Towards a Better Understanding of Misfit Through Explainable AI Techniques

Corine Boon, Erdinç Durak, and Ş. İlker Birbil

In the person-environment fit literature, originally, most research attention was devoted to fit, but the concept of misfit has received increased attention in recent years (De Cooman et al., 2019). Several studies have shown that misfit has negative consequences for individuals (e.g., Billsberry et al., 2023; Follmer et al., 2018). At the same time, not all misfits are problematic; for example, increased diversity in perspectives can be very beneficial (Doblhofer et al., 2019; Vogel et al., 2016). Overall, research shows that misfit is important, but the concept is 'far more complex and multifaceted' than has been studied so far (De Cooman et al., 2019, p. 649), and at the same time, misfit has been underresearched (De Cooman et al., 2019; Williamson & Perumal, 2021). So far, we lack a common understanding of what misfit is, whether different forms of misfit exist, and what factors determine misfit (Billsberry et al., 2023; Englert et al., 2024).

Researchers conceptualise person-environment fit and misfit in different ways. First, a distinction is made between perceived (mis)fit and calculated (or actual) (mis)fit (Kristof-Brown et al., 2023; Van Vianen, 2018). Perceived (mis)fit reflects a direct self-report of the degree of (mis)fit someone experiences (Kristof-Brown et al., 2023). In contrast, calculated (mis)fit is an indirect form of (mis)fit which focuses on the discrepancy between attributes of the person (P) and the environment (E) that are measured separately. These separate measures are then used to calculate (mis)fit (Kristof-Brown et al., 2023). Relatedly, Edwards et al. (2006) made a distinction between three approaches to study person-environment (mis)fit: molar, molecular, and atomistic. The molar approach is used to measure perceived (mis)fit, as it implies directly measuring the perceived fit between the person and the environment, usually by taking the average of, for example, three direct questions about fit. The molecular approach is a direct way of measuring the perceived discrepancy (or misfit) between the person and the environment, and the atomistic approach aims to assess calculated

C. Boon (✉) · E. Durak · Ş. İ. Birbil
University of Amsterdam, Amsterdam, Netherlands
e-mail: C.T.Boon@uva.nl

J. Billsberry and D. L. Talbot (eds.), *Employee Misfit*,
https://doi.org/10.1007/978-981-96-8208-9_12

(mis)fit and implies separate measurement of attributes of the person and the environment, and combining them in some way. In this chapter, we focus primarily on calculated misfit, using the atomistic approach. In addition, we use perceived misfit as an outcome, following the molar approach.

Recently, there has been increasing attention for the nature of misfit and how it is conceptually different from fit. For *calculated (mis)fit*, basic assumptions in the field of (mis)fit are that deficiency (person (P) > environment (E)) and excess (P < E) are similarly harmful for individuals, and that absolute levels of P and E do not affect outcomes (Edwards, 2008; Van Vianen, 2018). However, researchers have indicated that both assumptions are questionable as both the direction of misfit, and the absolute levels of P and E are likely to matter (Edwards, 2008; Van Vianen, 2018). This suggests that there may be multiple forms of misfit, each having different antecedents and outcomes (Billsberry et al., 2023). Also, we lack knowledge on whether misfit on some attributes have larger consequences than misfit on other attributes, and research into boundary conditions of fit and misfit relationships suggests that context affects whether (mis)fit matters (Kristof-Brown et al., 2023). This highlights the need for a more detailed approach to (mis)fit, and for context to be taken into account. For example, for newcomers, misfit may be more detrimental than for people with longer tenure in the organisation (Chi et al., 2020).

Perceived misfit can also be operationalised in different ways. In most studies, perceived misfit is measured as low levels of perceived fit. However, researchers argue that low levels of fit could indicate misfit, but does not necessarily do so (Sun & Billsberry, 2023). Instead, they recommend operationalising misfit as high levels of misfit or incongruence. In this chapter, we measure perceived misfit as low levels of perceived fit in line with most research in this area, but our proposed method is also applicable to other measures of misfit.

Several researchers have suggested that we need better methods to examine the complex nature of (mis)fit, and more attention should be given to the precise forms of (mis)fit and their consequences (De Cooman et al., 2019; Kristof-Brown et al., 2023; Sun & Billsberry, 2023). In this chapter, we propose the use of artificial intelligence (AI) methods as explorative techniques to address the complexities of misfit, help to detect misfit, and to build more specific theory on misfit, its antecedents, and effects. AI techniques allow for building and testing more complex and complete models (Tonidandel et al., 2018). Such techniques are not only suitable for analysing big data, but they can also be used in analysing the typically smaller datasets that are common in organisational research. Therefore, AI techniques are very suitable for exploring calculated misfit where personal and environmental attributes are measured separately and the discrepancy is calculated (Van Vianen, 2018) as well as for exploring the complex interplay of antecedents of (perceived) misfit.

The growing interest in AI and the increasing model complexities have triggered a new research area coined as explainable AI (XAI), where the main objective is to understand the inner workings of trained machine learning models (e.g., Biran & Cotton, 2017; Doran et al., 2017; Doshi-Velez & Kim, 2017; Gilpin et al., 2018; Hassija et al., 2024; Li et al., 2023; Linardatos et al., 2021; Minh et al., 2022). XAI can be used as a tool to help to uncover which forms of misfit exist, which variables

predict misfit, and under what circumstances misfits are (in)effective. Doing so, such models can be used to build and refine theory (Shrestha et al., 2021; Tonidandel et al., 2018). In this chapter, we aim to give an overview of the possibilities that XAI offers for studying misfit, and how researchers might use XAI to build theory that addresses the complexities inherent in misfit. We do not intend to be exhaustive, but instead, we illustrate different types of XAI approaches and show their potential contribution to misfit research.

To illustrate the different XAI approaches and how they can be applied to the study of misfit, we will use a dataset containing personal and organisational attributes (which can be used for calculated misfit), and we include perceived person-organisation misfit as a dependent variable. In our example, we use perceived fit data and assume that misfit is the opposite of fit. Taken together, we will explain and show through examples how XAI techniques can help to address the complexity of misfit and answer important theoretical questions about misfit which have been left unanswered to date. We created the following GitHub page where we provide the data and code to reproduce the empirical illustrations or to apply XAI approaches on other data: https://github.com/erdincd/Misfit-through-XAI.

1 A Brief Introduction to XAI

Explainable Artificial Intelligence (XAI) is a growing field in AI that aims to make AI models more understandable. AI models range from relatively simple models such as linear regression to very complex models such as neural networks. Generally speaking, it is important to find a balance between the transparency and accuracy (or explained variance) of an AI model, as pointed out by Rudin (2019). This implies that as an AI model becomes more complex and, thus, potentially more accurate, it also tends to become less transparent and harder to interpret. Conversely, a simpler, more transparent model might not capture all the nuances of the data, potentially sacrificing accuracy. For instance, linear regression or logistic regression models are known for their interpretability. By utilising the coefficients provided by these models, one can easily explain the relationship between the independent variables and the outcome, making them transparent and easy to understand for people who apply these models. However, the explained variance of linear models such as linear regression is generally relatively low. Also, such models often fail to capture complex patterns in a dataset, as there might be nonlinearities in the data (Hastie et al., 2009). In contrast, more complex AI models provide high accuracy, but at the same time they are less easy to interpret. The outcomes produced by these complex models are intricate and difficult to comprehend, making it more challenging to derive conclusions and apply the findings broadly.

Most of the approaches in XAI research fall into one of two categories: (1) the ones that aim at designing inherently interpretable (transparent or glass-box) but accurate models, (2) the ones that develop post-hoc *explainer* techniques that aim at

interpreting trained complex (opaque or black-box) models. While inherently interpretable models offer global explanations that provide an overall understanding of the model's behaviour across the entire decision space, most state-of-the-art explainer techniques can only provide local explanations. Local explanations are explanations that clarify the model's reasoning for individual predictions or specific data points rather than across the whole dataset. These local explanations are often used in practice because decision-makers can focus on a specific instance and gain insights into why the complex model made a particular decision for that instance.

In this chapter, we discuss two approaches from each one of these categories and show how these approaches can help to increase our understanding of misfit. From the first category, besides the commonly used logistic regression model, we also cover decision trees. From the second category, we use two approaches, Shapley additive explanations (SHAP) and counterfactual explanations, to explain a trained gradient boosting machine model which is an AI model renowned for its high accuracy. Below, we first explain the different ways in which XAI can help inform misfit research. This is followed by an empirical illustration using four XAI approaches.

2 Using XAI for Identifying Types of Misfit

As explained above, there are different ways in which XAI techniques can help to identify types of misfit. Here, we discuss three ways in which XAI techniques can contribute to misfit research: they can provide a detailed view of person-environment (PE) misfit, they can allow for nonlinear and asymmetrical effects, and they can allow for contextual factors to be included.

A detailed view of misfit. Much of the research on calculated PE misfit has examined the fit between a set of attributes of the person (e.g., personal values, competencies, skills, etc.) and a set of attributes of the environment (e.g., organisational values, organisational goals, needs, and job competencies). In most cases, these attributes are collapsed into one variable reflecting the person, and one reflecting the environment, or studies use subdimensions of attributes and calculate misfit for each of these dimensions separately (for example, reviews and meta-analyses on calculated PE fit by Sun & Billsberry, 2023; Van Vianen, 2018; Yang et al., 2008). On the one hand, collapsing individual attributes to one or a few overall variables reduces complexity, and helps to interpret the results. However, on the other hand, collapsing individual attributes means a loss of information, and does not address possible differences and complex interactions among individual attributes. Misfit on some competencies, needs, or values may have larger consequences than misfit on other competencies, needs, or values, and for some attributes, achieving fit is perceived as a necessary condition to fit well, whereas for other attributes, a misfit might not affect a person that much. For example, someone might care more about results orientation than about honesty, as a result of which they might be more negatively affected by a misfit on results orientation than a misfit on honesty values.

XAI techniques provide the opportunity to include a large number of attributes (i.e., items or features) of P and E in one analysis and help to explore which attributes among this large set are most important in explaining the outcome. Thus, using such techniques can help select the most relevant features from a large set of features, which will address the complexity inherent in misfit, and at the same time it will help in enhancing interpretation and in detecting patterns in the results.

Nonlinear and asymmetrical effects. One of the implicit assumptions of theory on (mis)fit is that absolute levels of personal and environmental attributes do not matter, and instead deficiency (person > environment, or *nongain*) and excess (person < environment, or *loss*) are more salient to people. Empirical research shows that the level and direction of misfit matters, however, misfit may be far more complex (e.g., effects might be nonlinear and/or asymmetrical). For example, building on the idea that consequences of deficiency and excess are likely not to be symmetrical, it might be that the consequences of deficiency versus excess depend on the type of attributes, or there may be other forms of nonlinear relationships, where deficiency or excess only start having certain consequences after a certain threshold level. To address this, XAI techniques can explore complex interactions and hint at nonlinear relationships between attributes of the person and the environment. This can help to uncover different forms of misfit and to build theory on the differential consequences of these forms of misfit.

Contextual factors. Research has suggested that the effect sizes of fit and misfit differ across studies and contexts, which suggests that contextual factors are likely to affect the consequences of misfit (Edwards, 2008; Kristof-Brown et al., 2023). Including contextual factors at different levels, such as country, industry, career stage, work experience, tenure, and personality types, could provide a more accurate view of how these factors affect misfit. So far, contextual factors have been included in studies as either control variables or moderators, however, that does not account for possible nonlinear or other types of influences of contextual factors. XAI techniques help in identifying those contextual factors that matter for misfit, as well as how and at what levels they matter.

Below we illustrate how logistic regression, decision trees, SHAP, and counterfactual explanation techniques can be used to provide a detailed view of misfit, for identifying types of misfit and to explore contextual factors that play a role in misfit, and we discuss the implications of these techniques and their results.

3 An Empirical Illustration

Below, we present different XAI approaches for exploring misfit. To illustrate these approaches, we focus on calculated person-organisation (PO) misfit (and more specifically, value incongruence), and we use perceived misfit as an outcome. Our example illustrates two different ways in which researchers might want to explore misfit: first, the input variables consist of a set of personal and organisational attributes, which

enable exploration of calculated misfit (which combination of personal and organisational attributes affects the outcome, and how). Second, we use perceived misfit as an outcome, which enables prediction of misfit, using a set of input variables.

We used a survey-based dataset (n = 249). A cross-sectional questionnaire was sent to 696 employees and their managers working in four health providers in the Netherlands, and 290 employees (a response rate of 42%) working in 42 teams responded to the survey. The number of respondents per organisation is well-distributed (60–80 employees per organisation). All responses were fully confidential. We removed 41 respondents (employees) who had missing data, resulting in a final sample of 249 employees. 88% of 249 respondents were female, 8% were male, and 4% did not answer. The average age of the employees is 40.5 years (SD = 10.8), and their average tenure is 7.3 years (SD = 6.6). 30% of the employees work for 20 h or less per week, and 37% work for four days or more per week. We used the following measures:

3.1 Calculated PO Misfit

As a basis for the calculated fit analyses, we measured (1) a set of values rated by the person, (2) the same set of values for the organisation, based on Ostroff et al. (2005). The personal and organisational values were measured using a five-point Likert-type scale with response options ranging from 'strongly disagree' (1) to 'strongly agree' (5).

The values are as follows:

- Team oriented
- Reputation
- Improvement
- Information sharing
- Professionalism
- Self-directed
- Supportive
- Client convenience
- Initiative
- Flexibility
- Client service
- Results focus
- Adaptability
- Honesty
- Responsibility
- Innovation
- Integrity
- Performance

3.2 Perceived PO Misfit

We measured perceived misfit using Cable and DeRue's (2002) three-item measure for PO fit defined as value congruence. This measure was designed to measure fit rather than misfit. For this illustrative example, we interpreted low fit as misfit. To be able to apply prediction models, we converted the perceived misfit measure into a dichotomous measure; if the perceived PO fit value was less than or equal to 3, we coded it as perceived misfit (1), and when it was greater than 3, we coded it as perceived fit (0). This serves the purpose of our illustrative example, but researchers might find other ways in which they label people as misfits or fits.

3.3 Tenure

We included tenure as a contextual variable, measured as the number of years the respondents have worked at their current organisation.

4 Models

Below, we first build two inherently interpretable models (logistic regression and decision trees) on the dataset and interpret the results. Then, we build one complex model (gradient boosting machine) and use two tools to interpret this complex model: SHAP and Counterfactual Explanations.

4.1 Logistic Regression

Logistic regression is an inherently explainable binary classification approach (Hastie et al., 2009). Like the well-known linear regression, logistic regression uses a linear combination of input features (independent variables; e.g., attributes of the person and the organisation) to obtain an outcome showing the probability of the selected sample being in one of the two classes: perceived misfit or perceived fit.

Figure 1 shows the results. The coefficient of each of the values is depicted for the person (in dark grey) and for the organisation (in light grey), and the likelihood of perceiving a misfit is used as the outcome. A negative coefficient means that the corresponding feature is associated with lower likelihood of being a misfit. For example, the personal value 'results focus' has the strongest negative association with misfit, so a 'results focus' of a person lowers misfit. In contrast, a positive coefficient implies that the corresponding feature is associated with a higher likelihood of being a misfit. When looking at the coefficients of both personal (P) and organisational (O)

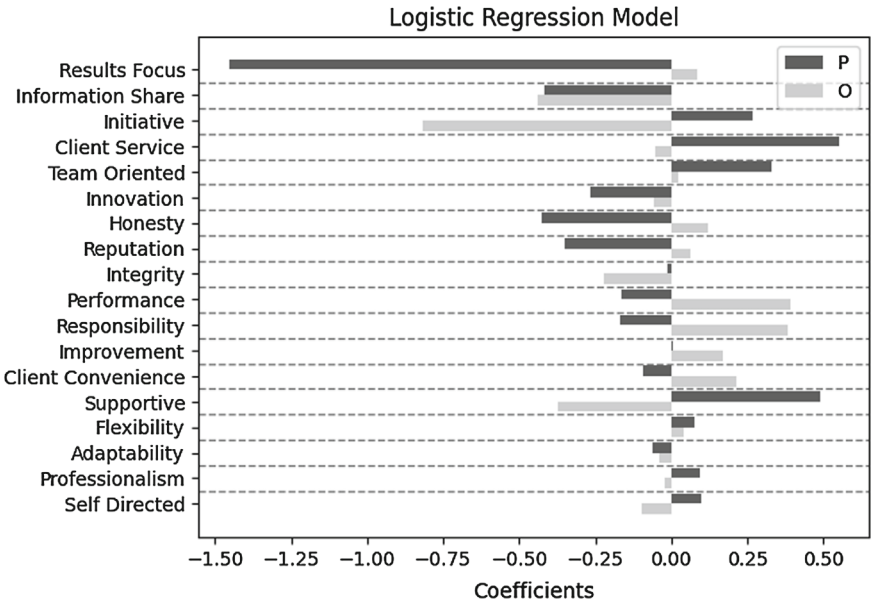

Fig. 1 Logistic regression model coefficients

values, one can draw conclusions about how calculated fit relates to perceived misfit. For example, Fig. 1 shows that for initaitive, client service, supportive, and self-directedness, P has a positive, but O has a negative coefficient. So when people value these *more* than their organisation, they are more likely to be a misfit. When they value results focus, honesty, reputation, performance, and individual responsibility *less* than their organisation, they are also more likely to be a misfit. At the same time, some values, such as flexibility, adaptability, and professionalism are not strongly related to misfit perceptions. Overall, these results show that there is a (large) difference between the different personal and organisational values in predicting misfit, showing that a detailed view—including individual attributes rather than an aggregate measure—provides additional insight into the nature and antecedents of misfit.

4.2 Decision Trees

Another approach to obtain inherently interpretable models is based on decision trees. A decision tree is a predictive model that maps input features to an outcome by using a tree structure. This tree consists of nodes representing decisions based on values of the features, and branches representing the resulting splits of the samples. At each node, the model evaluates all possible splits across all features and selects the one that best separates the data. This process is repeated recursively, splitting the data into increasingly homogeneous subsets until a stopping criterion is met, such

as a minimum number of observations per node. The leaves of the tree contain the final predictions, or the value of the outcome. This way, decision trees can be useful in misfit research by revealing important factors, providing per-sample rules, and revealing combinations of attributes that relate to an outcome, such as a combination of person and environment attributes that represent misfit. In addition, the visual nature of decision trees allows for a straightforward interpretation and communication of complex decision-making processes. Decision trees can vary in depth, from shallow trees representing only a few decisions to deep trees that include a large number of decisions. The interpretability of the decision trees can thus be adjusted by changing their depth level, as shallow trees are more interpretable than deep trees.

We constructed a decision tree on the dataset, including the set of personal and organisational values listed above, predicting misfit. Figure 2 presents the results. At the top of the figure, the first node—also called the root node—is represented (results focus of the person). Below this root node, the decision tree model results indicate that there is a 'split' (indicated by the arrows) into P-results focus of smaller than 3.5 (the left side of the figure) and larger than 3.5 (the right side of the figure). This means that when explaining whether someone perceives a misfit or not, two groups can be distinguished: one group which scores lower than 3.5 on results focus, and another group that scores higher than 3.5 on results focus. Next, for the group with a results focus score of lower than 3.5, the organisational value being supportive is the next important factor for determining whether someone perceives a misfit or not. As can be seen in Fig. 2, the decision tree results show that the organisational value being supportive is split into smaller than, and larger than 3.5.[1] At each of the decision nodes, we included bar charts. These bar charts show the distribution of the feature values (e.g., how many respondents scored 1, 2, 3, 4, and 5 on P-results focus, and how many of these are labelled as misfit). At the bottom of the figure (the leaf nodes), the percentage of fit and misfit employees is shown by pie charts. For example, at the bottom left of Fig. 2, you see the respondents that score lower than 3.5 on (P) results focus, their organisations score lower than 3.5 on (O) support and lower than 2.5 on (O) flexibility. The majority of the people grouped in the leaf node are labelled as misfits, therefore, any individual falling into this leaf node is classified as misfit with a probability that is equal to the proportion of misfits, that is, 4/5. On the other hand, at the bottom right there is a larger group ($n = 147$) of employees of which most perceive a fit. This group is characterised by a results focus (P) of higher than 3.5, their organisations score higher than 2.5 on (O) innovation, they score higher than 3.5 on (P) innovation, and higher than 3.5 on (P) reputation.

It is also possible to include contextual variables in XAI models as input variables. To consider the contextual variable tenure, we ran another decision tree model including employee tenure (in years) as a contextual variable. Figure 3 shows that this new model is similar to the decision tree presented in Fig. 2, except for tenure appearing in the last split on the right-hand side. By doing so, we obtain two groups which are both labelled as fit but the one with lower tenure (less than 1.25 year)

[1] The value 3.5 comes back more often in these results, but this is a coincidence, and it emerged from the data.

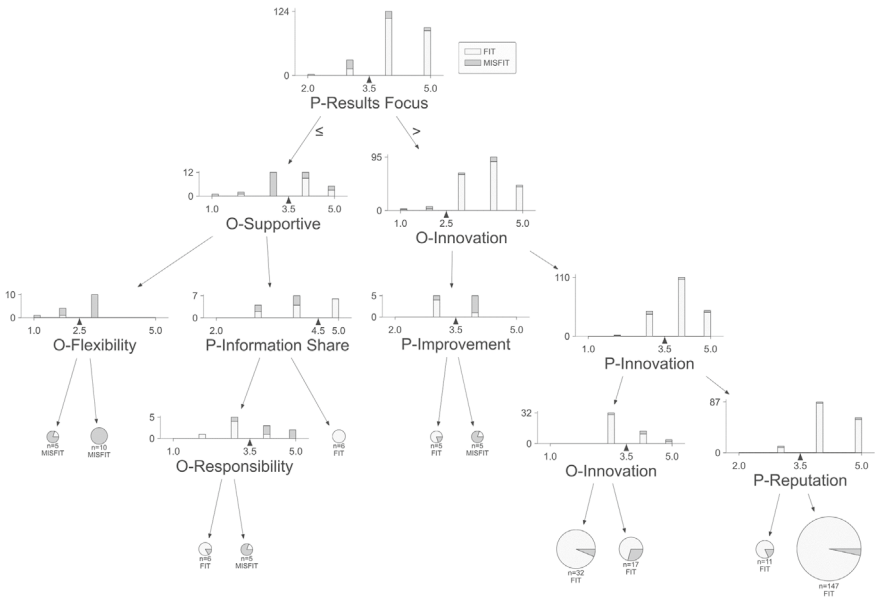

Fig. 2 Decision tree results

has a greater share of misfits. The results thus show for which employees' tenure matters in predicting misfit. In this example, tenure matters only for those employees who score higher than 3.5 on results focus (P), higher than 2.5 on organisational (O) innovation, and higher than 3.5 on personal (P) innovation.

Benefits. There are several benefits of using decision tree models. First, the results of the decision tree model show which features are the most important drivers of being a misfit. Similar to the logistic regression results, the results show that the first split, or the most important feature that determines whether someone is a misfit or not, is a person's results focus. Second, beyond the logistic regression results, the results of the decision tree also give insight into conditional effects, for example, a person's improvement value is only relevant when the results orientation value is high and the organisation's innovation value is low. For the other groups (e.g., that score low on results focus), the improvement value does not matter for their perception of misfit. In a decision tree, several values that are less important for explaining the outcome might not show up, particularly if the depth is limited. As the values of the person and the organisation are included as separate features in the model, this means that for some values only the personal value, or only the organisational value, is included (as is the case in our example), showing it might not be fit or misfit that is important in explaining the outcome, but the organisational or personal value itself. Instead, if both a personal and an organisational value show up in the same branch (e.g., innovation value of the person and of the organisation in our empirical example), it is possible to draw conclusions about calculated misfit. For example, Fig. 2 shows that if P-innovation is lower than 3.5, and O-innovation is higher than 3.5 (thus a type

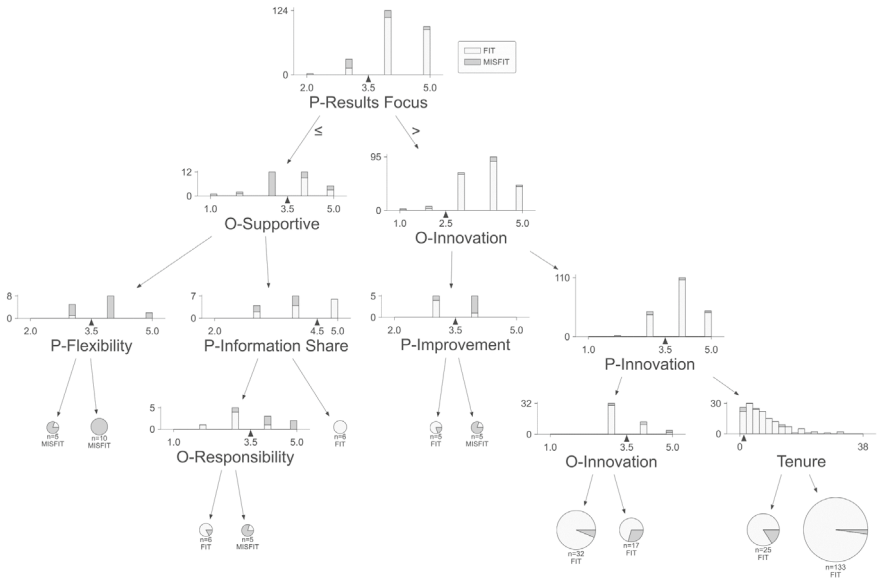

Fig. 3 Decision tree results including tenure as a contextual variable

of excess misfit), the chance of perceiving a misfit is higher than when O-innovation is between 2.5 and 3.5, and P-innovation is lower than 3.5 (indicating a fit). This way, you can detect which types of misfit matter in explaining the outcome.

A third benefit of decision trees is that the splits are based on a concrete value (e.g., 2.5 or 3.5), which can be used to identify and describe different subgroups of people who show similar characteristics. For each of the leaf nodes, a description of scores for the corresponding group can be given. Consider, for instance, the second-left leaf node in Fig. 2. The group in this node scores lower than 3.5 on P-results focus, lower than 3.5 on O-supportive, and higher than 3.5 on P-flexibility. This group consists only of misfits, implying that this combination of values relates to misfit.

Lastly, contextual variables can also be included in the model. This way it is possible to go beyond 'only' controlling for demographics or contextual variables to identify groups of people with certain characteristics (e.g., tenure) that show a different pattern of results, similar to conditional effects but allowing for nonlinearity.

Here, we presented decision trees with depth 4 but deeper and shallower trees are also possible. This usually comes with a trade-off: deeper trees tend to have high accuracy, especially for the training set. However, they are more difficult to explain. One possibility is to produce a set of decision trees with different depth levels and compare the results. Another option is to set a minimum number of data points for each node. In the example, we set the minimum to 5, which helps to identify (small) groups rather than focusing on individuals. However, depending on the aim, setting a lower minimum could help to address unique differences between individual (misfit) employees.

4.3 SHAP Explanations (of Gradient Boosting Machine)

Next, we used a gradient boosting machine (GBM) model and used SHAP to interpret the model. GBM is an ensemble approach that combines the predictions of multiple weak learners (usually shallow decision trees) to form a strong predictive model. In essence, a GBM model learns to improve its predictions by focusing on the mistakes made in earlier steps. While the results of a GBM model are less interpretable than a decision tree model, the opaque model can capture complex patterns, which can be used for obtaining highly accurate results. To be able to interpret the results obtained with GBM, we can use explainer techniques which elucidate the GBM model's reasoning behind individual predictions. Here, we employ two approaches: SHAP (this section) and Counterfactual Explanation (CE; the next section).

SHAP provides insights into complex machine learning models by offering SHAP values for each feature that is included in the model. These values, ranging from -1 to $+1$, denote the contribution of each feature to the model's prediction. They are calculated by comparing the change in prediction when a feature is included versus excluded, considering all possible combinations. The sum of SHAP values gives the precise predicted value, illustrating the total influence of all features on the outcome. In a binary classification model, like one that distinguishes between misfit and fit, this predicted value represents the probability of belonging to one of the classes. Below we illustrate SHAP.

First, we depict the SHAP values for a particular employee. To do so, we selected one employee who perceives themselves to be a misfit, and for whom GBM also predicted them to belong to the misfit category rather than fit. Table 1 shows the personal and organisational values of this employee, and Fig. 4 reflects the SHAP values that show the influence each feature has in predicting them to be a misfit. Figure 4 shows the nine values having the most significant influence on the misfit prediction of this individual, with the personal value results focus and the organisation's support value having the strongest impact. Only for adaptability do both the personal and the organisational values appear in the results, which means that only for adaptability does (calculated) fit matter for explaining misfit. For the other values, it is the personal or organisational values themselves, and not the misfit that matters. Looking at misfit, Fig. 4 shows that the SHAP value for the personal value adaptability is $+0.03$, and the SHAP value for the organisational value adaptability is -0.02. This implies that for this particular employee, having a personal value of adaptability which is higher than the organisational value of adaptability has negative consequences (i.e., increases the likelihood of misfit).

Figure 5 shows the SHAP values for all employees in the dataset where each circle corresponds to a contribution of the corresponding feature for an observation (employee). Please note that Fig. 5 only presents the 20 features that have the highest contribution, so not all personal and organisational values are represented in the results. The colours of the dots represent the score on each of the features (which varies between 1 and 5). For example, in the first row (results focus of the person), the light grey dots at the right represent a low value (i.e., the person has a low results

Table 1 Personal and organisational values of a misfit employee

Values	Person	Organisation
Team Oriented	4	3
Information Sharing	4	4
Supportive	4	3
Flexibility	3	3
Adaptability	3	2
Innovation	4	3
Reputation	4	3
Professionalism	4	4
Client Convenience	4	4
Client Service	4	3
Honesty	4	4
Integrity	4	4
Improvement	3	3
Self-directed	4	3
Initiative	4	3
Results Focus	3	5
Responsibility	4	4
Performance	4	5

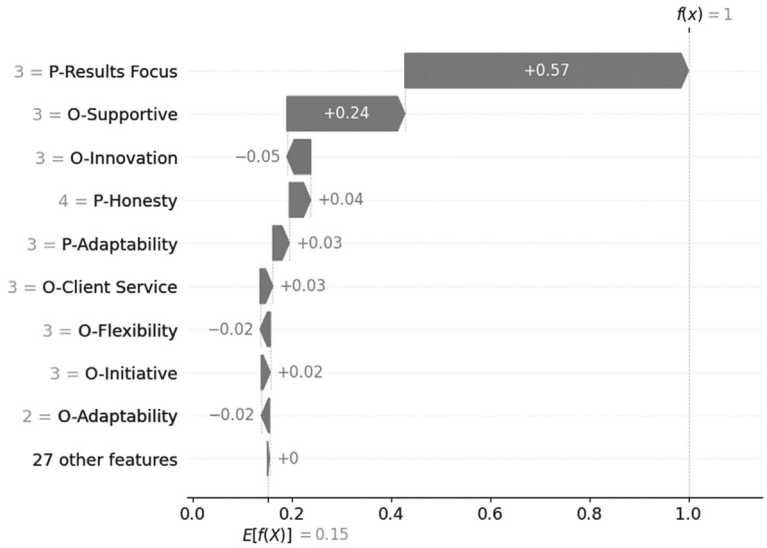

Fig. 4 SHAP values for a misfit employee

focus), and the dot's position on the X-axis is almost 0.5, which implies that it contributed with 0.5 to predict this employee as misfit.

Figure 5 shows that in line with the decision tree results, P-results focus, and O-innovation and support are the features with highest impact on the prediction of misfit. Again, we see that several values appear only once, which shows that only the personal or organisational value matters in explaining misfit. In addition, we find that some pairs of values, such as personal and organisational innovation, information sharing, and adaptability show up as important features, so for these values, calculated

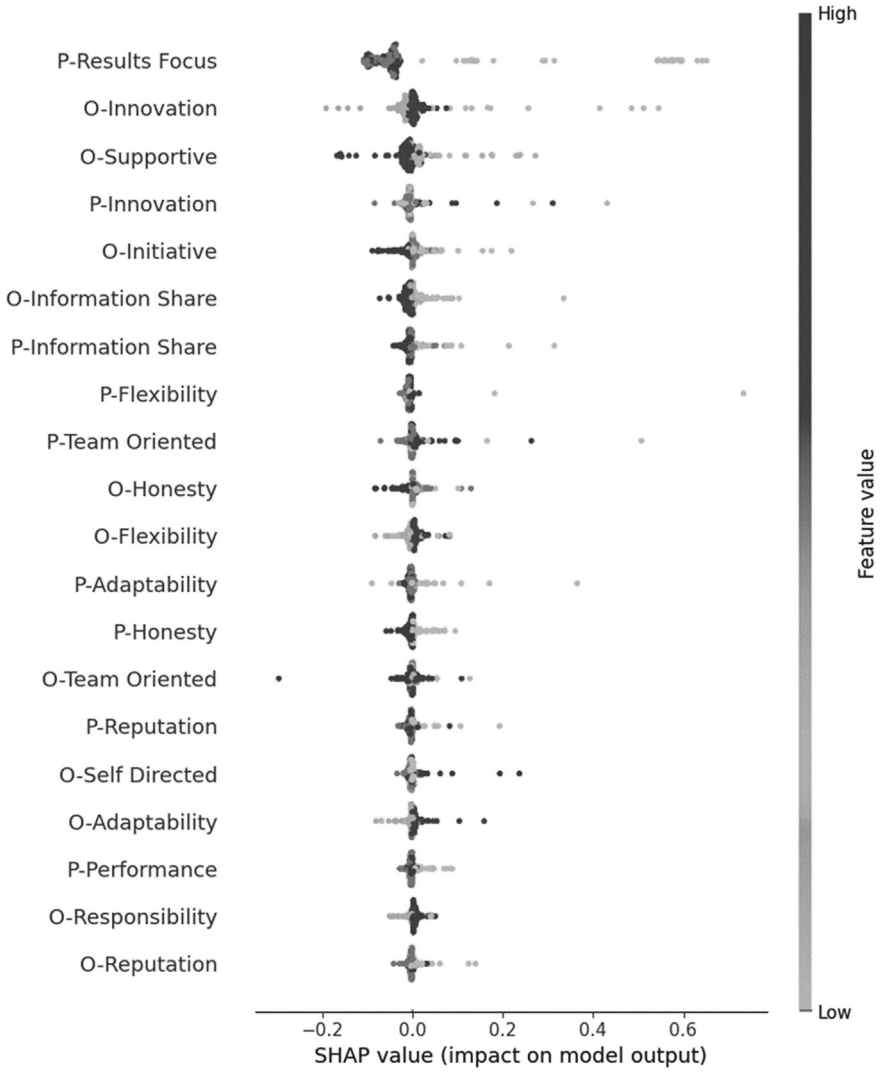

Fig. 5 Gradient Boosting Machine results (explained by SHAP)

misfit matters in explaining perceived misfit. For example, Fig. 5 shows, with this illustrative data, that when employees score low on adaptability (light grey dots on P-adaptability—on the right side) they are more likely to be classified as a misfit. However, if they score high on their organisation's adaptability (O-adaptability), employees are also more likely to be classified as a misfit. Taken together, we may conclude that when employees value adaptability less than their organisation, they are more likely to perceive that they are a misfit. Comparing Figs. 4 and 5, the most impactful values such as a person's results focus and the organisation's support value are found to be (most) important both in the complete sample, and for the specific individual. However, the organisational value client service does not occur in Fig. 5, but does occur in Fig. 4. This implies that there are individual differences in the values that impact someone's misfit perception.

Benefits. There are several benefits of SHAP explanations. First, since SHAP can be employed on AI model results of varying complexity, it enables us to work with models (such as GBM) whose accuracy is typically higher than logistic regression and decision trees, which can explain more variance in misfit than other models. Second, the results show which features are the most important drivers of being in the misfit category, in line with the logistic regression and decision tree approaches. Beyond the decision tree results, SHAP shows a larger number of features (in our example, personal and organisational values) and their impact on the prediction of the outcome (in this case, misfit). Third, the possibility of looking more closely at specific individuals is useful, particularly in the misfit domain. As misfits are relatively underresearched, (De Cooman et al., 2019; Williamson & Perumal, 2021), closer inspection of relatively rare cases can help to explore whether patterns occur, and whether different clusters or types of misfit can be found. For example, SHAP values of different employees can be plotted to compare their results and to help detect patterns that are not visible when looking at the results for the full sample.

4.4 Counterfactual Explanations (by Employing Gradient Boosting Machine)

The final approach we illustrate is counterfactual explanations. As a different approach to interpreting the results of GBM, counterfactual explanations give insight in what change in a value is necessary to change someone from a misfit to a fit, or the other way around. For example, in a scenario where an employee is classified as misfit by a trained GBM model, this approach can provide the minimum change in their values to be classified as fit instead of a misfit. To illustrate this approach, we selected the same employee as above in Table 1 and Fig. 4. Table 2 shows two alternative counterfactual explanations, indicating the minimum change in values to change a classification from a misfit to a fit. For this particular employee, if their results focus would change from 3 to 3.5, this employee would change from a perceived misfit to a fit. Similarly, if their organisation's support value would increase from 3 to 3.5, the

Table 2 Counterfactual explanations for a misfit employee

	Person	Organisation
Team Oriented	4	3
Information Sharing	4	4
Supportive	4	**3 --> 3.5**
Flexibility	3	3
Adaptability	3	2
Innovation	4	3
Reputation	4	3
Professionalism	4	4
Client Convenience	4	4
Client Service	4	3
Honesty	4	4
Integrity	4	4
Improvement	3	3
Self-directed	4	3
Initiative	4	3
Results Focus	**3 --> 3.5**	5
Responsibility	4	4
Performance	4	5

prediction would change from misfit to fit. This suggests that only a relatively small change in some of the personal or organisational values could make a difference between someone perceiving as a misfit versus a fit. The counterfactual explanations also show that for the other person and organisation values, small changes will not change the prediction, so it provides insight in which of the person and organisation values can make a difference in being a misfit or fit.

Benefits. The use of counterfactual explanations can bring the following benefits. It helps to identify how close specific misfit individuals are to being a fit, and which of the personal and organisational values make a difference between being a misfit or a fit. This helps to prioritise certain values over others in trying to reduce misfit, and it helps to gain insight in how much change in a feature is needed in order to turn a misfit into a fit. Similarly, counterfactual explanations can also be applied to address the opposite question: what changes are required to switch someone's classification from a fit into a misfit? Third, looking at counterfactuals for several individuals, it is possible on the one hand to detect patterns, and on the other hand, to gain insights into which factors seem to be unique to each individual misfit.

When using this method, it is also possible to 'fix' some of the feature values, meaning that they are set not to change. This can for example be useful when exploring what changes in organisational values are needed. In that case, all personal values could be fixed, in order to see what the results of the counterfactual explanations suggest in terms of changes only in organisational values to change from a

misfit to a fit. Another possibility is to take a stepwise approach, requesting only a change in one feature per step. By doing so, features that will and will not cause a change from misfit to fit can be identified.

5 Discussion: The Theoretical Value of XAI Approaches for Misfit

In this chapter, we explained how XAI can be used as an exploratory approach to building and refining theory on misfit. XAI approaches can be used to detect complex patterns that have not been examined before, and therefore help to build theory that can be tested in future studies. Through an illustrative example, we showed the different ways in which different XAI approaches can increase our insight into the nature of misfit, its antecedents and consequences. Through our example, we highlighted both the potential of XAI for exploring calculated misfit, and for exploring the prediction of misfit as a (dichotomous) outcome. Below, we will discuss different ways in which XAI approaches can have theoretical value for the misfit literature.

5.1 The Nature of Misfit

First, XAI approaches can help explore the nature of calculated misfit. In our example, for calculated misfit, we have taken an approach which is free from assumptions about the nature of misfit: we did not calculate any difference scores or other measures of misfit, but instead, we used organisational and personal values separately. This approach makes it possible to explore if and how personal and organisational values interact in explaining the outcome.

In previous research, personal and environmental attributes were often collapsed into one overall construct (e.g., personal values and organisational values), or a few sub-factors (e.g., rational goal values, open system values) (e.g., Ostroff et al., 2005). XAI approaches can include all attributes separately, even if they correlate highly, which is often the case when data is collected through surveys. Also, XAI approaches can address complex interactions that might occur and might not have been visible when traditional methods were used. For example, our results show that across the different XAI approaches, a few values seem to be driving the effect, and others are less important for explaining the outcome. For example, when one value consistently shows higher coefficients in the results across the different XAI approaches, this suggests that this value, or misfit on this value, plays an important role in explaining results. Other values might not show up in the results of any of the models. This indicates that they play a less important role in explaining results. These findings can be used to build theory on which specific characteristics or values

drive the effect and why, which goes beyond previous research that typically does not examine fit in such a detailed way due to lack of theory and/or statistical challenges.

Our example also shows that it is not always fit or misfit that matters. In some cases, the presence of personal or organisational values helps to explain the outcomes more than the fit or misfit *per se*. As the four XAI approaches include personal and organisational values separately, the results show that in some cases both personal and organisational values play a role in explaining the outcome, which means that conclusions can be drawn about misfit. In other cases, however, only the personal value or the organisational value helps to explain the outcome, which implies that it is not (mis)fit that matters, but rather the presence of either personal or organisational values.

In exploring the specific nature of misfit, decision trees can be particularly useful. One of the benefits of decision trees is that it is possible to identify clusters of people who, for example, have a high score on a personal value, and a low score on an organisational value, representing a specific type or magnitude of misfit. The decision tree results are very specific and will give specific values above or below which someone ends up in a certain cluster. That way, we can not only draw conclusions about what happens in case of higher or lower misfit, but we can also identify specific 'turning points' or critical values at which misfit starts to become a problem (e.g., only when the personal value is higher than X and the organisational value lower than Y, will it have consequences for the outcome), going beyond linear and quadratic relationships to include other nonlinear or asymmetrical options. Another way to identify critical values is the counterfactual explanations, which identify what change in personal or organisational values changes the prediction from misfit to fit. That shows for example, that reducing a misfit for some values matters more for the outcome than reducing a misfit on other values (see, for example, Table 3, where the counterfactual explanations indicated a reduced misfit for the values of support and results focus).

Thus, these approaches shed light on the complexity of misfit and help to address questions such as:

For which personal and environmental attributes does misfit matter?

What is the nature of misfit?

Can we identify types of misfit?

Which types of misfit are important and which ones are less important in determining the outcome of interest?

5.2 Misfit as an Outcome

We also looked at misfit as an outcome and focused on the question of which factors predict misfit. To enable the prediction of misfit, we took an approach that is different from most studies in the misfit area. For the purpose of our illustrative example, we used a survey-based measure of perceived fit, and based on the scores, we labelled individuals as perceived misfit or perceived fit. There are limitations to this approach.

However, it is also possible to use other measures or ways to classify individuals as a misfit, such as asking people directly whether they perceive themselves to be a misfit. The four methods we propose apply to any data in which individuals are labelled as misfit or fit, or in other words, put into a 'misfit' or 'fit' category. We would recommend considering different approaches and comparing the results, which will help to detect patterns and enable theory building.

The different XAI approaches show that a combination of factors predicts misfit. Both decision trees and GBM took into account complex interactions between factors and provided insight into which factors were the main drivers between being in the misfit versus the fit category. In addition, counterfactual explanations give insight into what changes in the predictors are needed to turn a misfit into a fit, or the other way around. For example, counterfactual explanations help to shed light on for which of the organisational and personal values a change matters for being a misfit or fit, and for which of the values, a change does not make a difference. Identifying main drivers in determining whether someone is a misfit, might allow the exploration of questions such as which factors are basic determinants for being a misfit, and which factors make the difference between being a misfit or a fit.

5.3 Addressing the Uniqueness of Misfit

So far, misfits have been studied less often (De Cooman et al., 2019; Williamson & Perumal, 2021). So there is still a lot unknown about the nature of misfits, while at the same time there might be more variation within the misfits category than the fit category, as the person score can be higher or lower than the environment score, and the magnitude of the difference can differ too. As Tolstoy (1878) wrote in Anna Karenina: 'Happy families are all alike; every unhappy family is unhappy in its own way'. The literature and our empirical illustration suggest that this might also be true for misfit. To take this into account, examining specific individuals through, for example, GBM and counterfactuals and detecting patterns can help in shedding light on which factors apply to a broad group of individuals that are categorised as misfits, and which factors are more unique to some specific individuals that are categorised as misfits. In line with this, we recommend running analyses not only on the complete dataset, but also taking a closer look at a selection of misfit individuals, to find patterns in the results, addressing both the similarities and the uniqueness of misfits. This will help to build and refine theory on types of misfit and their characteristics.

5.4 Contextual Factors

So far, contextual factors have been included in quantitative studies of fit and misfit mostly as control variables or as moderators. Using XAI approaches makes it possible to include more contextual factors if needed, and these approaches provide a more

detailed view on if and when such factors are impacting misfit. For example, we showed that tenure was only relevant for some employees (with a certain value profile) but not for others. One of the benefits of incorporating contextual variables and exploring their role is that it helps to identify groups of employees with similar characteristics for whom misfit plays a different role. For example, that for people with tenure shorter than a certain number of years, misfit might have a different effect than for people with longer tenure. This gives the opportunity to build more specific theory and to give more targeted practical insights in misfits that are more relevant or more detrimental for some groups than for others. Researchers might use this to explore the impact of different contextual factors, such as tenure, age, educational background, profession, organisational unit, sector, country, etc., to answer questions such as:

Do types of misfit differ depending on these contextual factors?

Does misfit impact some people more than others (e.g., are newcomers or people in certain professions more impacted by misfit)?

Do contextual factors help to predict misfit?

5.5 Comparison of the Models

The four models employed in this study—logistic regression, decision trees, and gradient boosting machines (with SHAP and counterfactual explanations)—vary significantly in terms of complexity. As model complexity increases, from logistic regression to gradient boosting machines, there is an associated increase in predictive power and the ability to capture complex nonlinear relationships. However, this comes at the cost of longer training times and reduced interpretability. For instance, while logistic regression is straightforward to interpret and computationally efficient, it is constrained by its linearity assumption. In contrast, decision trees and gradient boosting machines do not impose such assumptions on the data. However, these models require careful tuning of multiple hyperparameters (e.g., maximum depth, minimum number of observations per node) to achieve optimal performance. Despite their advantages, the results of gradient boosting machines necessitate additional approaches (e.g., SHAP and Counterfactual Explanations) to be interpreted. These techniques, while helpful, are insufficient to provide a comprehensive understanding of the model's behaviour across the whole dataset.

Accordingly, the choice of the appropriate model should be guided by the nature of the research questions or aims, and/or the nature of the dataset. Due to its simplicity and transparency, logistic regression is the best choice for smaller datasets where interpretability is crucial. This model is particularly suitable when the relationship between variables is expected to be linear. Gradient boosting machines are recommended in case of large and complex datasets that exhibit significant nonlinear relationships, and if understanding the model's behaviour across the whole dataset is not crucial, but rather if the researcher is interested in exploring specific individuals or cases (or in other words, local explanations). In these cases, for a better understanding

of a gradient boosting machine model's behaviour, we recommend using both SHAP and counterfactual explanations. On the other hand, if it is necessary to understand how the model works on the entire dataset, but there is data that contains more complex relationships than logistic regression can handle, or if this understanding is not required but we do not have the training time required for the gradient boosting machine, we would recommend the use of decision trees. Decision trees offer a balance between complexity and interpretability, making them suitable for medium-sized datasets with moderately complex relationships, where some interpretability is still required. In summary, each model has distinct strengths and weaknesses. The choice of the appropriate model depends on the specific question, complexity of the data, the need for interpretability, and the available computational resources.

6 Conclusion

In this chapter, we showed the potential of XAI approaches to explore and build theory on misfit. Using a data-driven approach to theory building, with techniques that can address the complexity and diversity of misfit, can help to move the misfit field forward. Our empirical illustrations focused on a set of values, misfit perceptions, and tenure as a contextual variable, and were meant as a starting point to illustrate the potential of XAI in this field. Taken together, there is a wealth of opportunities for explorative work in the area of misfit when using XAI.

References

Billsberry, J., Hollyoak, B. M., & Talbot, D. L. (2023). Insights into the lived experience of misfits at work: A netnographic study. *European Journal of Work and Organizational Psychology, 32*(2), 199–215.

Biran, O., & Cotton, C. (2017). Explanation and justification in machine learning: A survey. *IJCAI-17 Workshop on Explainable AI (XAI), 8*(1), 8–13.

Cable, D. M., & DeRue, D. S. (2002). The convergent and discriminant validity of subjective fit perceptions. *Journal of Applied Psychology, 87*(5), 875–884.

Chi, N.-W., Fang, L.-C., Shen, C.-T., & Fan, H.-L. (2020). Detrimental effects of newcomer person-job misfit on actual turnover and performance: The buffering role of multidimensional person-environment fit. *Applied Psychology, 69*(4), 1361–1395.

De Cooman, R., Mol, S. T., Billsberry, J., Boon, C., & Den Hartog, D. N. (2019). Epilogue: Frontiers in person–environment fit research. *European Journal of Work and Organizational Psychology, 28*(5), 646–652.

Doblhofer, D. S., Hauser, A., Kuonath, A., Haas, K., Agthe, M., & Frey, D. (2019). Make the best out of the bad: Coping with value incongruence through displaying facades of conformity, positive reframing, and self-disclosure. *European Journal of Work and Organizational Psychology, 28*(5), 572–593.

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives. *arXiv Preprint*. arXiv:1710.00794

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning. *arXiv Preprint*. arXiv:1702.08608

Edwards, J. R. (2008). Person–environment fit in organizations: An assessment of theoretical progress. *The Academy of Management Annals, 2*(1), 167–230.

Edwards, J. R., Cable, D. M., Williamson, I. O., Schurer Lambert, L., & Shipp, A. J. (2006). The phenomenology of fit: Linking the person and environment to the subjective experience of person–environment fit. *Journal of Applied Psychology, 91*(4), 802–827.

Englert, B., Sievert, M., Helmig, B., & Jansen, K. (2024). The incongruity of misfit: A systematic literature review and research agenda. *Human Relations, 77*(9), 1306–1332.

Follmer, E. H., Talbot, D. L., Kristof-Brown, A. L., Astrove, S. L., & Billsberry, J. (2018). Resolution, relief, and resignation: A qualitative study of responses to misfit at work. *Academy of Management Journal, 61*(2), 440–465.

Gilpin, L. H., Bau, D., Yuan, B. Z., Bajwa, A., Specter, M., & Kagal, L. (2018). Explaining explanations: An overview of interpretability of machine learning. *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA),* 80–89.

Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., & Hussain, A. (2024). Interpreting black-box models: A review on explainable artificial intelligence. *Cognitive Computation, 16*(1), 45–74.

Hastie, T., Tibshirani, R., Friedman, J. H., & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2, pp. 1–758). Springer.

Kristof-Brown, A., Schneider, B., & Su, R. (2023). Person-organization fit theory and research: Conundrums, conclusions, and calls to action. *Personnel Psychology, 76*(2), 375–412.

Li, B., Qi, P., Liu, B., Di, S., Liu, J., Pei, J., Yi, J., & Zhou, B. (2023). Trustworthy AI: From principles to practices. *ACM Computing Surveys, 55*(9), 1–46.

Linardatos, P., Papastefanopoulos, V., & Kotsiantis, S. (2021). Explainable AI: A review of machine learning interpretability methods. *Entropy, 23*(1), Article 1.

Minh, D., Wang, H. X., Li, Y. F., & Nguyen, T. N. (2022). Explainable artificial intelligence: a comprehensive review. *Artificial Intelligence Review,* 1–66.

Ostroff, C., Shin, Y., & Kinicki, A. J. (2005). Multiple perspectives of congruence: Relationships between value congruence and employee attitudes. *Journal of Organizational Behavior, 26*(6), 591–623.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Shrestha, Y. R., He, V. F., Puranam, P., & von Krogh, G. (2021). Algorithm supported induction for building theory: How can we use prediction models to theorize? *Organization Science, 32*(3), 856–880.

Sun, Y., & Billsberry, J. (2023). An investigation into how value incongruence became misfit. *Journal of Management History, 29*(3), 423–438.

Tolstoy, L. (1878). Anna Karenina.

Tonidandel, S., King, E. B., & Cortina, J. M. (2018). Big data methods: Leveraging modern data analytic techniques to build organizational science. *Organizational Research Methods, 21*(3), 525–547.

Van Vianen, A. E. M. (2018). Person–environment fit: A review of its basic tenets. *Annual Review of Organizational Psychology and Organizational Behavior, 5*(1), 75–101.

Vogel, R. M., Rodell, J. B., & Lynch, J. W. (2016). Engaged and productive misfits: How job crafting and leisure activity mitigate the negative effects of value incongruence. *Academy of Management Journal, 59*(5), 1561–1584.

Williamson, M. K., & Perumal, K. (2021). Exploring the consequences of person–environment misfit in the workplace: A qualitative study. *SA Journal of Industrial Psychology, 47*(1), 1–12.

Yang, L.-Q., Levine, E. L., Smith, M. A., Ispas, D., & Rossi, M. E. (2008). Person-environment fit or person plus environment: A meta-analysis of studies using polynomial regression analysis. *Human Resource Management Review, 18*(4), 311–321.

Corine Boon is Professor of HRM and People Analytics at the University of Amsterdam Business School and is the Director of the Amsterdam People Analytics Centre (APAC). Her work focuses on strategic HRM, people analytics, and person-environment fit. Her research has been published in journals such as the *Journal of Management*, *Human Resource Management*, and the *International Journal of Human Resource Management*, among others. She also serves as the editor-in-chief of *Human Resource Management Review*, is a member of several editorial boards, and she has been elected to the leadership track of the HR division of the Academy of Management.

Erdinç Durak is a Ph.D. candidate at the University of Amsterdam Business School and a member of the Amsterdam People Analytics Centre (APAC). His research explores the application of explainable artificial intelligence in understanding person-environment fit. His broader interests include machine learning, explainable AI, operational research, and their practical applications in people analytics and organizational science.

Ş. İlker Birbil is a Professor of AI & Optimization Techniques for Business and Society in University of Amsterdam, where he is the head of the Business Analytics section of the Amsterdam Business School. In the past, he had served for three years as a Professor of Data Science and Optimization at the Department of Econometrics of Erasmus University, and before that he had been a Professor of Optimization at the Industrial Engineering Department of Sabancı University for more than a decade. His research interests centre around optimization methods in data science and decision making. Lately, he is working on explainable artificial intelligence, optimization for machine learning, and data privacy in operations research.