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

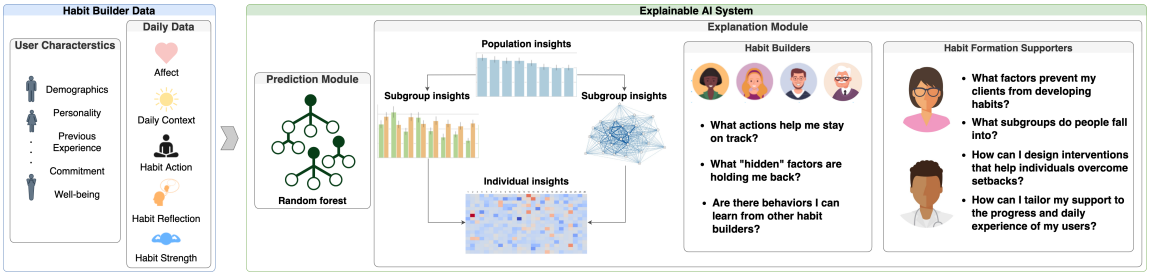


Fig. 1. Conceptual design of our explainable AI system to assist habit builders and their supporters.

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:
Robert Lewis, Yuanbo Liu, Matthew Groh, and Rosalind Picard. 2021. Shaping Habit Formation Insights with Shapley Values: Towards an Explainable AI-system for Self-understanding and Health Behavior Change. In *Proceedings of CHI '21: Realizing AI in Healthcare: Challenges Appearing in the Wild Workshop (CHI '21 WS AI Health)*. ACM, New York, NY, USA, 9 pages.

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

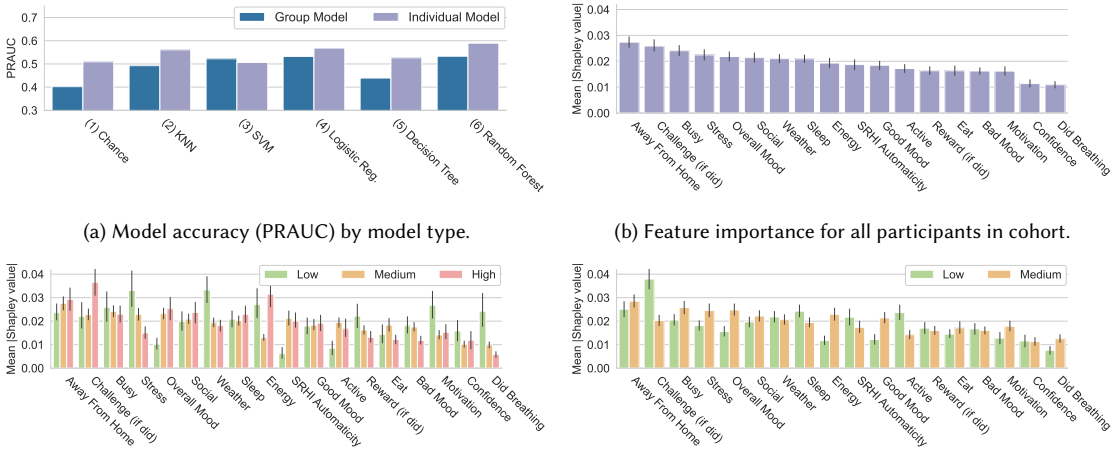
We have created an XAI model that learns how to predict whether the user will practice the breathing exercise at the next opportunity (the next day). The prediction model is a binary classifier defined by: $\hat{y}_{t_1,i} = f(X_{t_0,i})$, where $\hat{y}_{t_1,i}$ is the prediction of whether user i will practice the exercise tomorrow, using the information $X_{t_0,i}$ 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 intraindividual and interindividual, 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, 13, 24].

3.3 System Predictive Accuracy



(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.

Fig. 2. An overview of the system prediction and explanation capabilities. (a): Model accuracy (area under the precision-recall curve, PRAUC) by model type, where *Chance* is a baseline representing random prediction given the training set's class distribution. (b-d): Feature importance (mean absolute Shapley value) from the individual random forest models for all participants (b) and different participant subgroupings (c-d). All results are reported on data from the 26 participants in the *warm-start* cohort. The features are on a 7-point *likert* scale except for "Did Breathing" which is Boolean. Appendix A describes the features in further detail.

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 *warm-start* scenario encountered in practice, where users have previous interactions with the system. The results are reported on the *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 *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 *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[5].

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 subgroupings 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

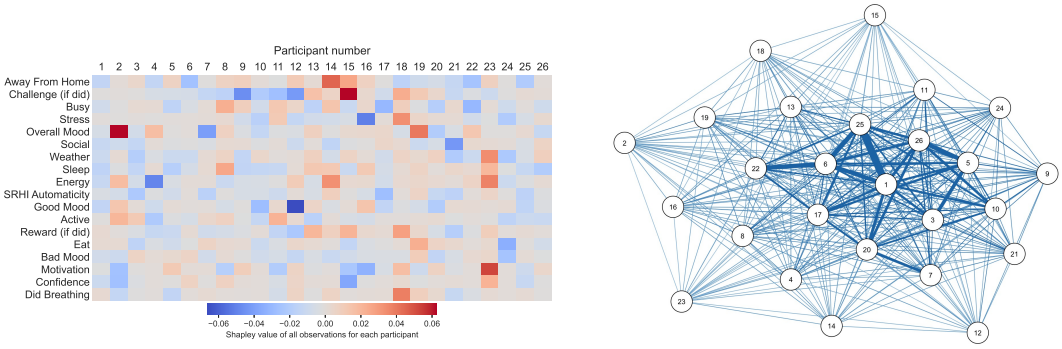
In this paper we have outlined the design and results of an explainable AI system to help users build mindfulness habits. We show how Shapley values can cast light on the internal workings of *black-box* models such as the random forest, allowing system designers to learn how various factors influence 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

[5]NB: These observations are specific to the sample of users in the observational study, and performing statistical tests to make population level conclusions / inferences about habit formation is beyond the scope of this position paper.

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.



(a) Median Shapley values by participant.

(b) Users organised by Shapley value similarity.

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

[1] Emma Beede, Elizabeth Baylor, Fred Hersch, Anna Iurchenko, Lauren Wilcox, Paisan Ruamviboonsuk, and Laura M. Vardoulakis. 2020. A Human-Centered Evaluation of a Deep Learning System Deployed in Clinics for the Detection of Diabetic Retinopathy. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–12. https://doi.org/10.1145/3313831.3376718

[2] Hyunju Cho, Seokjin Ryu, Jeeae Noh, and Jongsun Lee. 2016. The effectiveness of daily mindful breathing practices on test anxiety of students. *PLoS ONE* 11, 10 (2016), 1–10. https://doi.org/10.1371/journal.pone.0164822

[3] Angèle Christin. 2017. Algorithms in practice: Comparing web journalism and criminal justice. *Big Data and Society* 4, 2 (2017), 1–14. https://doi.org/10.1177/2053951717718855

[4] J. David Creswell. 2017. Mindfulness Interventions. *Annual Review of Psychology* 68 (2017), 491–516. https://doi.org/10.1146/annurev-psych-042716-051139

[5] Brian Galla and Angela Duckworth. 2015. More Than Resisting Temptation: Beneficial Habits Mediate the Relationship Between Self-Control and Positive Life Outcomes. *Journal of personality and social psychology* 109 (02 2015). https://doi.org/10.1037/pspp0000026

[6] Benjamin Gardner. 2015. A review and analysis of the use of 'habit' in understanding, predicting and influencing health-related behaviour. *Health Psychology Review* 9, 3 (2015), 277–295. https://doi.org/10.1080/17437199.2013.876238 PMID: 25207647.

[7] Benjamin Gardner and Amanda L. Rebar. 2019. Habit Formation and Behavior Change. https://doi.org/10.1093/acrefore/9780190236557.013.129
[8] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Dino Pedreschi, and Fosca Giannotti. 2018. A survey of methods for explaining black box models. *arXiv* 51, 5 (2018). arXiv:1802.01933
[9] Stefan G. Hofmann, Alice T. Sawyer, Ashley A. Witt, and Diana Oh. 2010. The Effect of Mindfulness-Based Therapy on Anxiety and Depression: A Meta-Analytic Review. *Journal of Consulting and Clinical Psychology* 78, 2 (2010), 169–183. https://doi.org/10.1037/a0018555
[10] Elizabeth Kaziunas, Mark S. Ackerman, Silvia Lindtner, and Joyce M. Lee. 2017. Caring through Data. (2017), 2260–2272. https://doi.org/10.1145/2998181.2998303
[11] Phillippa Lally and Benjamin Gardner. 2013. Promoting habit formation. *Health Psychology Review* 7, sup1 (2013), S137–S158. https://doi.org/10.1080/17437199.2011.603640 arXiv:https://doi.org/10.1080/17437199.2011.603640
[12] Scott M. Lundberg, Gabriel Erion, Hugh Chen, Alex DeGrave, Jordan M. Prutkin, Bala Nair, Ronit Katz, Jonathan Himmelfarb, Nisha Bansal, and Su In Lee. 2019. Explainable AI for trees: From local explanations to global understanding. *arXiv* 2, January (2019). https://doi.org/10.1038/s42256-019-0138-9 arXiv:1905.04610
[13] Scott M. Lundberg, Bala Nair, Monica S. Vavilala, Mayumi Horibe, Michael J. Eisses, Trevor Adams, David E. Liston, Daniel King Wai Low, Shu Fang Newman, Jerry Kim, and Su In Lee. 2018. Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nature Biomedical Engineering* 2, 10 (2018), 749–760. https://doi.org/10.1038/s41551-018-0304-0
[14] Robert R McCrae and Paul T Costa Jr. 2008. The five-factor theory of personality. In *Handbook of personality: Theory and research, 3rd ed.* The Guilford Press, New York, NY, US, 159–181.
[15] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007 arXiv:1706.07269
[16] Christoph Molnar. 2019. *Interpretable Machine Learning*. https://christophm.github.io/interpretable-ml-book/.
[17] Elizabeth L. Murnane, Xin Jiang, Anna Kong, Michelle Park, Weili Shi, Connor Soohoo, Luke Vink, Iris Xia, Xin Yu, John Yang-Sammataro, Grace Young, Jenny Zhi, Paula Moya, and James A. Landay. 2020. Designing Ambient Narrative-Based Interfaces to Reflect and Motivate Physical Activity. *Conference on Human Factors in Computing Systems - Proceedings* (2020), 1–14. https://doi.org/10.1145/3313831.3376478
[18] David T. Neal, Wendy Wood, and Jeffrey M. Quinn. 2006. Habits - A repeat performance. *Current Directions in Psychological Science* 15, 4 (2006), 198–202. https://doi.org/10.1111/j.1467-8721.2006.00435.x
[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016.* 1135–1144.
[20] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *AAAI*.
[21] Takaya Saito and Marc Rehmsmeier. 2015. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS ONE* 10, 3 (2015), 1–21. https://doi.org/10.1371/journal.pone.0118432
[22] Benjamin Schöne, Thomas Gruber, Sebastian Graetz, Martin Bernhof, and Peter Malinowski. 2018. Mindful breath awareness meditation facilitates efficiency gains in brain networks: A steady-state visually evoked potentials study. *Scientific Reports* 8, 1 (2018), 13687. https://doi.org/10.1038/s41598-018-32046-5
[23] L Shapley. 1953. A Value for n-Person Games. (1953), 307–318. https://doi.org/10.1515/9781400881970-018
[24] Erik Štrumbelj and Igor Kononenko. 2014. Explaining prediction models and individual predictions with feature contributions. *Knowledge and Information Systems* 41, 3 (2014), 647–665. https://doi.org/10.1007/s10115-013-0679-x
[25] Ruth Tennant, Louise Hiller, Ruth Fishwick, Stephen Platt, Stephen Joseph, Scott Weich, Jane Parkinson, Jenny Secker, and Sarah Stewart-Brown. 2007. The Warwick-Edinburgh Mental Well-being Scale (WEMWBS): Development and UK validation. https://doi.org/10.1186/1477-7525-5-63
[26] John Torous, Mathew V Kiang, Jeanette Lorme, and Jukka-Pekka Onnela. 2016. New Tools for New Research in Psychiatry: A Scalable and Customizable Platform to Empower Data Driven Smartphone Research. *JMIR Mental Health* 3, 2 (2016), e16. https://doi.org/10.2196/mental.5165
[27] Bas Verplanken and Sheina Orbell. 2003. Reflections on Past Behavior: A Self-Report Index of Habit Strength. *Journal of Applied Social Psychology* 33 (05 2003), 1313 – 1330. https://doi.org/10.1111/j.1559-1816.2003.tb01951.x
[28] Christine T. Wolf. 2019. Explainability scenarios: Towards scenario-based XAI design. *International Conference on Intelligent User Interfaces, Proceedings IUI* Part F1476 (2019), 252–257. https://doi.org/10.1145/3301275.3302317
[29] Yongfeng Zhang, Xu Chen, and Boston Delft. 2020. Foundations and Trends ® in Information Retrieval Explainable Recommendation: A Survey and New Perspectives. *Foundation and Trends in Information Retrieval* 14, 1 (2020), 1–101. https://doi.org/10.1561/1500000066.Yongfeng

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.

A. Daily survey items	
1. Completion	Whether or not the participant did the mindful breathing exercise.
2. SRHI habit automaticity	3 questions from the SRHI [27] scale related to habit automaticity. On a 7-point scale from "Extremely Inaccurate" to "Extremely Accurate", participants rate the extent mindful breathing is something that: i) I do automatically, ii) I do without having to consciously remember, iii) I would find hard not to do. Note: on every seventh day, participants complete the full 12-item SRHI.
3. Other habit reflections	Participants rate (7-point scale) their i) motivation and ii) confidence for the building the habit. Additionally, if they did the breathing exercise they rate how iii) rewarding and iv) how challenging it felt.
4. Mood	For the past 24 hours, participants rate (7-point scale) how often they felt in i) a good mood and ii) a bad mood; the extent they felt iii) calm or stressed and iv) lethargic or energetic; and v) their overall rating of mood from extremely unpleasant to extremely pleasant.
5. Daily context	Participants rate (7-point scale) i) how busy their day was, ii) how well they slept, iii) how physically active they have been, iv) how well they ate, v) how much they interacted with other people, vi) how much they enjoyed the weather, and vii) how much time they spent away from their home residence.
B. Pre-survey items	
1. Demographics	Various items of information on how participants identify (for example age, gender and ethnicity).
2. Past experience	Participants rate (7-point scale) how experienced they are at mindfulness.
3. Commitment	Participants rate (7-point scale) how committed they are to forming the daily mindful breathing habit during the study.
4. Habit strength	Participants complete the 12-item SRHI [27] to survey the strength of their mindful breathing habit at study initiation.
5. Well-being	Participants complete the The Warwick-Edinburgh Mental Well-being Scale survey (WEMWBS [25]).
6. Personality	Participants complete the Five Factor Personality Model survey [14].
C. Mid- and post-survey items	
1. Well-being	Participants complete the the WEMWBS survey again [25].
2. Habit formation reflections	Participants are prompted to rate (7-point scale) how i) rewarding, ii) challenging, and iii) frustrating their habit formation experience has been. There is also space for participants to provide open-ended reflections on their experiences.
D. Passive smartphone usage data	
1. Smartphone usage	The Beiwe *digital phenotyping* platform [26] was also used to passively collect data on participant daily smartphone usage for the duration of the study period. Data includes location, accelerometer, and screen lock/unlock time.

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 ϕ_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, $E_X(\hat{f}(x))$

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

(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.

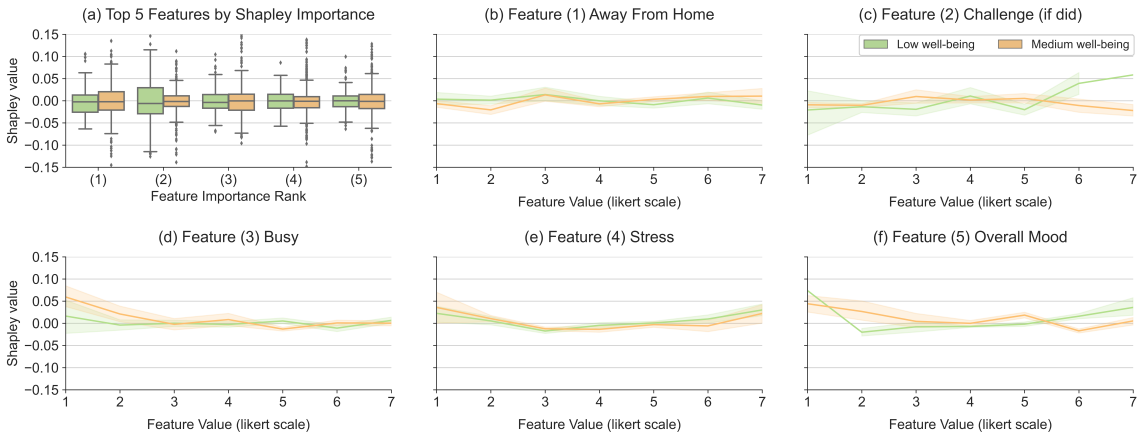


Fig. 4. More granular explanatory insights using Shapley value signs by well-being subgroups. (a): Overall distribution in Shapley values for the 5 most important features. (b-f): Average Shapley value by underlying feature value (where feature values are on a 7-point *likert* scale). Higher values for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

[6]https://github.com/slundberg/shap