

Investigating Feature Substitution in Algorithmic Worker-Scoring Models

Abdulrahman Ababneh

The University of Sydney
Sydney, Australia
aaba0704@uni.sydney.edu.au

Abstract

Algorithmic management systems increasingly use operational data to predict, rank, allocate, and evaluate work. This paper investigates whether worker-relevant predictive capacity persists after explicit worker-identifying and monitoring features are removed from algorithmic scoring models. Using a public warehouse task-execution dataset, we train classification models to predict whether tasks fall above the median observed duration, a worker-relevant operational outcome that may inform allocation, performance management, or staffing decisions. We compare model performance across full-feature, direct-monitoring-excluded, operational/proxy-only, and stricter proxy-ablation settings. The results show that removing explicit worker-identifying fields only modestly reduces predictive performance and the direct-monitoring-excluded model retains most of the full model’s ROC-AUC, while operational proxy models remain meaningfully predictive even after removing workzone and priority variables. These findings suggest that feature removal may reduce formal observability without eliminating worker-relevant scoring capacity, because operational context variables can preserve predictive signal. We contribute a sociotechnical framework for analysing feature substitution in algorithmic management and argue that audits should examine retained predictive capacity, not merely the presence of direct monitoring features.

Introduction

Algorithmic management systems increasingly structure how work is assigned, paced, evaluated, and rewarded. In platform-mediated labour, ride-hailing, logistics, and warehouse work, computational systems can assign tasks, provide informational support, evaluate performance, and shape workers’ access to future opportunities (Lee et al. 2015; Möhlmann and Zalmanson 2017; Kellogg, Valentine, and Christin 2020; Rosenblat and Stark 2016; Wood et al. 2019). In warehouse and delivery environments, these systems are not limited to visible monitoring devices: work can be organised through barcodes, task records, location data, workflow systems, and other operational traces that make labour digitally legible (Delfanti 2021; Newlands 2021). These systems are often justified as tools for optimisation, but they also create new forms of workplace visibility.

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

A common response to concerns about surveillance, discrimination, or excessive profiling is to remove direct identifiers or overt monitoring variables from model inputs. In a workplace setting, this might mean excluding an employee identifier, tracking variable, or explicit productivity measure. This intervention is intuitively appealing because it appears to reduce the model’s access to surveillance-relevant information. However, prior work on algorithmic decision-making shows that removing a sensitive or contested feature does not necessarily remove the model’s ability to reconstruct or act on correlated information (Barocas and Selbst 2016; Dwork et al. 2012; Datta et al. 2017; Prince and Schwarcz 2020). In workplace systems, task category, location, priority, quantity, workflow stage, or scheduling context may act as operational proxies. Even when such variables are not intended to identify or evaluate a worker, they may retain enough information to support worker-relevant prediction.

This paper investigates this problem as *feature substitution*: the persistence of worker-relevant predictive capacity after direct monitoring or worker-identifying features are removed. A system may therefore appear less visibly surveillant without becoming substantially less capable of worker-relevant scoring. The concept is related to concerns about proxy discrimination and fairness through unawareness, but the workplace setting gives it a distinct institutional meaning. The issue is not only whether a model reconstructs a protected demographic attribute, but whether ordinary operational data can preserve the capacity to rank, allocate, or evaluate work.

This distinction matters because many governance responses to workplace AI focus on visible inputs. A system that does not use an employee identifier, direct productivity score, or explicit monitoring stream may appear less intrusive than one that does. Yet workers are affected by model outputs, not only by the formal labels attached to model inputs. This is consistent with broader critiques of transparency and accountability in automated decision systems: input disclosure alone does not necessarily reveal how a system functions, how it affects people, or whether affected individuals can meaningfully contest it (Ananny and Crawford 2018; Kroll et al. 2017; Wachter, Mittelstadt, and Russell 2018; Citron and Pasquale 2014). If ordinary operational variables preserve the ability to predict worker-relevant out-

comes, then a system may satisfy a narrow form of data minimisation while continuing to support allocation, ranking, or managerial attention.

This creates a gap between *formal feature removal* and *functional reduction* in scoring capacity. It also connects to contextual accounts of privacy and sociotechnical accounts of fairness: whether a data practice is acceptable depends not only on the type of data used, but also on the institutional context, purpose, information flow, and downstream consequences of prediction (Nissenbaum 2010; Selbst et al. 2019). In algorithmic management, operational data may be collected for coordination, scheduling, or inventory control, but later used to support worker-relevant prediction. This paper treats that shift from formal input restriction to retained functional capacity as the central governance problem.

We study feature substitution using a public warehouse task-execution dataset. We construct a binary prediction task that classifies whether a task falls above the median observed task duration. Task duration is not treated as a direct measure of worker productivity. It is a worker-relevant operational outcome: in algorithmic management systems, predicted task duration may inform staffing, allocation, and performance expectations. This makes it an appropriate setting for examining whether predictive capacity persists after removing direct features.

We therefore ask the following research questions:

1. **RQ1:** To what extent does worker-relevant predictive capacity persist after direct worker-identifying or monitoring features are removed from an algorithmic management model?
2. **RQ2:** Which operational proxy features preserve predictive capacity after direct worker-identifying fields are excluded?
3. **RQ3:** Does retained predictive capacity persist under stricter ablations and temporal validation?

To examine these questions, we conduct a feature-ablation study using a public warehouse task-execution dataset. We model whether tasks fall above the median observed duration and compare predictive performance across full, direct-monitoring-excluded, operational/proxy-only, and stricter proxy-ablation settings. Our results show that worker-relevant predictive capacity persists after direct worker-identifying fields are removed. The direct-monitoring-excluded model retains most of the full model’s ROC-AUC, and stricter ablations remain meaningfully above chance even after removing salient operational proxies.

This paper makes three contributions:

1. **Conceptual:** We define *retained predictive capacity* as a diagnostic for feature substitution in algorithmic management.
2. **Empirical:** We provide a feature-ablation study showing that operational warehouse variables can sustain worker-relevant prediction after explicit worker identifiers are excluded.
3. **Governance:** We argue that audits of workplace AI should evaluate functional scoring capacity, regardless of the presence or absence of direct monitoring features.

The paper therefore treats feature removal as an empirical audit question rather than a formal compliance step: the relevant issue is whether the model’s worker-relevant functionality remains after it is removed.

Related Work

Algorithmic Management and Worker Scoring

Algorithmic management research examines how computational systems increasingly perform managerial functions such as assigning tasks, directing behaviour, or even shaping access to work. Kellogg, Valentine, and Christin describe algorithmic control as operating through six mechanisms: restricting, recommending, recording, rating, replacing, and rewarding (Kellogg, Valentine, and Christin 2020). This framing is useful because it shows that algorithmic management is not limited to explicit surveillance; it can also operate through routine allocation, evaluation, and feedback systems. In platform work, Rosenblat and Stark show how information asymmetries allow platforms to exercise indirect control over workers while maintaining narratives of flexibility and autonomy (Rosenblat and Stark 2016). Wood et al. similarly find that algorithmic control in online labour platforms can coexist with flexibility while producing overwork, uncertainty, and irregular schedules (Wood et al. 2019).

This literature establishes that algorithmic systems can function as managerial infrastructure, but it generally focuses on organisational control, worker experience, or platform governance. Less attention has been paid to the model-level question this paper examines of whether worker-relevant scoring capacity still persists after explicit worker-identifying or monitoring features are removed.

Workplace Surveillance and Operational Data

Workplace surveillance research shows that employers can collect and analyse increasingly granular data about workers and work processes. Ajunwa, Crawford, and Schultz argue that advances in analytics, mobile devices, biometrics, and productivity applications have expanded the scale and convenience of worker surveillance (Ajunwa, Crawford, and Schultz 2017). Ravid et al. define electronic performance monitoring as technology-enabled collection and analysis of information linked to job performance, including indirect performance-related traces (Ravid et al. 2020). Data & Society’s work on workplace monitoring identifies prediction and flagging tools, remote monitoring, time tracking, gamification, and algorithmic management as major trends in contemporary workplace surveillance (Mateescu and Nguyen 2019). Ajunwa’s later work on the quantified worker further emphasises how data-driven workplace systems can shift risk and intensify managerial visibility (Ajunwa 2023).

These accounts are central for understanding why operational data matters in the workplace. However, much of the surveillance literature treats monitoring as a question of data collection. What remains underexplored is the predictive substitutability of operational variables.

Proxy Variables and Feature Substitution

The problem of feature substitution is related to long-standing concerns in algorithmic fairness about proxy variables and fairness through unawareness. Barocas and Selbst argue that data mining can reproduce discrimination even when sensitive attributes are not explicitly used, because models may discover regularities that reflect existing patterns of exclusion (Barocas and Selbst 2016). Dwork et al. respond to the limitations of simple unawareness approaches by proposing fairness through awareness, requiring a task-specific account of similarity between individuals (Dwork et al. 2012). Datta et al. formalise proxy discrimination as the use of protected-class correlates that have causal influence on model outputs, showing that proxy effects can be difficult to detect by inspection alone (Datta et al. 2017). Prince and Schwarcz similarly argue that removing obvious proxies may simply lead models to locate less intuitive proxies in large datasets (Prince and Schwarcz 2020).

This literature demonstrates that excluding sensitive attributes is not sufficient to prevent indirect inference. The workplace setting, however, raises a related but distinct problem. The concern is not only whether models reconstruct protected demographic attributes, but whether operational variables substitute for direct monitoring features in ways that preserve evaluation capacity. This paper extends proxy-variable concerns from protected-attribute inference to algorithmic management.

Auditing, Explainability, and Retained Predictive Capacity

Algorithmic accountability research has developed tools and frameworks for documenting, auditing, and explaining machine learning systems. Mitchell et al. propose model cards as a structured form of model reporting that documents intended uses, performance characteristics, and ethical considerations (Mitchell et al. 2019). Raji et al. argue for end-to-end internal algorithmic auditing across the AI development lifecycle, emphasising that harms can be difficult to trace once systems are deployed (Raji et al. 2020). Ananny and Crawford caution that transparency alone is insufficient for algorithmic accountability: seeing a system does not necessarily mean understanding or governing it (Ananny and Crawford 2018). Recent work on the algorithmic auditing ecosystem further argues that audits require clearer standards, stakeholder involvement, and attention to real-world harms rather than simple formal evaluation (Costanza-Chock et al. 2023).

These approaches help in shifting accountability from transparency to concrete evaluation. Yet existing audit practices often focus on documentation or input disclosure. They less often explore how much predictive capacity remains after contested features are removed. A model card may identify intended uses, and an audit may disclose which variables are included, but neither necessarily address if removing direct monitoring or worker-identifying features reduces system capacity.

Problem Formulation

Feature substitution can be understood as a functional problem rather than only a feature-label problem. In a workplace model, removing an employee identifier or direct monitoring field changes the input space, but it does not necessarily change what the model can infer. Other variables may preserve similar predictive information because they encode the organisation of work such as where tasks occur, how tasks are prioritised, what type of work is being performed, and how workflows are structured. The object of analysis is whether the model retains the capacity to support worker-relevant prediction after direct features are removed.

We distinguish between three types of features. *Direct monitoring features* are variables that explicitly identify or directly track a worker, such as employee identifiers or resource-level fields. *Operational proxy features* are variables that describe the work context rather than the worker directly, such as task category, workzone, priority, quantity, or workflow type. *Derived scoring capacity* refers to the ability of a model to predict a worker-relevant operational outcome even when direct monitoring features are absent.

The central object of analysis is not whether a specific feature is normatively acceptable in isolation, but whether the model retains predictive capacity after feature-removal. Let $AUC(F)$ denote the predictive performance of a model trained on a full non-leakage feature set, and let $AUC(R)$ denote the performance of a model trained on a reduced feature set after direct monitoring variables are removed. We define retained predictive capacity as:

$$\text{Retention}(R) = \frac{AUC(R)}{AUC(F)}.$$

Retention should be interpreted comparatively. It is not an absolute measure of harm. A high retention value does not show that a system is unlawful, discriminatory, or individually surveillant. It shows that the feature-removal intervention has not substantially reduced the model’s predictive functionality. In an audit context, this is useful because it identifies cases where formal input restrictions may not correspond to meaningful functional change. This provides an empirical signal that the model’s scoring capacity may be functionally substitutable through other variables.

This formulation is deliberately narrower than a full theory of workplace harm. Retention does not measure whether a system is lawful, fair, or democratically legitimate. Instead, it measures whether a feature-removal intervention produces a meaningful reduction in model functionality. This is important because algorithmic accountability mechanisms often focus on documentation, explanation, or disclosure, while affected workers may care more about what the system can still do after apparently protective changes are made (Kroll et al. 2017; Mitchell et al. 2019; Raji et al. 2020; Wachter, Mittelstadt, and Russell 2018).

We examine this through a warehouse task-execution setting. The prediction target is whether a task falls above the median observed task duration. Task duration is worker-relevant because it may inform staffing, allocation, expected pace, or managerial intervention. This makes it an appropri-

ate operational outcome for studying how scoring capacity can persist through indirect features.

This formulation also separates three claims that are often collapsed in debates about workplace AI. The first is an *input claim*: whether a system uses direct worker-identifying or monitoring variables. The second is a *performance claim*: whether removing those variables reduces predictive capacity. The third is a *governance claim*: whether the remaining predictive capacity is acceptable or institutionally constrained. This paper focuses on the second claim while explaining why it matters for the third.

The distinction between input, performance, and governance claims is especially important in workplace settings. Direct identifiers may be removed to satisfy privacy, compliance, or data-minimisation concerns, but the remaining variables may still encode the organisation of work. Workzone, priority, and task category are not merely neutral descriptors; they are produced by managerial choices, infrastructure, workflow design, and historical patterns of task allocation. A high-retention reduced model therefore raises a sociotechnical question of whether the workplace system has reduced surveillance capacity, or merely shifted that capacity into less obvious operational features (Selbst et al. 2019; Newlands 2021; Delfanti 2021).

The formulation also clarifies the paper’s governance claim. The issue is not only whether a system contains direct worker identifiers. A model may become less visibly surveillant while still retaining the capacity to rank, allocate, or evaluate work in ways that affect workers. Audits of algorithmic management systems should therefore ask not only what features are present, but how much worker-relevant predictive capacity remains after contested features are removed.

Methodology

Figure 1 summarises the study design. The analysis constructs a worker-relevant duration target, compares progressively restricted feature settings, and evaluates retained predictive capacity as an audit signal for feature substitution.

Dataset and Prediction Target

We use a public warehouse task-execution dataset associated with workforce allocation under uncertain activity duration (Derkinderen, Bekker, and Smet 2023a,b). The dataset contains records of warehouse tasks and associated operational fields, including task type, category, priority, quantity, workzone, employee identifier, and timestamp information. We use the data as a retrospective case study of operational prediction in algorithmic management. The dataset is not treated as a deployed surveillance system, and the analysis does not make claims about individual worker quality.

We construct the prediction target from task start and stop timestamps. For each valid task record, we compute task duration as the difference between the stop and start time. Records with missing, invalid, or non-positive duration are removed. To reduce the influence of extreme system artefacts or non-standard task records, we remove observations in the bottom and top one percent of the

duration distribution. We then define a binary outcome, `high_duration_task`, equal to 1 when a task’s duration is above the median valid task duration and 0 otherwise. This produces a balanced prediction task and avoids treating task duration as a precise continuous measure of worker performance.

Feature Settings

We evaluate predictive performance across progressively restricted feature settings. The *full non-leakage model* includes all usable non-leaky variables after excluding direct timestamp and duration-derived fields. This setting retains the employee identifier where available. The *direct-monitoring excluded model* removes worker-identifying or direct monitoring variables, including the employee identifier. The *operational/proxy-only model* uses operational task-context variables that do not directly identify workers, such as task type, category, priority, quantity, and workzone.

We also include stricter robustness ablations to test whether predictive capacity is concentrated in a small number of salient proxies. The *no workzone proxy model* removes workzone variables from the operational proxy set. The *no priority proxy model* removes priority variables. The *no workzone or priority model* removes both. Finally, where available, we train an *employee ID only model* to estimate how much predictive capacity is associated with the direct worker identifier alone.

This design allows us to distinguish three possibilities. First, if predictive performance collapses after removing direct worker-identifying features, then the direct monitoring variables carry most of the scoring capacity. Second, if performance remains high only when workzone or priority are retained, then feature substitution is concentrated in salient operational proxies. Third, if performance remains meaningfully above chance even after removing these proxies, then worker-relevant predictive capacity is distributed across a broader set of operational variables.

Models and Preprocessing

We train two standard classification models: logistic regression and histogram-based gradient boosting. Logistic regression provides an interpretable linear baseline, while gradient boosting captures non-linear relationships among operational features. Numerical variables are median-imputed and standardised. Categorical variables are imputed using the most frequent value and one-hot encoded. All preprocessing steps are fit only on the training data within each evaluation split to avoid leakage.

We do not optimise hyperparameters extensively. This choice reflects the purpose of the study. The goal is not to maximise predictive performance, but to compare relative performance across feature settings. Because the research question concerns retained predictive capacity, the main quantity of interest is the change in performance between full, reduced, and proxy-only models under the same modelling pipeline.



Figure 1: Overview of the feature-substitution audit design. The figure summarises the modelling pipeline.

Evaluation

We use stratified three-fold cross-validation as the primary evaluation procedure. The primary metric is ROC-AUC, because algorithmic management systems often operate as ranking, allocation, or prioritisation systems rather than only as fixed-threshold classifiers. We also report PR-AUC, F1, accuracy, and balanced accuracy as secondary metrics.

To test whether the results are driven by random-fold similarity between records, we also conduct temporal validation. In this evaluation, task records are sorted by start time. Models are trained on earlier records and evaluated on later records. This stricter validation setting tests whether predictive capacity persists under temporal distribution shift.

Finally, we compute permutation importance for the strongest reduced model to identify which operational proxy variables preserve predictive signal. It helps interpret which task-context variables contribute most to retained predictive capacity after direct worker-identifying fields are removed.

Results

Table 1 reports predictive performance across the feature-ablation settings. The full non-leakage model achieved a ROC-AUC of 0.822. Removing direct worker-identifying fields reduced ROC-AUC to 0.802, a drop of 0.020. This means that the direct-monitoring-excluded model retained approximately 97.6% of the full model’s predictive performance. The operational/proxy-only model achieved the same ROC-AUC, indicating that the remaining task-context variables preserved most of the worker-relevant predictive signal.

The small performance drop between the full and direct-monitoring-excluded models is the central empirical result. If the explicit employee identifier carried most of the worker-relevant predictive signal, removing it would be expected to substantially reduce performance. Instead, the model retained nearly all of its predictive capacity. This suggests that direct worker identity was not necessary for the model to predict whether a task would fall above the median observed duration. The identical performance of the direct-monitoring-excluded and operational/proxy-only models further indicates that the predictive signal was preserved by operational task-context variables rather than by explicitly worker-identifying fields.

The robustness ablations show that predictive capacity is not solely explained by the direct employee identifier. The employee-ID-only model achieved a ROC-AUC of 0.699,

which is meaningfully above chance but below the operational proxy models. This suggests that worker identity contains signal, but that operational context explains more of the prediction task than the employee identifier alone. Removing workzone reduced ROC-AUC from 0.802 to 0.745, indicating that location or workflow structure is an important proxy. However, performance remained above chance. Removing priority had a much smaller effect, with ROC-AUC remaining at 0.798. Even when both workzone and priority were removed, ROC-AUC remained at 0.738. This stricter ablation weakens the possibility that the result is driven by a single obvious proxy variable.

These ablations also clarify the structure of feature substitution in this setting. Workzone appears to be the strongest individual operational proxy, since removing it produced the largest decline among the proxy ablations. Priority, by contrast, contributed less to retained performance. The no-workzone-or-priority model is therefore especially important: it removes both a salient location/workflow proxy and a scheduling proxy, yet still preserves meaningful predictive capacity. This suggests that worker-relevant prediction is distributed across several operational fields rather than being reducible to one replacement variable.

Temporal validation produced the same qualitative pattern. Models trained on earlier task records and evaluated on later records continued to perform above chance. The direct-monitoring-excluded and operational/proxy-only models each achieved a temporal ROC-AUC of 0.781, while the no-workzone-or-priority model achieved 0.730. The employee-ID-only model achieved 0.649. The temporal setting is stricter than random cross-validation because it tests performance under temporal distribution shift. The persistence of predictive capacity under this setting supports the interpretation that the result is not merely an artefact of random-fold similarity between task records.

Permutation importance further clarifies which operational variables preserve predictive capacity. Figure 2 shows that workzone and task category are among the most important features in the reduced model, followed by quantity, priority, and task type. This pattern is consistent with the feature-substitution claim: predictive signal is preserved not through an explicit worker identifier, but through operational context variables that describe where, what, and how work is performed.

Taken together, the results support the central claim that removing direct worker-identifying features does not nec-

Feature setting	Best model	ROC-AUC	Retention	AUC drop	Features	Temporal AUC
Full non-leakage	HistGradientBoosting	0.822	1.000	0.000	6	–
Direct-monitoring excluded	HistGradientBoosting	0.802	0.976	0.020	5	0.781
Operational/proxy-only	HistGradientBoosting	0.802	0.976	0.020	5	0.781
No workzone proxy	HistGradientBoosting	0.745	0.906	0.077	4	–
No priority proxy	HistGradientBoosting	0.798	0.971	0.024	4	–
No workzone or priority	HistGradientBoosting	0.738	0.898	0.084	3	0.730
Employee ID only	HistGradientBoosting	0.699	0.850	0.123	1	0.649

Table 1: Predictive capacity retained across direct-feature and proxy-feature ablations. Retention is computed as ROC-AUC divided by the full non-leakage model’s ROC-AUC. Temporal AUC is reported for models evaluated using earlier task records for training and later task records for testing.

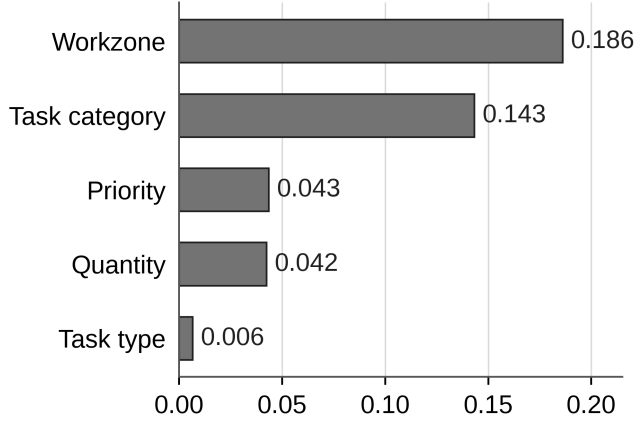


Figure 2: Permutation importance for the strongest reduced model. Importance is measured as the decrease in ROC-AUC when each feature is permuted.

essarily eliminate worker-relevant predictive capacity. The strongest evidence is the small performance drop from the full model to the direct-monitoring-excluded model, combined with the persistence of above-chance performance under stricter proxy ablations and temporal validation. The result is not that operational features are inherently inappropriate, nor that this dataset represents a deployed surveillance system. Rather, the result is that operational data can preserve the capacity for worker-relevant scoring even when direct identifiers are excluded.

Discussion

The results complicate an input-centred approach to governing algorithmic management. A model that excludes explicit worker-identifying fields may appear less intrusive, but our findings show that this does not necessarily remove worker-relevant predictive capacity. This distinction matters because algorithmic management systems often exercise control through allocation, ranking, recording, rating, and rewarding rather than through a single explicit act of surveillance (Kellogg, Valentine, and Christin 2020). In such systems, the governance question is not only whether an employee identifier or direct monitoring variable is present, but whether the model can still support decisions that affect

workers’ managerial attention.

The persistence of predictive capacity in the reduced models suggests that operational data can function as a form of indirect workplace visibility. Workzone, task category, quantity, priority, and task type are not facially personal attributes. They describe the organisation of work, where a task occurs, what kind of task it is, how much is involved, and how it is prioritised. Yet these variables preserve substantial signal about task duration. This aligns with scholarship on workplace datafication and electronic performance monitoring, which shows that managerial visibility can be produced not only through explicit surveillance tools but also through routine digital records generated in the course of work (Ajunwa, Crawford, and Schultz 2017; Ravid et al. 2020; Mateescu and Nguyen 2019). The key point is that operational data exhaust can become analytically worker-relevant even when it is collected for coordination, scheduling, inventory, or workflow management.

This finding also extends concerns about proxy variables beyond the conventional fairness-through-unawareness frame. In fairness research, removing a sensitive attribute does not necessarily prevent a model from reconstructing or acting on correlated information (Barocas and Selbst 2016; Dwork et al. 2012; Datta et al. 2017). In workplace settings, a similar mechanism can occur around monitoring rather than only demographic status. The relevant proxy may not be a protected attribute; it may be a structural feature of the job. A workzone, priority label, or task category can encode organisational conditions that remain predictive of worker-relevant outcomes. Feature substitution in algorithmic management therefore also involves the persistence of managerial scoring capacity through ordinary operational variables.

The robustness ablations help clarify the structure of this substitution. Removing workzone produces a noticeable decline in performance, indicating that location carries important predictive signal. However, performance remains meaningfully above chance even when both workzone and priority are removed. This matters because it weakens a narrow explanation in which the result is driven by one obvious proxy. Instead, the evidence suggests a more distributed form of substitutability: predictive capacity can be spread across multiple operational traces. For governance, this means that removing a single contested variable may not be enough if the model can recover similar functionality

through the remaining feature space.

The results also show why transparency alone is an incomplete accountability strategy. Input disclosure can tell workers, auditors, or regulators which variables a system uses, but it does not necessarily reveal what the model can still infer. Ananny and Crawford argue that transparency does not automatically produce understanding or control (Ananny and Crawford 2018). Our findings show that a system may transparently disclose that it does not use an employee identifier while still retaining the capacity to support worker-relevant prediction through task-context variables.

This has implications for algorithmic auditing. Existing accountability tools such as model cards, internal audits, and algorithmic impact assessments encourage documentation of intended uses, model performance, risks, and lifecycle controls (Mitchell et al. 2019; Raji et al. 2020; Costanza-Chock et al. 2023). These are important, but our results suggest that audits of workplace AI should also include counterfactual feature-removal tests. Instead of asking only whether a model contains direct monitoring fields, auditors should ask how much worker-relevant predictive capacity remains when those fields are removed, and how much remains after salient proxy variables are also excluded. Retained predictive capacity can therefore function as a diagnostic for feature substitution. It identifies when feature removal has failed to substantially reduce the model’s worker-relevant functionality.

This diagnostic is especially relevant for notice and contestability. In platform and digitally mediated work, information asymmetries already limit workers’ ability to understand how algorithmic systems shape their opportunities and evaluations (Rosenblat and Stark 2016; Wood et al. 2019). Feature substitution can deepen this problem because the system’s practical capacity is not obvious from the surface meaning of its inputs. A worker may be able to contest the use of direct monitoring data, but it is harder to contest a model that relies on apparently neutral variables.

The broader implication is that algorithmic management should be governed as a sociotechnical system, not merely as a set of model inputs. Operational variables are produced by workplace design, managerial choices, infrastructure, and historical patterns of task allocation. When these variables become predictive, they may reflect organisational structure as much as worker behaviour. This is why retained predictive capacity is a useful but limited audit signal as it shows that the model can still score a worker-relevant outcome, while leaving open the further institutional question of how that score should be constrained. The contribution of this paper is to make that diagnostic step explicit. Feature removal may reduce apparent observability, but governance should also ask whether the scoring function remains.

Limitations

This study has several limitations. First, the dataset is a retrospective warehouse task-execution dataset rather than a live algorithmic management system. The models studied here therefore do not directly represent an employer’s deployed scoring, discipline, or allocation system. The analysis should

instead be read as a mechanism study. It tests whether operational data can preserve worker-relevant predictive capacity after direct worker-identifying features are removed.

Second, the prediction target is task duration, not individual worker productivity. Task duration may reflect worker behaviour, but it may also reflect task complexity, warehouse layout, equipment availability, congestion, priority rules, batching decisions, or other organisational conditions. This limitation is important because worker-relevant prediction does not necessarily imply worker-caused variation. At the same time, this ambiguity is central to the governance problem as algorithmic management systems may act on outcomes that blend individual behaviour with structural workplace conditions.

Third, the feature categories used in the ablation design are researcher-defined. We classify employee identifiers as direct worker-identifying features and variables such as workzone, priority, category, quantity, and task type as operational proxy features. These distinctions are analytically useful, but they may not fully capture how a warehouse operator, worker, auditor, or regulator would understand the same variables in practice. Future work should validate feature categories with domain experts, workers, and organisational stakeholders.

Fourth, the analysis evaluates predictive persistence rather than causal effects, legal compliance, or distributive fairness. We do not claim that any particular feature causes longer task duration, nor do we evaluate whether model outputs disadvantage specific demographic groups. The dataset does not contain the demographic or organisational context needed for those claims. The contribution is narrower: retained predictive capacity is used as an audit signal for feature substitution.

Finally, this paper examines one dataset and one worker-relevant operational outcome. Feature substitution may behave differently across workplaces, industries, and prediction targets. Error prediction, absence prediction, productivity ranking, safety risk scoring, or shift allocation may involve different proxy structures. The results therefore motivate broader evaluation rather than establishing a universal pattern.

Conclusion

This paper examined whether worker-relevant predictive capacity persists after direct worker-identifying features are removed from algorithmic management models. Using a public warehouse task-execution dataset, we found that excluding explicit worker identifiers only modestly reduced predictive performance, while operational proxy variables retained substantial capacity to predict whether tasks would exceed median duration. Robustness ablations and temporal validation showed that this capacity was not solely attributable to a single obvious proxy.

The central implication is that feature removal should not be treated as sufficient evidence that a system no longer supports worker-relevant scoring. In AI-mediated work, operational data can preserve predictive functionality even when direct monitoring features are absent. We therefore propose retained predictive capacity as a practical audit signal for

feature substitution. Audits of algorithmic management systems should ask not only what features a model uses, but what the model can still infer after contested features are removed.

Ethical Statement

This work studies the governance risks of algorithmic management systems that use operational data to support worker-relevant prediction. The analysis uses a public, retrospective warehouse task-execution dataset and does not involve new data collection, intervention with workers, or identification of individuals. The paper does not claim that the dataset represents a deployed surveillance system, nor does it evaluate individual worker productivity. Task duration is treated as a worker-relevant operational outcome.

The main ethical concern addressed by the paper is that feature removal may create a misleading sense of protection if operational proxy variables preserve worker-relevant scoring capacity. A system may exclude explicit worker identifiers or direct monitoring fields while still retaining the functional ability to support allocation, ranking, or managerial attention. This raises concerns for transparency, notice, and contestability, because affected workers may not be able to infer a system's practical capabilities from a list of model inputs alone.

The findings could be used constructively to improve audits, impact assessments, procurement review, and organisational accountability for workplace AI systems. However, they could also be misused by organisations seeking to preserve scoring functionality while avoiding more visible/contestable monitoring features. For this reason, we frame retained predictive capacity as an audit signal rather than as a design recommendation. The purpose of the analysis is not to help employers build less visible scoring systems, but to help auditors, researchers, regulators, and worker advocates identify when feature removal has failed to meaningfully reduce worker-relevant predictive capacity.

We also recognise that the feature categories used in this paper are researcher-defined. The intended contribution is to support more careful evaluation of algorithmic management systems, especially where predictions may affect assignments, expectations, performance evaluation, or managerial intervention.

References

- Ajunwa, I. 2023. *The Quantified Worker: Law and Technology in the Modern Workplace*. Cambridge University Press.
- Ajunwa, I.; Crawford, K.; and Schultz, J. 2017. Limitless Worker Surveillance. *California Law Review*, 105: 735–776.
- Ananny, M.; and Crawford, K. 2018. Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability. *New Media & Society*, 20(3): 973–989.
- Barocas, S.; and Selbst, A. D. 2016. Big Data's Disparate Impact. *California Law Review*, 104: 671–732.
- Citron, D. K.; and Pasquale, F. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review*, 89: 1–33.
- Costanza-Chock, S.; Harvey, E.; Raji, I. D.; Czernuszenko, M.; and Buolamwini, J. 2023. Who Audits the Auditors? Recommendations from a Field Scan of the Algorithmic Auditing Ecosystem. arXiv:2310.02521.
- Datta, A.; Fredrikson, M.; Ko, G.; Mardziel, P.; and Sen, S. 2017. Proxy Non-Discrimination in Data-Driven Systems. arXiv:1707.08120.
- Delfanti, A. 2021. Machinic Dispossession and Augmented Despotism: Digital Work in an Amazon Warehouse. *New Media & Society*, 23(1): 39–55.
- Derkinderen, V.; Bekker, J.; and Smet, P. 2023a. Optimizing Workforce Allocation under Uncertain Activity Duration. *Computers & Industrial Engineering*, 179: 109228.
- Derkinderen, V.; Bekker, J.; and Smet, P. 2023b. Replication Data for: Optimizing Workforce Allocation under Uncertain Activity Duration. KU Leuven RDR, V1. <https://doi.org/10.48804/YHMU7R>.
- Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness Through Awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, 214–226.
- Kellogg, K. C.; Valentine, M. A.; and Christin, A. 2020. Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals*, 14(1): 366–410.
- Kroll, J. A.; Huey, J.; Barocas, S.; Felten, E. W.; Reidenberg, J. R.; Robinson, D. G.; and Yu, H. 2017. Accountable Algorithms. *University of Pennsylvania Law Review*, 165(3): 633–705.
- Lee, M. K.; Kusbit, D.; Metsky, E.; and Dabbish, L. 2015. Working with Machines: The Impact of Algorithmic and Data-Driven Management on Human Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 1603–1612.
- Mateescu, A.; and Nguyen, A. 2019. Explainer: Workplace Monitoring & Surveillance. Data & Society. <https://datasociety.net/research-library/explainer-workplace-monitoring-surveillance/>. Accessed: 2026-08-08.
- Mitchell, M.; Wu, S.; Zaldivar, A.; Barnes, P.; Vasserman, L.; Hutchinson, B.; Spitzer, E.; Raji, I. D.; and Gebru, T. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229.
- Möhlmann, M.; and Zalmanson, L. 2017. Hands on the Wheel: Navigating Algorithmic Management and Uber Drivers' Autonomy. In *ICIS 2017 Proceedings*. Paper 3. <https://aisel.aisnet.org/icis2017/DigitalPlatforms/Presentations/3/>.
- Newlands, G. 2021. Algorithmic Surveillance in the Gig Economy: The Organization of Work through Lefebvrian Conceived Space. *Organization Studies*, 42(5): 719–737.
- Nissenbaum, H. 2010. *Privacy in Context: Technology, Policy, and the Integrity of Social Life*. Stanford University Press.
- Prince, A. E. R.; and Schwarcz, D. 2020. Proxy Discrimination in the Age of Artificial Intelligence and Big Data. *Iowa Law Review*, 105: 1257–1318.

- Raji, I. D.; Smart, A.; White, R. N.; Mitchell, M.; Gebru, T.; Hutchinson, B.; Smith-Loud, J.; Theron, D.; and Barnes, P. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 33–44.
- Ravid, D. M.; Tomczak, D. L.; White, J. C.; and Behrend, T. S. 2020. EPM 20/20: A Review, Framework, and Research Agenda for Electronic Performance Monitoring. *Journal of Management*, 46(1): 100–126.
- Rosenblat, A.; and Stark, L. 2016. Algorithmic Labor and Information Asymmetries: A Case Study of Uber’s Drivers. *International Journal of Communication*, 10: 3758–3784.
- Selbst, A. D.; Boyd, D.; Friedler, S. A.; Venkatasubramanian, S.; and Vertesi, J. 2019. Fairness and Abstraction in Sociotechnical Systems. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 59–68.
- Wachter, S.; Mittelstadt, B.; and Russell, C. 2018. Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR. *Harvard Journal of Law & Technology*, 31(2): 841–887.
- Wood, A. J.; Graham, M.; Lehdonvirta, V.; and Hjorth, I. 2019. Good Gig, Bad Gig: Autonomy and Algorithmic Control in the Global Gig Economy. *Work, Employment and Society*, 33(1): 56–75.