. The purpose of this second approach is to make the users aware of why additional items are dynamically added to their checklist (justification), and how these are related to the answered part of the checklist (transparency), which could promote trustworthiness [39, 2]. A formula for finding the most important answer is derived as follows: Let x be a target organisation for an ongoing inspection and $\mathbf{y}^{cnd}$ be a candidate checklist that a user interacts with during the inspection. Let's assume that an item $\hat{e} \notin \mathbf{y}^{cnd}$ is considered as a candidate to be dynamically added to the checklist. Let $(e_i, l_i) \in \mathbf{y}^{cnd}$ be pairs of existing items and given answers in the checklist, respectively, with the position in $\mathbf{y}^{cnd}$ indexed by i. The probability for finding non-compliance ($L = 1$) for any candidate item $\hat{e}$, given x and every pair (e_i, l_i), can be estimated via [15]:

$$\theta^{be}(L = 1|x, \hat{e}, \mathbf{y}^{cnd}) = \frac{\beta_{L=1|x,\hat{e}} + \sum_{(l_i, e_i \in \mathbf{y}^{cnd})} p(1, x, \hat{e}, e_i, l_i)}{\sum_{\hat{l}=0}^{1} \beta_{L=\hat{l}|x,\hat{e}} + \sum_{(l_i, e_i \in \mathbf{y}^{cnd})} p(\hat{l}, x, \hat{e}, e_i, l_i)}, \quad (1)$$

which CBCBR relies on to dynamically update the checklists and is the mean of a posterior beta distribution (see Flogard et al. [15] for more details). Given this information, we seek to find the index of the pair (e_i, l_i) that has the most impact on an $\hat{e}$ being selected for a dynamic update to the checklist. The index i can be found by altering Equation 1 to depend on only single pairs (e_i, l_i) as follows:

$$\arg\max_i \theta^{be}(L = 1|x, \hat{e}, e_i, l_i) = \arg\max_i \frac{\beta_{L=1|x,\hat{e}} + p(1, x, \hat{e}, e_i, l_i)}{\sum_{\hat{l}=0}^{1} \beta_{L=\hat{l}|x,\hat{e}} + p(\hat{l}, x, \hat{e}, e_i, l_i)}. \quad (2)$$

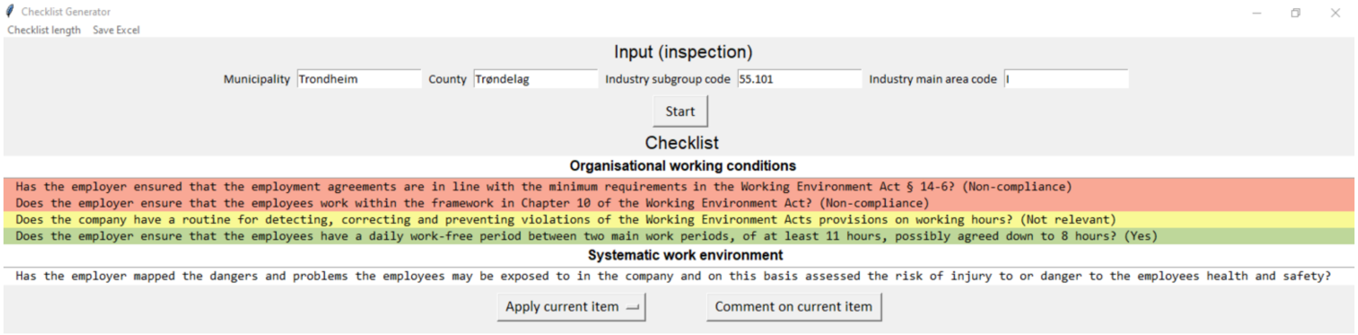
The right hand side of Equation 2 can be reduced to $\arg\max_i \frac{p(1,x,\hat{e},e_i,l_i)}{\sum_{\hat{l}=0}^{1} p(\hat{l},x,\hat{e},e_i,l_i)}$. In some cases $\arg\max_i \theta^{be}$ may have multiple solutions. In that case, we select one of them randomly. We compute $\arg\max_i \theta^{be}$ via sequential search in the checklist $\mathbf{y}^{cnd}$, when a dynamic update takes place. This runs quite fast as checklists are relatively short and because we calculate the parameters p and store them in tables immediately each time an item on the checklist is answered, to reduce computational costs. After finding i, an explanation text for the item $\hat{e}$ in the dynamic update is generated. A demonstration of the text is presented in Section 4.

4 Implementation of Dynamic Checklists

For the field study, we have created a prototype based on the CBCBR framework introduced by Flogard et al. [15]. The prototype is developed based on the Minimum Viable Product (MVP) scheme, which is a lean and cost-effective way to confirm or refute hypotheses about a product's benefits or values [28]. The prototype is also designed according to the Human-AI interaction guidelines proposed by Amershi et al. [2]. The details regarding the interface, functionality, and configuration of the prototype are described below. We also demonstrate a comparison between a dynamic and traditional checklist.

4.1 Prototype Interface and Functionality

A screenshot of a short checklist generated with the prototype is shown in Figure 3. The graphical interface of the software consists of



Input (inspection)

Municipality Trondheim County Trøndelag Industry subgroup code 55.101 Industry main area code I

Checklist

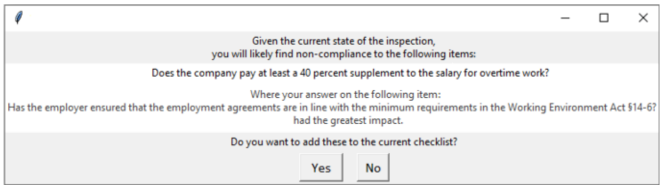
Organisational working conditions

Has the employer ensured that the employment agreements are in line with the minimum requirements in the Working Environment Act § 14-6? (Non-compliance)
Does the employer ensure that the employees work within the framework in Chapter 10 of the Working Environment Act? (Non-compliance)
Does the company have a routine for detecting, correcting and preventing violations of the Working Environment Acts provisions on working hours? (Not relevant)
Does the employer ensure that the employees have a daily work-free period between two main work periods, of at least 11 hours, possibly agreed down to 8 hours? (Yes)

Systematic work environment

Has the employer mapped the dangers and problems the employees may be exposed to in the company and on this basis assessed the risk of injury to or danger to the employees health and safety?

Figure 3. A screenshot of the CBCBR prototype with a short checklist of $K = 5$ items, generated for a labour inspection of a hotel in Trondheim, Norway. As a demonstration, we answered four of the items, marked in red, yellow and green. The colors are used to highlight the answers so that they are easy to recognize.


Given the current state of the inspection,
you will likely find non-compliance to the following items:

Does the company pay at least a 40 percent supplement to the salary for overtime work?

Where your answer on the following item:
Has the employer ensured that the employment agreements are in line with the minimum requirements in the Working Environment Act §14-6?
had the greatest impact.

Do you want to add these to the current checklist?

Figure 4. A pop-up with recommendation of an additional item and an explanation, based on the answers from the checklist in Figure 3.

three parts: the input fields, the checklist and pop-up windows with dynamic updates to the checklist. The functionality of these are described below.

Input Fields. The input fields, shown at the top in Figure 3, are used to specify the features of the organisation that is targeted for the inspection. The features describe the location (municipality) and industry[2] of the target organisation. The fields are used to build the query in Figure 2. The query is executed once the user (inspector) presses the start button. After the user presses the start button, an initial checklist that matches the query appears below the input fields. Different input values to the prototype can yield very different checklists.

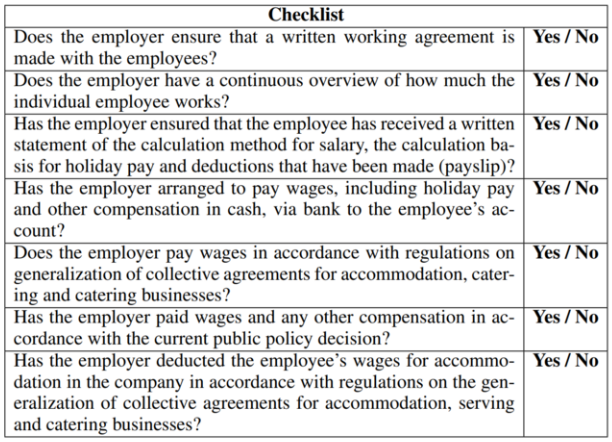
Checklist. The checklist-part of Figure 3 corresponds to the candidate checklist in Figure 2. The checklist is generated for a hotel (ISC 55.101) located in Trondheim municipality, with an initial length of $K = 5$ items. It is possible to adjust the initial checklist length in the upper left corner. To enhance readability for longer checklists, items are grouped under headlines according to the main working environment factor each belongs to. The groups and items within are also initially sorted alphabetically, but users can easily reorder them. The checklist can be saved and opened in Excel format (Save Excel button), where the estimated probability for finding non-compliance is listed for each item to serve as an explanation. We decided not to list the estimates (explanations) on the main GUI to avoid cluttering. As mentioned in Section 3, the purpose of the estimates is to provide ML model transparency and justification regarding the selected items.

The checklist is answered chronologically from top-to-bottom, as shown by the partially filled-out checklist in Figure 3. The user can select answers from a drop-down menu by clicking on the “apply current item” button. The options are “non-compliance” (red color), “yes” (green), “not relevant” (yellow), “not controlled” (yellow), “follow up later” (yellow), and “regulation already checked” (yellow). A yes-answer means that the regulation for the corresponding item is compliant; “not relevant” is used for items that do not relate to the target organisation’s operations; and “not controlled” or “follow up later” are used if the inspector does not have time or lacks the knowledge/information to follow up the item immediately during the inspection. Finally, “regulation already checked” means that the checklist contains another item that corresponds to the same regulation. Since CBCBR relies on binary target labels for training and prediction, we have designated the “non-compliance” answer as 1 (positive) and the rest of the answers as 0 (negative). The logic is that 1 means that non-compliance is found and 0 means that non-compliance is not found. The resulting binary values, mapped from the answers, are also used to update CBCBR and provide dynamic updates with additional items for the checklist.

Dynamic Updates to the Checklist. Dynamic updates to the checklists are implemented as dynamic recommendations for additional items to use during the inspections. Figure 4 demonstrates a recommendation of an additional item, based on all the answers from Figure 3. The recommended item is appended to the bottom of the checklist if the user presses the “yes” button in the dialogue, and it is answered in the same manner as ordinary checklist items. In our implementation, dynamic updates are attempted each time all the checklist items grouped under a headline have been answered. This is the case in Figure 3, where all items under “organisational working conditions” have been answered. It is also worth noting that sometimes no recommendations are made if there are no eligible items outside the checklist that have received sufficiently increased posterior probability estimates to appear in a recommendation. Below the recommended item in Figure 4, an explanation is also provided. The explanation shows the answer on the checklist (Figure 3) that had the most impact on the recommendation of the item, as described in Section 3. For the purpose of the field study, we have encouraged inspectors to accept any dynamic items that are recommended for the checklists. We have chosen not to dynamically remove checklist items, or forcibly append new items to the checklists in the software without user approval as this could strain or confuse users [15].

Using the Prototype in Inspections. We intentionally designed the prototype to operate in a wide range of labour inspections. This is important as the execution of labour inspections is contextual and varies [33, 38]. Inspections can be based on conversations in an office, or perception-based where inspectors walk around and inspect working areas or equipment. Before an inspection starts, an initial checklist is created with the prototype. As the inspection progresses, the checklist (including any dynamic items) is filled out as described above. After the inspection is completed, the inspector saves the checklist to an Excel spreadsheet and then manually uploads it into the case management system where a draft for an inspection report is automatically generated.

[2] An overview of Norwegian industry codes can be found at: https://www.ssb.no/en/klass/klassifikasjoner/6



Checklist	
Does the employer ensure that a written working agreement is made with the employees?	Yes / No
Does the employer have a continuous overview of how much the individual employee works?	Yes / No
Has the employer ensured that the employee has received a written statement of the calculation method for salary, the calculation basis for holiday pay and deductions that have been made (payslip)?	Yes / No
Has the employer arranged to pay wages, including holiday pay and other compensation in cash, via bank to the employee's account?	Yes / No
Does the employer pay wages in accordance with regulations on generalization of collective agreements for accommodation, catering and catering businesses?	Yes / No
Has the employer paid wages and any other compensation in accordance with the current public policy decision?	Yes / No
Has the employer deducted the employee's wages for accommodation in the company in accordance with regulations on the generalization of collective agreements for accommodation, serving and catering businesses?	Yes / No

Figure 5. A simplified version of a traditional checklist currently used by NLIA for inspections in hotels and restaurants. A yes-answer on the checklist means that the inspected organisation is compliant with the regulation in question, while no means that it is non-compliant.

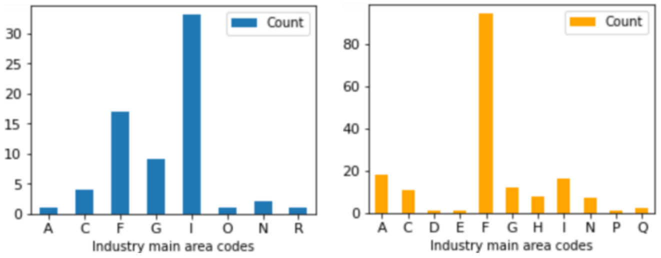
4.2 *Comparison of Dynamic vs. Traditional Checklist*

An example of a traditional checklist is presented in Figure 5, to highlight the difference from the dynamic checklist in Figure 3. Figure 5 shows one of many different checklists that are typically used for an inspection at a hotel (ISC 55.101). The checklist contains 7 items that focus only on working agreements, working hours, and wage requirements. The dynamic checklist in Figure 3 is generated for ISC 55.101 and has 5 items, plus the dynamic recommended item in Figure 4. It covers working agreements, working hours, systematic HSE risk assessments, and overtime payments. The content of the dynamic checklist is broader and more relevant to the inspected organisation and addresses more specific violation risks than the traditional checklist. More specifically, both checklists cover working agreements and salaries but the dynamic checklist also focuses on routines to minimize risk of injuries and ensure decent working hours.

It should be noted that the relevant content of a checklist varies depending on the inspected organisation, as the applicable HSE risks and regulations also vary between organisations. This is a challenge with the current traditional checklists, as hundreds of predefined checklists need to be maintained (NLIA has 369) [13]. Selecting the correct traditional checklist for each specific inspection can therefore be difficult. Such variations are not a problem for dynamic checklists, since they can be created on-demand specifically for each inspected organisation.

4.3 *Prototype Configurations and Setup*

We are using the same configurations as Flogard et al. proposed for their example demonstration of CBCBR, where the components of the query q are $\theta = 100\%$, $\kappa = 70$. The NBI model is implemented via MSSQL2019 and uses the same fixed prior parameter values as Flogard et al. [15]. CBCBR's similarity-based retrieval is implemented via myCBR [3]. To generate the dynamic checklists, the prototype relies on the dataset that Flogard et al. introduced for generating labour inspection checklists [14]. The dataset contains $N = 1967$ different unique checklist items from checklists used in 59989 past labour inspections. Another related dataset exists [13], but it is not suitable for creating new checklists from scratch. We are using the municipality/county and industry code hierarchy from the dataset as features to represent organisations (x) [14, 15]. The GUI of the prototype is implemented via TKinter and installed on Microsoft Surface Pro tablets. CBCBR also runs locally on each tablet. Response times for generating an initial checklist and for dynamic updates are circa 10 and 5 seconds, respectively.



(a) Test group distribution of inspections using dynamic checklists.

(b) Control group distribution of inspections using paper-based checklists.

Figure 6. The diagrams show the number of inspections conducted within each industry[2]. For the test group, most of the inspections were conducted within the Accommodation & Food (I) and Construction industries (F). The majority of the inspections of the control group were carried out in Construction businesses (F).

5 Field Study with Dynamic Checklists

In this section, we present the results from testing our implementation of explainable dynamic checklists in labour inspections. The purpose of the field study is to test the assumption that the checklists increase labour inspection efficiency and the number of violations found by inspectors.

5.1 *Method and Design.*

Design of a Test and Control Group. For the study, seven of NLIA's inspectors volunteered to participate. The field study is conducted as a paired test where the same seven inspectors participate in both a test group and a control group. The test group consists of 69 inspections conducted between March 1. 2022 and October 1. 2022, using the CBCBR prototype. The control group consists of 171 ordinary inspections that the same inspectors conducted within the same period using NLIA's standard paper-based checklists, without any interventions from us. Figure 6a shows how the 69 inspections in the test group are distributed. Most of the efforts are concentrated on Accommodation & Food (I, with 33 inspections), Construction (F, with 17 inspections), and Wholesale & Retail (G, with 9 inspections).[2] Figure 6b shows a different distribution for the control group, where most of the inspections are conducted within the Construction industry. For both the test and control groups, the inspections in Accommodation & Food (F) were carried out by four of the seven inspectors. The inspections in Construction were distinctly carried out by further two of the seven of inspectors. The inspections in Wholesale & Retail and the remaining industries were carried out by the one remaining inspector, in addition to three of the inspectors from Accommodation & Food. The inspections were divided in this manner to avoid disrupting NLIA's inspection efforts and allow inspectors to freely select their targets (discussed in the ethical statement), as inspectors tend to be specialists and therefore carry out most of their inspections within a few industries that are familiar to them. We reviewed other factors such as the size (number of employees) and location of the inspected organisations and found no significant differences between the test and control groups.

Measuring Results. Because labour inspections are industry oriented and due to the differences between the distributions of inspections in the test and control group in Figure 6, we report the results from the study by the following categories to address bias: Accommodation & Food (I), Construction (F), Others and All inspections. This is important as there are substantial differences in how checklists are used and inspections are carried out between Construction and Accommodation & Food. There are relatively few inspections in Wholesale & Retail (G) and the other industries in our study. We therefore grouped these industries into the category named Others, as the inspection results in these industries are similar. The results are reported for each category in terms of average relative frequency of checklist answers per inspection, average number of discovered violations (Avg_v), and average length (Avg_l) of the checklists used. We also use two different average precision scores [14, 15]. These are calculated from a set of N completed inspections as follows: $Prec_v = \frac{1}{N}\sum_{i=1}^{N}\frac{v_i}{|y_i|}$ and $Prec_r = \frac{1}{N}\sum_{i=1}^{N}\frac{r_i}{|y_i|}$. $|y_i|$ is the number of items in each completed checklist y_i (predicted positives), $v_i \in \mathbb{R}_{\geq 0}$ and $r_i \in \mathbb{R}_{\geq 0}$ is the number of violations and the number of reactions in the i-th inspection (true positives), respectively. Each checklist can at most have one violation/reaction per checklist item, so that $v_i \leq |y_i|$ and $r_i \leq |y_i|$ always holds. In words, $Prec_v$ can be explained as the average number of violations per checklist item and $Prec_r$ as the average number of reactions per checklist item. We also use an additional statistic $D\,Prec_v$, which is $Prec_v$ calculated exclusively on the dynamically added part of the checklists. Ideally, it would be beneficial to have more statistics such as recall or accuracy in the study. However, the ground truth (negatives) needed to calculate these is not feasible to obtain [14]. To compare the overall results from the study with current cross-validation scores, we use an industry-weighted average precision score from all inspections in the study to remove bias (see Fig. 6): $Weighted\,Precision = \sum_j w_j\,Prec_v(j)$, where $Prec_v(j) = \frac{1}{N_j}\sum_{i=1}^{N_j}\frac{v_i}{|y_i|}$ is calculated on the set of all (N_j) completed inspections within each industry j (see $Prec_v$). The weights w satisfy $\sum_j w_j = 1$ and each w_j is calculated using Flogard et al.'s dataset [14] $\mathcal{D}$ as follows: $w_j = \frac{S_j}{S_{\mathcal{D}}}$, where S_j is the number of inspections in $\mathcal{D}$ within industry j and $S_{\mathcal{D}}$ is the total number of inspections in $\mathcal{D}$.

5.2 Qualitative Results and Discussions.

We conducted both conversational and structured qualitative interviews with the inspectors in the study, which are summarized due to space restrictions: Overall, the CBCBR prototype is mostly well-received. Most of the inspectors reported an increase in the number of significant working environment violations they found in the inspections when using dynamic checklists, but they also had to spend a bit more time on case management afterward to follow up on the extra violations. The inspectors also perceive dynamic checklists as more relevant to the target organisations, in comparison to the existing paper-based checklists. This is because the content of each checklist is tailored to match the working environment risks in each organisation. The inspectors also reported that they found violations on items that they normally would not think of, especially among the dynamically added items. They found the explanatory probability estimates (explanation method 1) helpful to understand the model's confidence in finding violations to items on the checklists. The explanations for the dynamic checklist updates (explanation method 2) were also useful, both for understanding how and why they should be used. On the negative side, the inspectors reported that dynamic checklists are more difficult to memorize than their paper-based counterparts due to their uniqueness. The checklists also require more attention from the inspectors when operated due to the dynamic updates. The prototype's GUI also needs some improvements.

Table 1. Quantitative results from the test and control group of inspections conducted in the study.

Test Group - Dynamic Checklists

	Acc&Food	Construction	Others	All
$\mathbf{Avg_v}$	9.03 ± 0.52	2.94 ± 0.76	3.53 ± 0.59	6.02 ± 0.54
$\mathbf{Avg_l}$	17.0 ± 0.49	17.5 ± 0.49	16.2 ± 0.58	17.0 ± 0.31
$\mathbf{Prec_v}$	0.53 ± 0.02	0.17 ± 0.02	0.22 ± 0.02	0.35 ± 0.01
$\mathbf{Prec_r}$	0.37 ± 0.02	0.11 ± 0.02	0.18 ± 0.02	0.25 ± 0.01
$\mathbf{D\,Prec_v}$	0.71 ± 0.10	0.24 ± 0.09	0.12 ± 0.06	0.49 ± 0.08

Control Group - Traditional Checklists

	Acc&Food	Construction	Others	All
$\mathbf{Avg_v}$	7.50 ± 0.86	2.88 ± 0.26	2.78 ± 0.38	3.70 ± 0.26
$\mathbf{Avg_l}$	17.6 ± 1.05	22.1 ± 0.78	19.1 ± 0.86	20.6 ± 0.56
$\mathbf{Prec_v}$	0.42 ± 0.03	0.14 ± 0.01	0.15 ± 0.01	0.15 ± 0.01
$\mathbf{Prec_r}$	0.30 ± 0.03	0.09 ± 0.01	0.10 ± 0.01	0.12 ± 0.01

5.3 Quantitative Results and Discussions.

Overall Results from the Field Study. The results in Table 1 show significant increases in the average number of violations found per inspection (Avg_v) between the test and control groups. For all inspections, the increase is 62.7% (6.02 vs 3.70), but some of the difference can be explained by the fact that the test group contains more inspections conducted in Accommodation & Food than the control group (see Figure 6), where inspectors find more violations than any other industries. For Accommodation & Food and Others, the increases are 20.5% (9.03 vs 7.50) and 27% (3.53 vs 2.78) respectively, which are significant and likely attributed to the dynamic checklists. It seems that the dynamic checklists are less effective in Construction, as the increase in Avg_v is insignificant. However, the average precision per checklist item for violations ($Prec_v$) for Construction is still significantly higher in the test group, as the average checklist size (Avg_l) is lower. $Prec_v$ is also significantly higher in Accommodation & Food and in Others as well, as more violations are found. The benefits of finding more violations are discussed in Section 1, and shorter checklists may also decrease time spent on the inspections and the cognitive load for the inspectors [12]. Thus, the overall increases in $Prec_v$ in the test group can therefore be seen as indicators for increased labour inspection efficiency [14, 15]. The results in Table 1 also support Flogard et al.'s claim that dynamic updates to the checklists increase overall precision [15]. $D\,Prec_v$ is the precision score exclusively for the dynamic part of checklists and the overall score is 49% (0.49), which is significantly higher than the $Prec_v$ score of 35% for the full dynamic checklists in the test group. Without the checklist updates, the overall $Prec_v$ would have been 33%. The effectiveness of the dynamic items varies among the industries, and they seem to be less effective in the Others category (test group) as $D\,Prec_v$ is 9 percentage points less than $Prec_v$.

Distribution of Checklist Answers. Table 1 also shows variations in the scores between different industries, which are similar for both the test and control groups. Therefore, we look more closely into how the checklists are used in the different industries for the test group only. Table 2 shows the distribution of checklist answers in the test group. There are clear differences in how inspectors interact with the checklists, based on the industries. For all industries combined,

Table 2. Average relative frequency distribution of checklist answers per inspection from the test group in the field study.

	Acc&Food	Construction	Others	All
Non-compliance	0.53	0.17	0.22	0.35
Yes	0.42	0.52	0.59	0.49
Not controlled	0	0.06	0.04	0.03
Not relevant	0.04	0.25	0.15	0.13
Follow up later	0.01	0	0	0
Total	1.00	1.00	1.00	1.00

35% of the checklist items are non-compliant, 49% are compliant, and the rest are either not controlled or not relevant. The Construction industry has the highest share of checklist items marked as "not relevant" and "not controlled" (31%), which is interesting as Construction also has the lowest $Prec_v$ scores in Table 1 for both the test and control groups. In contrast, the shares of these answers in the Others and Accommodation & Food categories are only 19% and 4% respectively. Therefore, it seems that inspectors generally find checklists less relevant for inspections in Construction, compared to other industries. There are also considerably less yes-answers and more non-compliance-answers in Accommodation & Food compared to the Others category, which indicates lower compliance with overall working environment regulations for this industry. However, some of these observed differences between the industries may not be industry-specific but could be caused by individual differences between inspectors of the different industries. This is because inspection efforts in each industry are distinctly divided between the 7 inspectors participating in the study, pair-wise for both test and control groups as mentioned earlier.

Field Study versus Cross-Validation Performance. Figure 7 shows the average precision performance scores for dynamic vs. traditional checklists for different experimental setups. The right plots show $Prec_v$ scores based on estimates, using empirical distributions from the validation parts of the dataset [15]. The middle plots show cross-validation scores using only available ground truth labels [15]. The left plots show the overall $Prec_v$ scores from this field study, which are weighted according to how inspections are distributed among industries in Flogard et al.'s dataset, for comparisons with the other plots. For the baseline traditional checklists (orange), all the scores are nearly identical with 0.17 for the field study and 0.18 for the data sets. For the dynamic checklists (blue), the weighted score from the field study (0.24) is much lower than the cross-validation scores of 0.32 and 0.50. This indicates that the cross-validations are unable to accurately estimate the field performance of dynamic checklists. The discrepancies might be attributed to confounding factors related to the use of checklists in the real world that are not accounted for in the cross-validation, such as how users interact with their checklists [16]. Cross-validation may still not be completely unreliable, as Figure 7 shows that dynamic checklists consistently outperform traditional checklists in both the cross-validations and field study. Yet, the differences between the field study and cross-validation results highlight the importance of field-testing ML methods.

6 Conclusion

In this paper, we developed a prototype based on the current state-of-the-art ML method for generating dynamic checklists (CBCBR). We also propose two different approaches for explaining the content of the checklists to inspectors, which are implemented into the prototype. The prototype is tested in a field study of real-world labour inspections. The results indicate that the *efficiency of labour inspections significantly increases with explainable dynamic checklists*. The way checklists are answered varies based on the industry where the inspections are carried out. Some of these variations could be caused by individual differences in inspection practices between inspectors. Our findings also suggest that current cross-validation methods [15] do not accurately reflect real-world performances of ML-based checklists. Field testing is therefore essential for obtaining fully reliable estimates of checklist performance.

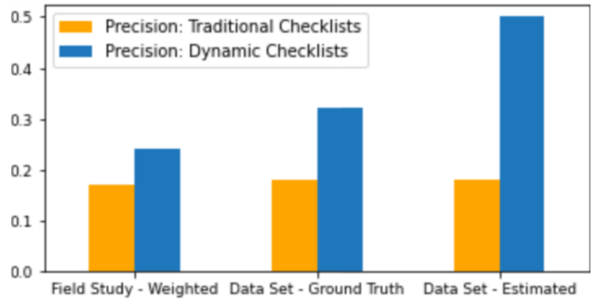


Figure 7. Weighted average precision scores from this field study (left) versus ground truth precision (middle) and estimated precision scores (right) from cross-validations done in previous work [15]. The standard error is 0.01 for the field study and 0 for the rest.

Research on using ML for generating checklists is in an early stage. Despite this, we believe that the results from the field study are strong enough to encourage labour inspection authorities to adopt and further develop ML methods for creating checklists, which could increase national and global levels of compliance with labour rights and reduce injuries (SDG indicators 8.8.1 and 8.8.2). NLIA has already plans to further develop our prototype into a system that can be used nationwide in Norway, replacing the 369 different traditional checklists currently being used [13]. To accomplish this, improving the user interface of the prototype is important. Based on the results from the study and the feedback from the inspectors, it is likely that doing so could further increase inspection performance. The dataset used for the prototype does not have many features, so adding more features and using feature selection could be an option to optimize performance [13, 26, 24, 25]. Another direction for future work is to take advantage of the transparency of the CBCBR method, and develop more explanation methods to promote and increase the inspectors' trust in the system. In particular, inspectors have requested methods that provide counterfactual explanations for why certain items have not made it into their checklists.

Ethical Statement

We seek to avoid conducting research in ways that can have negative impacts. An ethical concern for this field study is that the time the inspectors spend on inspections for the study could be spent on something else. Thus, we designed the study to avoid disrupting inspectors from their daily tasks or degrading the quality of the inspections. Some of the design choices may therefore not be optimal from a scientific point of view, such as letting the inspectors select organisations for inspections based on their own decisions (in both test and control groups) or that appending extra dynamic items to the checklists is not done automatically without approval from the user. Privacy for the participating inspectors is also a possible concern, and we have therefore collected an informed consent from the participating inspectors for the purpose of this paper. We have also taken care to not provide any results or information in this paper that can be used to identify any businesses subjected to the labour inspections.

References

[1] Agnar Aamodt and Enric Plaza, 'Case-based reasoning: Foundational issues, methodological variations, and system approaches', *AI communications*, **7**(1), 39–59, (1994).

[2] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al., 'Guidelines for human-ai interaction', in *Chi conference on human factors in computing systems*, pp. 1–13, (2019).

[3] Kerstin Bach, Bjørn Magnus Mathisen, and Amar Jaiswal, 'Demonstrating the mycbr rest api.', in *ICCBR Workshops*, pp. 144–155, (2019).

[4] Hubo Cai, JungHo Jeon, Xin Xu, Yuxi Zhang, Liu Yang, et al., 'Automating the generation of construction checklists', Technical report, Purdue University. Joint Transportation Research Program, (2020).

[5] Ken Catchpole and Stephanie Russ, 'The problem with checklists', *BMJ quality & safety*, **24**(9), 545–549, (2015).

[6] Stefan C Christov, Heather M Conboy, Nancy Famigletti, George S Avrunin, Lori A Clarke, and Leon J Osterweil, 'Smart checklists to improve healthcare outcomes', in *International Workshop on SEHS*, pp. 54–57, (2016).

[7] European Commission. Proposal for a regulation of the european parliament and of the council laying down harmonised rules on artificial intelligence (artificial intelligence act) and amending certain union legislative acts, 2021.

[8] Øyvind Dahl and Marius Søberg, 'Labour inspection and its impact on enterprises' compliance with safety regulations', *Safety Science Monitor*, **17**(2), 1–12, (2013).

[9] Jesus M Darias, Marta Caro-Martínez, Belén Díaz-Agudo, and Juan A Recio-Garcia, 'Using case-based reasoning for capturing expert knowledge on explanation methods', in *ICCBR*, pp. 3–17. Springer, (2022).

[10] Adnan Darwiche, *Modeling and reasoning with Bayesian networks*, Cambridge university press, 2009.

[11] AJR De Bie, S Nan, LRE Vermeulen, PME Van Gorp, RA Bouwman, AJGH Bindels, and HHM Korsten, 'Intelligent dynamic clinical checklists improved checklist compliance in the intensive care unit', *BJA: British Journal of Anaesthesia*, **119**(2), 231–238, (2017).

[12] Ashley JR De Bie, Eveline Mestrom, Wilma Compagner, Shan Nan, Lenneke van Genugten, Kiran Dellimore, Jacco Eerden, Steffen van Leeuwen, Harald van de Pol, Franklin Schuling, et al., 'Intelligent checklists improve checklist compliance in the intensive care unit: a prospective before-and-after mixed-method study', *British Journal of Anaesthesia*, **126**(2), 404–414, (2021).

[13] Eirik Lund Flogard and Ole Jakob Mengshoel, 'A dataset for efforts towards achieving the sustainable development goal of safe working environments', in *Thirty-sixth Conference on Neural Information Processing Systems (NeurIPS)*, (2022).

[14] Eirik Lund Flogard, Ole Jakob Mengshoel, and Kerstin Bach, 'Bayesian feature construction for case-based reasoning: Generating good checklists', in *International Conference on Case-Based Reasoning*, pp. 94–109. Springer, (2021).

[15] Eirik Lund Flogard, Ole Jakob Mengshoel, and Kerstin Bach, 'Creating dynamic checklists via bayesian case-based reasoning: Towards decent working conditions for all', in *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pp. 5108–5114, (2022).

[16] Rodrigo J Daly Guris and Meghan B Lane-Fall, 'Checklists and cognitive aids: underutilized and under-researched tools to promote patient safety and optimize clinician performance', *Current Opinion in Anaesthesiology*, **35**(6), 723–727, (2022).

[17] Daniel E Ho, Sam Sherman, and Phil Wyman, 'Do checklists make a difference? a natural experiment from food safety enforcement', *Journal of Empirical Legal Studies*, **15**(2), 242–277, (2018).

[18] Qixuan Jin, Haoran Zhang, Thomas Hartvigsen, and Marzyeh Ghassemi, 'Fair multimodal checklists for interpretable clinical time series prediction', in *NeurIPS 2022 Workshop on Learning from Time Series for Health*, (2022).

[19] Nektarios Karanikas and Sikder Mohammad Tawhidul Hasan, 'Occupational health & safety and other worker wellbeing areas: Results from labour inspections in the bangladesh textile industry', *Safety Science*, **146**, 105533, (2022).

[20] Eoin M Kenny and Mark T Keane, 'Twin-systems to explain artificial neural networks using case-based reasoning: Comparative tests of feature-weighting methods in ann-cbr twins for xai', in *Twenty-Eighth International Joint Conferences on Artifical Intelligence (IJ-CAI), Macao, 10-16 August 2019*, pp. 2708–2715, (2019).

[21] Leah Kulp, Aleksandra Sarcevic, Megan Cheng, and Randall S Burd, 'Towards dynamic checklists: Understanding contexts of use and deriving requirements for context-driven adaptation', *ACM TOCHI*, **28**(2), 1–33, (2021).

[22] Yukti Makhija, Edward De Brouwer, and Rahul G Krishnan, 'Learning predictive checklists from continuous medical data', *arXiv preprint arXiv:2211.07076*, (2022).

[23] Aditya Mate, Lovish Madaan, Aparna Taneja, Neha Madhiwalla, Shresth Verma, Gargi Singh, Aparna Hegde, Pradeep Varakantham, and Milind Tambe, 'Field study in deploying restless multi-armed bandits: Assisting non-profits in improving maternal and child health', in *AAAI Conference on Artificial Intelligence*, pp. 12017–12025, (2022).

[24] Ole Jakob Mengshoel, Eirik Flogard, Jon Riege, and Tong Yu, 'Stochastic local search heuristics for efficient feature selection: An experimental study', in *Norsk IKT-konferanse for forskning og utdanning*, pp. 58–71, (2021).

[25] Ole Jakob Mengshoel, Eirik Lund Flogard, Tong Yu, and Jon Riege, 'Understanding the cost of fitness evaluation for subset selection: Markov chain analysis of stochastic local search', in *Genetic and Evolutionary Computation Conference*, pp. 251–259, (2022).

[26] Ole Jakob Mengshoel, Tong Yu, Jon Riege, and Eirik Flogard, 'Stochastic local search for efficient hybrid feature selection', in *Genetic and Evolutionary Computation Conference*, pp. 133–134, (2021).

[27] Christoph Molnar, Gunnar König, Julia Herbinger, Timo Freiesleben, Susanne Dandl, Christian A Scholbeck, Giuseppe Casalicchio, Moritz Grosse-Wentrup, and Bernd Bischl, 'General pitfalls of model-agnostic interpretation methods for machine learning models', in *xxAI-Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020*, pp. 39–68. Springer, (2022).

[28] Dobrila Rancic Moogk, 'Minimum viable product and the importance of experimentation in technology startups', *Technology Innovation Management Review*, **2**(3), (2012).

[29] World Health Organization, 'Joint estimates of the work-related burden of disease and injury, 2000-2016: Global monitoring report', Technical report, (2021).

[30] Guglielmo Papagni, Jesse de Pagter, Setareh Zafari, Michael Filzmoser, and Sabine T Koeszegi, 'Artificial agents' explainability to support trust: considerations on timing and context', *AI & SOCIETY*, 1–14, (2022).

[31] Judea Pearl, 'The seven tools of causal inference, with reflections on machine learning', *Communications of the ACM*, **62**(3), 54–60, (2019).

[32] Andi Peng, Besmira Nushi, Emre Kiciman, Kori Inkpen, and Ece Kamar, 'Investigations of performance and bias in human-ai teamwork in hiring', in *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 12089–12097, (2022).

[33] Roberto Pires, 'Promoting sustainable compliance: Styles of labour inspection and compliance outcomes in brazil', *International Labour Review*, **147**(2-3), 199–229, (2008).

[34] Juan A Recio-García, Belén Díaz-Agudo, and Victor Pino-Castilla, 'Cbr-lime: a case-based reasoning approach to provide specific local interpretable model-agnostic explanations', in *ICCBR*, pp. 179–194. Springer, (2020).

[35] Frode Sørmo, Jörg Cassens, and Agnar Aamodt, 'Explanation in case-based reasoning–perspectives and goals', *Artificial Intelligence Review*, **24**(2), 109–143, (2005).

[36] Baptiste Vasey, Myura Nagendran, Bruce Campbell, David A Clifton, Gary S Collins, Spiros Denaxas, Alastair K Denniston, Livia Faes, Bart Geerts, Mudathir Ibrahim, et al., 'Reporting guideline for the early stage clinical evaluation of decision support systems driven by artificial intelligence: Decide-ai', *bmj*, **377**, (2022).

[37] David Walters, 'Labour inspection and health and safety in the eu', *The European Trade Union Institute's (ETUI) Health and Safety at Work Magazine,(14)*, 12–17, (2016).

[38] David Weil, 'Improving workplace conditions through strategic enforcement', *Boston U. School of Management Research Paper*, (2010-20), 2–3, (2010).

[39] Chathurika S Wickramasinghe, Daniel L Marino, Javier Grandio, and Milos Manic, 'Trustworthy ai development guidelines for human system interaction', in *2020 13th International Conference on Human System Interaction (HSI)*, pp. 130–136. IEEE, (2020).

[40] Haoran Zhang, Quaid Morris, Berk Ustun, and Marzyeh Ghassemi, 'Learning optimal predictive checklists', *NeurIPS*, **34**, (2021).