Shaping Habit Formation Insights with Shapley Values: Towards an Explainable AI-system for Self-understanding and Health Behavior Change

ROBERT LEWIS, Massachusetts Institute of Technology, USA
YUANBO LIU, Massachusetts Institute of Technology, USA
MATTHEW GROH, Massachusetts Institute of Technology, USA
ROSALIND PICARD, Massachusetts Institute of Technology, USA

This paper presents our ongoing work to design an explainable artificial intelligence (XAI) system that helps individuals to form new healthy habits. We are developing this system on data collected from our recent observational study in which 62 participants attempted to develop a new mindful breathing habit over 6 weeks. We discuss the design and empirical results of our system, which uses Shapley values to generate explanations for predictions about user behavior, and outline how our technical approach can enable adaptive and personalized intervention tools that assist users in realizing health behavior change in the wild.

CCS Concepts: • Human-centered computing → Empirical studies in HCI; Visualization toolkits; • Theory of computation → Machine learning theory.

Additional Key Words and Phrases: Explainable AI, Interpretable Machine Learning, Habit Formation, Health Behavior Change, Mindfulness, Affective Computing

ACM Reference Format:

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
© 2021 Association for Computing Machinery.
Manuscript submitted to ACM
1 INTRODUCTION

Mindfulness is widely recognized as an effective technique for regulating mental and physical health [2, 4, 9, 22]. While even a single practice can yield benefits, the full advantages of mindfulness are unlocked with regular repetitions over an extended period of time. Habit formation is an effective mechanism through which to enable such behavior change and has been associated with positive long-term health outcomes [6]. Building a habit allows us to transition behavior from a deliberation that requires motivation into an impulse that is automatic [7]. By doing so, habit serves as a form of self-control [5], enabling consistent performance of health behaviors even with inevitable motivation lapses.

However, habit formation is not simple and efforts to build healthy habits often end up unsuccessful. While previous studies have shown that successful habit formation trajectories are asymptotic and associated with consistent repetition [11, 18], not everyone is guaranteed to make their way onto one of these upward trends in the wild. Indeed, so many factors in our life can impact our behavior regulation – from our mood, motivation, and daily activities, to exogenous factors like the weather – and this in turn influences how successful our habit formation endeavors will be.

In this paper, we propose that an explainable artificial intelligence system (XAI) could help users on this journey, by generating accurate explanations that can be packaged into personalized interventions. We have two key user personas in mind while designing this system. First, the habit builders themselves, for whom we believe the system could generate insights that promote self-understanding. Second, we consider habit formation supporters – such as behavior change system designers and care professionals – as an important category of users. Accurate and explainable predictions at various levels of aggregation – from individual to subgroup to population – should enable these supporters to understand the habit building dynamics of their clients on an individual basis, thus aiding their design of interventions that target the factors that prevent new habits from being formed and maintained. A conceptual overview of the components and capabilities of our system is displayed in Figure 1. This report focuses on the insights we can generate for habit formation supporters; co-designing the user experience with habit builders is left as important future work.

2 PROGRESS TOWARDS REAL-WORLD EXPLAINABLE AI SYSTEMS

AI methods have improved productivity in many industries. However, recent trends to further increase accuracy have come at the expense of interpretability and user experience, which has stunted the real-world adoption of AI technologies in many domains. If the machine cannot be explained, it is often not useful; even worse, it may hide pernicious biases or safety flaws that could have disastrous consequences such as physical or mental harm to end-users.

Given these risks, skepticism has rightly been raised against black-box state-of-the-art machine learning technologies [8], with many real-world systems adopting simpler AI models or foregoing AI entirely. While some simple models are inherently interpretable (such as linear regression or decision trees), their explainability may come at the expense of accuracy. An interpretable model is only so useful if it achieves substandard accuracy, especially on out-of-sample data. Moreover, as low out-of-sample accuracy suggests a model is a poor descriptor of a system’s general behavior, conclusions drawn from the interpretation of this model’s parameters may be tangential to the system’s true nature.

Recent advances provide promising solutions to this barrier to progress by offering model-agnostic interpretability tools for black-box models [8, 16, 24]. Techniques such as Shapley values [12, 13, 23, 24], LIME [19], and Anchors [20] decouple the explanation generator from the underlying prediction model, an abstraction that allows the style and complexity of the prediction model to change without detriment to its interpretability. A benefit of these tools is if a developer discovers that a black-box model (such as a random forest or neural network) is more accurate than their
current model, then they can substitute it into their system without changing the explanation experience [13]. In this report we show how Shapley values generate system insights, saving comparison to other XAI methods for future work.

3 DESIGNING AN EXPLAINABLE AI SYSTEM TO PREDICT AND EXPLAIN TOMORROW’S BEHAVIOR

3.1 The Forming Healthy Habits Study

We are developing our explainable AI system on data collected from a six-week observational study we recently conducted, concluding in January 2021, that involved 62 participants who planned to adopt a new daily mindful breathing habit. Participants completed daily surveys, including whether they did the breathing exercise; how automatic, rewarding and challenging it felt; their confidence and motivation for building the habit; and questions about their mood and daily activities. Participants also completed pre-study and post-study surveys, providing information on their past mindfulness experience, their commitment to forming a new mindful breathing habit, their well-being [25], and their personality [14]. Overall, 47.4% (N=1,234) of daily surveys were completed and 41 participants completed the post-study survey. More details on the study protocol and data collected can be found in Appendix A.

3.2 Explainable AI System Design

We have created an XAI model that learns how to predict whether the user will practice the breathing exercise at the next opportunity (the next day). The prediction model is a binary classifier defined by: \( \hat{y}_{t, i} = f(X_{t, i}) \), where \( \hat{y}_{t, i} \) is the prediction of whether user \( i \) will practice the exercise tomorrow, using the information \( X_{t, i} \) collected from them today. Given the dataset is imbalanced – with more examples of an individual practicing the breathing exercise and completing the survey than not practicing but still reporting – we make missing the habit action the positive class in the model. Thus, the model output represents the probability a participant will miss tomorrow’s practice of the mindful breathing.

To add explanatory capabilities, we compute Shapley values [12, 13, 23, 24] for the predictions of the model. Shapley values are calculated for each feature on an individual data instance basis, and they represent the change in the value of the model prediction for the data instance, relative to the average prediction (or expected value), when the feature in question is added to the model. So, for example, in the binary classification case, if a feature receives a positive Shapley value, that indicates it is associated with increasing the probability that a participant will miss their next practice of the breathing exercise by an amount equivalent to the magnitude of the Shapley value (and vice versa for a feature with a negative Shapley value). Moreover, for a given prediction, adding all the Shapley values for all the features creates a sum that represents the difference between the predicted value and the expected value: thus, the contribution of all features are included in the explanation\(^1\). Shapley value theory is discussed further in Appendix B.

Shapley values offer several advantages to our system. First, as they are model-agnostic, they do not tie us into a specific modelling paradigm. Second, Shapley values can be aggregated to arbitrary levels, offering significant flexibility in how explanations are presented. For example, one can display Shapley values for a single prediction, all predictions, or any subgroups. This property enables interesting comparisons to be made, both intra-individual and inter-individual, and these can be used to formulate contrastive explanations, a style of explanation that is known to resonate with human users [15]. For example, one can ask the questions “How is the explanation for this prediction different from my past behavior?” or “How do my most predictive factors compare to those of other people?”.

\(^1\)This interpretation and additive property of Shapley values is analogous to that of feature contributions in linear models, such as linear regression, where the concept is referred to as the situational importance of the features. Indeed, because this property is so desirable for explaining predictions, replicating it for any prediction model was a design goal for the teams that proposed using Shapley values in a machine learning context [12, 15, 24].
3.3 System Predictive Accuracy

![Figure 2: An overview of the system prediction and explanation capabilities.](image)

(a) Model accuracy (PRAUC) by model type.  
(b) Feature importance for all participants in cohort.  
(c) Feature importance by experience subgroup in cohort.  
(d) Feature importance by well-being subgroup in cohort.

Figure 2a displays the accuracy\(^2\) of several machine learning algorithms on the prediction task. Models were trained and evaluated at both the group level (one model for all participants) and the individual level (one model per participant), using features collected from the participant daily surveys (Appendix A). To fit a model for each individual requires that they have a certain amount of historical data. As such, we exclude some users from the analysis\(^3\) and ensure these exclusions are consistent between the group and individual models. Our evaluation in this setting is thus equivalent to the \textit{warm-start} scenario encountered in practice, where users have previous interactions with the system. The results are reported on the \textit{hold-out folds} in a nested cross-validation evaluation scheme\(^4\), hence they are indicative of each model’s ability to generalize to previously unseen data (which is the expected setting when a model is used \textit{in the wild}).

From Figure 2a we see several emerging trends. First, it is clear that training a personalized model for each individual consistently leads to greater overall accuracy. This finding suggests there is heterogeneity in the factors across individuals that correlate with their ability to maintain their habit formation routines. Second, more complex models – notably the random forest – achieve greater accuracy compared to simpler baselines. Random forests place less inductive bias on the functional form of the mapping from features to predictions, by permitting nonlinear relations, for example. Therefore, their increased predictive power might suggest that future habit behavior is better explained by complex combinations of factors, rather than linear relations or simple heuristics (such as similarity to past experiences).

---

\(^2\)Given the imbalance in the dataset, area under the precision-recall curve (PRAUC) is used as the accuracy metric [21].

\(^3\)First, participants with less than 10 observations were excluded so that there were at least 2 observations per fold in the 5-fold cross validation scheme. Second, users were also excluded if they only had observations corresponding to a single outcome (i.e., if they always did the exercise or never did it), as several of the models are undefined if both classes are not available. The exclusions result in a set of 26 users, referred to as the \textit{warm-start} cohort.

\(^4\)5 outer folds are used and the average accuracy across the hold-out folds is reported. For the group folds, hyperparameters are optimized over 5 inner folds. However, hyperparameter tuning is not performed on the individual level models, given the small amount of data available per participant. At both levels, the folds are randomized across time, so that any bias arising from seasonal effects – such as the beginning of a holiday period – is mitigated.
3.4 Explaining Predictions To Generate Behavioral Insights
Figure 2b presents which features are most important to the best-performing prediction model. Specifically, it displays the mean absolute Shapley values aggregated over all participants in the warm-start cohort for the individual-level random forest in Figure 2a, with larger values indicating a factor is more important in determining the prediction that the habit will not be practiced tomorrow. We observe that how much a participant was away from home, how challenging the habit felt, and how busy they were, are the most influential factors in the model’s predictions.

Figures 2c-2d illustrate how the analysis of feature importance can be taken to a more granular level, with importance scores aggregated over participant subgroups determined by a participant’s level of pre-study mindfulness experience. We see interesting variation in important factors at this level. For example, from Figure 2c, it appears mood and context-related variables (e.g., the user’s stress and how good the weather was), as well as mindful breathing reward and habit building motivation and confidence, are comparatively more important for participants with lower levels of mindfulness experience. These insights, when presented to a care professional or system designer, may enable them to design customized interventions for less experienced clients that target the factors – such as behaviors and contexts – that are most likely to influence the probability they practice the exercise and strengthen the habit.

While there is not space in the main body, in Appendix C we briefly show how Shapley values can create further insight when we consider their signs in addition to their magnitudes. Insights that use the sign can further help a professional to fine-tune their support, indicating not just the factors that their interventions should target, but also the direction of change in which to guide their clients to increase the predicted probability they practice the exercise.

3.5 Subgrouping Users by their Observed Behavior
Finally, we propose that when Shapley values are aggregated to the individual user-level, they can be used to create novel user subgroups based on their observed behavior. When compared to subgrouping by a priori user characteristics (e.g., Figures 2c-2d), this approach may provide a useful lens to support professionals as it allows them to organize their clients into groups that reflect similarities in their experiences of developing the target behavior. Moreover, Shapley value subgroupings can be regenerated dynamically as more user data is collected, which may facilitate varying the content of interventions as users make progress on developing their habits. Figure 3a shows the heterogeneity in Shapley values (and thus observed behavior) between users, with Figure 3b proposing a way to organise this information that highlights users with similar behavioral characteristics.

4 CONCLUSIONS AND FUTURE WORK
In this paper we have outlined the design and results of an explainable AI system to help users build mindfulness habits. We show how Shapley values can cast light on the internal workings of black-box models such as the random forest, allowing system designers to learn how various factors influence the predicted probability of habit formation behaviors in real-world contexts. Furthermore, we show that system designers can take their user segmentation analyses further by creating novel user subgroups that use Shapley value vectors as representations of their observed behavior.

This work has limitations. First, we have not had the space here to compare Shapley value results to those produced by other XAI techniques. Second, model-agnostic interpretability techniques alone are not sufficient to solve the challenges related to the adoption of complex AI systems in the wild. While they are a critical system component, additional work is required to understand the user experience of receiving the explanations they produce [3, 15, 28]. It is one thing to...
generate an explanation that is faithful to the mechanics of the system; it is a significant further challenge to package this explanation into a personalized, engaging and actionable insight that resonates with the end-user. Recent studies have explored methodologies for the user-centric design of AI-systems in real-world settings, with several taking into account the system’s capacity to explain its decisions [1, 13, 28]. Furthermore, contemporary work emphasizes the importance of considering different narratives and aesthetics when presenting data-driven health insights, calling for designers to take into account the emotional and cognitive impacts of data monitoring, such as data-induced guilt, where a user feels shame about their actions relative to the trends displayed, and information overload, where users are presented with more information than they feel comfortable interpreting and making decisions from [10, 17].

Now that our observational study has concluded, an important next step is to engage participants in co-design of the explanation interface. We will also investigate advanced prediction methods to further improve accuracy, and analyze these methods in additional user scenarios including user cold-start. Finally, given many factors often explain a prediction, and different users will prefer different styles of explanation, we plan to assess recommender systems as a method for personalizing model explanations [29]. A recommender system learns to curate content based on the real-time feedback of users: we hope this mechanism will enable our system to customize its explanations in a way that engages, encourages and empowers its users.

Fig. 3. Creating novel user subgroupings based on observed behavior with Shapley values. (a): Median Shapley values for each feature by participant. (b): Organizing users by similarity in median Shapley value vectors using the Euclidean distance. Nodes represent users and thicker edge lines represent greater similarity between users.

REFERENCES


A THE FORMING HEALTHY HABITS STUDY

The initial phase of the Forming Healthy Habits Study consisted of an observational study, concluding in January 2021, that involved 62 participants who planned to adopt a new mindful breathing habit. Table 1 summarizes the data collected. Daily survey items (A1-A5) are used as features in the XAI model (described in Section 3.2). These features are on a 7-point likert scale with the exception of practicing the breathing exercise which is Boolean. No preprocessing is performed on the feature values. Pre-survey items (B1-B6) are not used as independent variables in the XAI system,
however they are used to aggregate the results to provide insights about participant subgroups. At the time of writing, mid- and post-survey items (C1-C2) are not used in the system, but will be incorporated as part of future work.

Table 1. Data collected from our six-week observational study in which 62 participants attempted to develop a new daily mindful breathing habit.

<table>
<thead>
<tr>
<th>A. Daily survey items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Completion</td>
</tr>
<tr>
<td>2. SRHI habit automaticity</td>
</tr>
<tr>
<td>3. Other habit reflections</td>
</tr>
<tr>
<td>4. Mood</td>
</tr>
<tr>
<td>5. Daily context</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>B. Pre-survey items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Demographics</td>
</tr>
<tr>
<td>2. Past experience</td>
</tr>
<tr>
<td>3. Commitment</td>
</tr>
<tr>
<td>4. Habit strength</td>
</tr>
<tr>
<td>5. Well-being</td>
</tr>
<tr>
<td>6. Personality</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>C. Mid- and post-survey items</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Well-being</td>
</tr>
<tr>
<td>2. Habit formation reflections</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>D. Passive smartphone usage data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Smartphone usage</td>
</tr>
</tbody>
</table>

**B SHAPLEY VALUES**

In our system we use Shapley values to represent the contribution of each feature to the prediction for each individual data instance. Shapley values (represented as \( \phi_j \) for feature \( j \)) have the following properties which make them powerful tools for explaining AI predictions:

1. **Efficiency**: which implies that the sum of Shapley values over all features must equal the difference between the predicted value, \( \hat{f}(x) \), for the given data instance, \( x \), and the expected value of the prediction model, \( E_X(\hat{f}(X)) \):

\[
\sum_{j=1}^{p} \phi_j = \hat{f}(x) - E_X(\hat{f}(X))
\]  

2. **Symmetry**: which implies that if two feature values influence the prediction to the same extent then they should receive the same Shapley value.

3. **Dummy**: which implies that if a feature has no influence on the predicted value (in other words, it is redundant) then it should receive a Shapley value of zero.
(4) **Additivity**: which implies that when calculating the Shapley values of submodels (or subgroups) one should be able to aggregate them (e.g., by averaging) to values that are consistent with the Shapley values for the overall model (or population).

Estimating Shapley values requires solving multiple integrations over coalitions of features and the details of this procedure are not required to understand the intuition behind Shapley values. As such we do not document these procedures in the interest of brevity, but refer the reader to the original papers that proposed using Shapley values in a machine learning context [12, 24], as well as a very thorough review of their theory by Christoph Molnar [16]. We use the Shapley Additive Explanations (SHAP) estimation method in our system [12], making use of the Python library\(^6\) created by the method’s authors.

**C FURTHER EXPLAINABLE AI RESULTS**

Figure 4 further illustrates the insights we can generate with Shapley values when we consider their signs in addition to their magnitudes. For well-being subgroups, it shows the overall distribution in Shapley values for the 5 most important features (a), and how these values change as the underlying feature value changes (b-f). We see, for example, that when the breathing exercise feels more challenging, this correlates with a larger increase in the probability that a user in a low well-being state will miss the exercise at the next opportunity, relative to a medium well-being user (represented by higher average Shapley values at higher levels of challenge). Similarly, we see that lower levels of busyness for medium well-being users correlate with a higher probability of missing the next exercise, relative to low well-being users. Insights at this level – that use the Shapley value sign as well as the magnitude – can further help a professional to fine-tune their support, indicating not just the factors that their interventions should target, but also the direction of change in which to guide their clients to increase the predicted probability they practice the exercise.

![Figure 4](https://example.com/figure4.png)

Fig. 4. More granular explanatory insights using Shapley value signs by well-being subgroups. (a): Overall distribution in Shapley values for the 5 most important features. (b-f): Average Shapley value by underlying feature value (where feature values are on a 7-point Likert scale). Higher values for the sleep and stress scales represent higher quality sleep and higher stress levels, respectively.

---

\(^6\)https://github.com/slundberg/shap