eXplainable AI
in Personalized Mental Healthcare

NICEDAY DEEPLOY COUNCYL

Table of contents

Introduction

As AI and ML systems gain importance in decision-making, the rising issue becomes transparency. The increasing complexity of AI algorithms has given rise to what is often referred to as "black box" models, where the inner workings remain inscrutable even to experts. This obscurity hampers trust and raises ethical questions about accountability in decisions made by AI systems.

In the medical field, decision support is vital. Current eXplainable AI (XAI) solutions focus on the integration of recommendation engines in the workflow of professionals, explaining why a certain recommendation has been made, before confirming which treatment process is suitable. However, clinicians lack a feedback loop to update the decision model based on whether recommendations were right or wrong.

This initiative employed XAI and Incremental Learning to provide decision support to mental health therapists. Specifically, this system was designed to help therapists prioritise clients who may benefit from outreach between therapy sessions – a daily decision that is made under time pressure and concerns large patient caseloads. The system design used a human-centred approach, incorporating end users into a) the selection of the specific decision to be supported, b) the development of the underlying recommendation engine, and c) the design of the interface to preserve explainability where users interact with the system (receiving recommendations and providing feedback for incremental learning).

Background

NiceDay is a leading (online) mental healthcare platform in the Netherlands, aiming to strengthen the mental wellbeing of as many people as possible.

NiceDay has introduced **the Data-supported Treatment** method, an innovative approach that uses technology to seamlessly combine evidence-based practices in psychopathology—such as CBT, Experience Sampling Methodology (ESM), and Feedback Informed Therapy—with sophisticated, data-driven decision support tools. These tools were created by leveraging insights from clients' historical data and combining this with the expertise of therapists. This method aims to enhance mental health outcomes and equip therapists with refined analytical insights. At the heart of this approach is the facilitation of treatment actions, ongoing client monitoring, and timely interventions.

Problem

Between-session care is an integral part of a data-supported treatment.

The inception of this decision support tool was motivated by challenges NiceDay therapists face in making optimal use of the between session care moments, and deciding which patients need extra support or a between session intervention.

01 Large client caseloads

Each NiceDay therapist is responsible for a client caseload of 30-40 clients. Clients have various complaints, different treatment goals, and are at varying stages of treatment, all of which are relevant to providing effective between-session care.

02 Many disparate data points

NiceDay provides unique data-driven therapy, which includes features that allow therapists to monitor and analyse patient data such as diary and feelings registrations, symptoms trackers, exposure exercises, psychotherapy documentation assignments, feedback informed therapy questionnaires, and more. However, treatment is personalized to each patient meaning patients in a therapists' caseload will use different trackers and exercises resulting in various data streams. This means that therapists find it extremely challenging not only to retain all of this data, but to effectively make use of it all.

03 Limited time

Therapists may have as little as 30 minutes per day to provide care to patients between their scheduled therapy sessions. Time spent reviewing patient registrations and the various data streams – tracking data, exercise completion, diary entries – to identify which patients require additional support means there is less time available for actually providing the between session care.

There is **no standard approach** in deciding who to intervene with during between session care moments; each individual therapist uses their own intuition to make optimal use of this time. There is also **no way to indicate** whether outreach was indeed necessary. As a result, there is **no historical data** available that can provide insight into whether when and if between session care was provided, and if it was 'correct' to do so. This gap in the process not only **places strain** on therapists and potentially **delays necessary interventions** for clients in need, but also **contributes to an absence of data** that could be leveraged to enable better-informed outreach decisions.

Solution

A prioritised client dashboard for therapists to review during designated between sessions care timeslots

In light of the pressing need to augment the effectiveness and efficiency of in between-session care in a mental healthcare setting, we devised an AI-based decision support application to be part of NiceDay.

The core of this application lies in facilitating the identification of clients who may be in immediate need of assistance, based on the output of the AI system, using recent client data. Therapists are presented with a ranked list of clients, accompanied by a set of reasons that underscore why a particular client might require immediate attention. This tool, thus, is expected to function as a vital piece of the data-supported treatment provided to clients via NiceDay.

Our decision support application integrates AI capabilities with the expertise of therapists, enhancing both the effectiveness and efficiency of the data-supported treatment. The AI assimilates recent historical data registered in the NiceDay platform with therapist input on possible courses of action, paving the path towards a more responsive and personalised healthcare approach. By incorporating incremental learning, the model constantly evolves based on therapist actions and feedback – a critical step in refining the model over time to yield progressively improved predictions and foster a culture of continual improvement and adaptation.

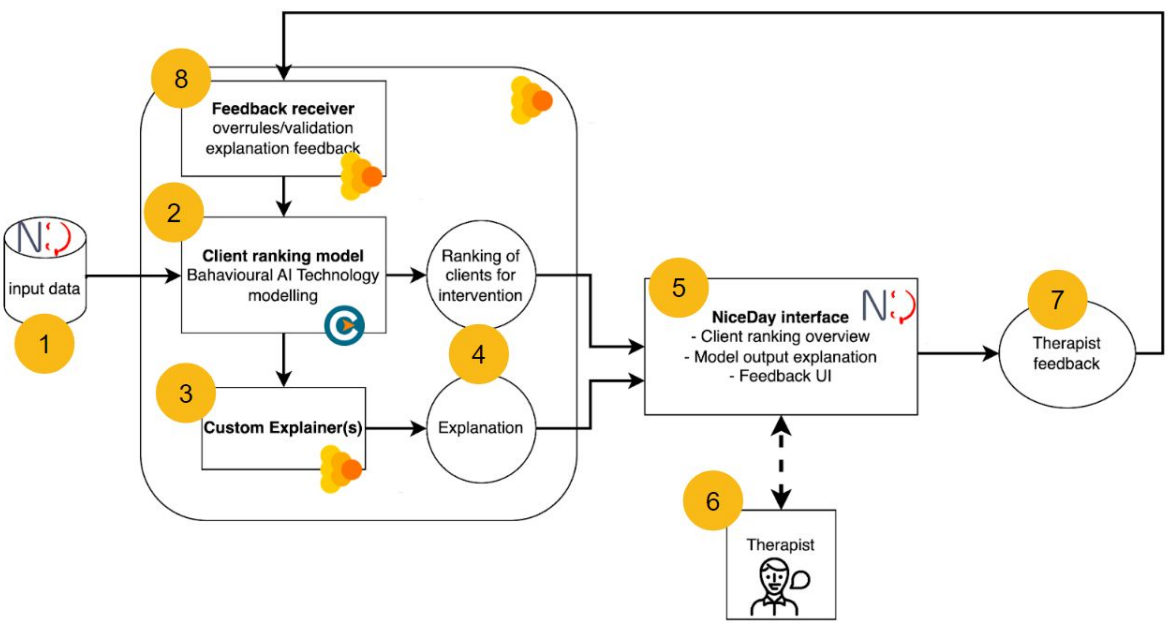
The rationale behind developing this application extends beyond alleviating the workload of therapists. It is designed to foster a collaborative environment where AI serves as an auxiliary support, enabling therapists to make informed decisions swiftly. It streamlines the process of monitoring and follow-up tasks, ensuring that therapists can proactively reach out to clients who may benefit significantly from timely intervention, without being overwhelmed by the extensive manual labour that was previously required.

Moreover, it fundamentally transforms the dynamic of between-session care, making it more adaptive, responsive, and data-driven. The utilisation of explainable AI in this scenario not only aids in identifying clients in need but also educates the therapists on the underlying reasons behind the recommendations, thereby nurturing a feedback loop for continuous improvement and learning. Lastly, by keeping therapists in the lead of decision-making and embedding a feedback mechanism for the recommendation engine, we are fostering a mutually beneficial relationship between technology and healthcare professionals. This relationship preserves the human touch in mental healthcare, while elevating it with the power of AI to process vast amounts of data quickly and accurately.

System Design

A 3-way collaboration to create an eXplainable human-centred decision support system with incremental learning.

NiceDay partnered with Deeploy (makers of an MLOps platform for serving, monitoring, and explaining AI models), and Councyl (makers of a decision support software that develops and serves client-specific expert choice models) to build the following complex of systems.



1. Input data used for the model flows from NiceDay's database
2. The BAIT model uses the input data to provide a ranking of the clients who might need intervention
3. Using custom explainers, the model output is made explainable for therapists
4. The model output and explanations are sent to NiceDay
5. The overview of clients and explanations are presented in NiceDay
6. The therapist can view the model output and explanations and give feedback in Niceday
7. The therapist feedback flows to Deeploy
8. The feedback receiver saves feedback collected, which can be used to improve the BAIT model

Key Components

The following elements provide the foundation for the system's operation, working in concert to deliver a useful decision support system where all observations are explainable and where incremental learning is facilitated with the same level of scrutiny and transparency as the initial system build.

01 BAIT model

BAIT is an approach to choice model development that coalesces domain-level input and observational-level choice data from subject matter experts. Because NiceDay lacks the historical data combining patient "case" (or data scenario at time of assessment) with therapist choice label (did the therapist provide between-session care or not, and was it correct to do so), a different approach that did not rely on this historical data was needed to develop the prioritisation model. Because NiceDay **does** have the domain expertise to make informed decisions regarding which clients require outreach in between session care, BAIT is a suitable solution.

02 Custom explainers

To enhance adoption and trust in the AI system, therapists must be able to understand on what basis recommendations are made. When therapists understand why a certain client is pointed out for outreach, he/she can provide meaningful feedback on the model and explanations. To facilitate this, custom explainers have been developed. Custom explainers – in this case, are natural language expressions of the values of the top three most important input criteria as determined by a calculation of feature importance – were developed to provide this understanding.

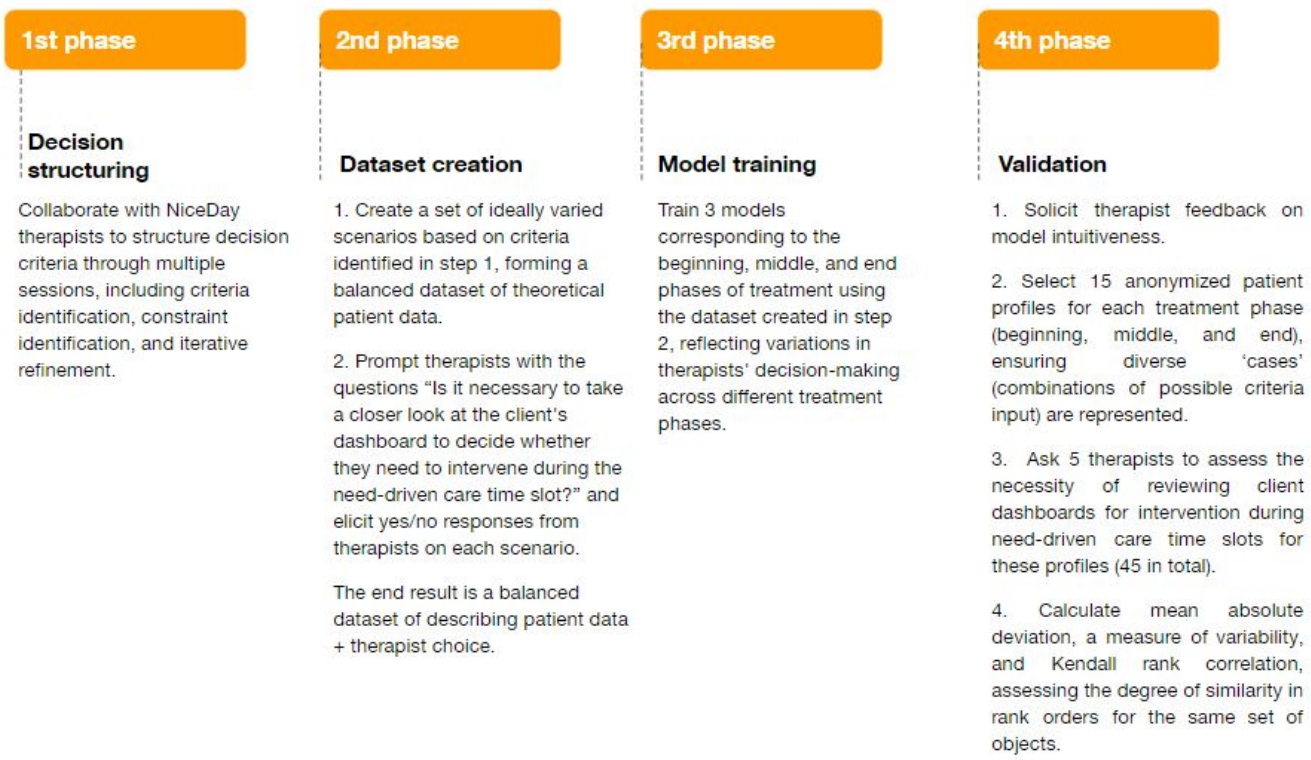
03 Feedback loop

Because the users of this system are also the domain experts, collection of observation level feedback on both model recommendations and model explanations can be used to iteratively improve the decision support system – including both the model powering the client ranking dashboard as well as the model explainers – thereby enhancing the system's performance, but also experts' trust in the system over time.

In addition to these system components, this project also explored and developed approaches to facilitate **incremental learning** of the recommendation model, ensuring learning remains in the control of system developers. All incremental changes should be understood before being incorporated into the system, resulting in improvement rather than degradation or drift. This includes when, by whom, and exactly where in the recommendation system improvement is needed (see Enabling a Human Expert Feedback Loop); it also explored approaches to quantify the effect of the model on expert behaviour and the effect, in turn, of those same experts providing feedback to retrain the system (see Quantifying Human in the Loop).

BAIT Model Development

Councyl worked closely with NiceDay therapists over the course of several months to structure and model the decision “Is it necessary to take a closer look at this client's dashboard to decide whether you need to intervene during the between session care time slot?”.



Councyl’s Behavioural AI Technology (BAIT) is a novel method of choice modelling beyond the state-of-the-art in supervised learning technology. The field of choice theory has traditionally focused on analysing and predicting the preferences of large groups of individuals (e.g. consumers) with respect to new options by developing micro-econometric models rooted in rational decision-making. With BAIT, it is possible to model expertise (rather than preferences) of small (rather than large) groups of experts (rather than consumers or citizens) in highly complex and dynamic (rather than simple and static) choice situations[1-4].

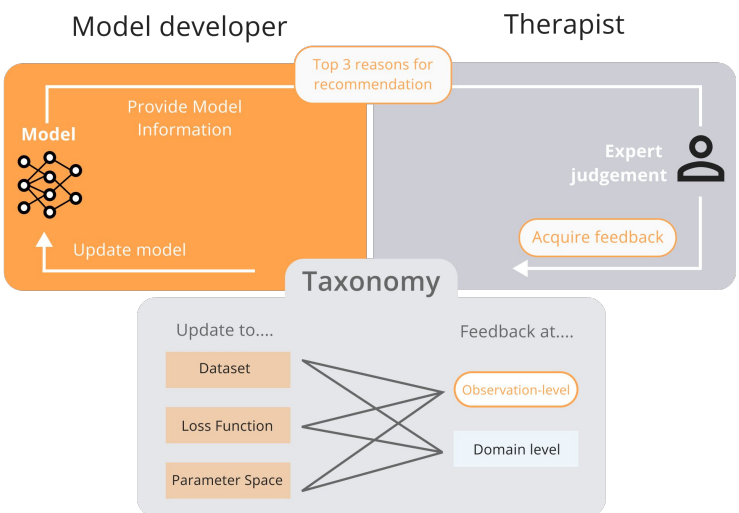
Importantly, because these models are developed using expert choice data (as described in the figure above), no historical data is required. Instead, an ideally balanced set of choice scenarios are presented to subject matter experts – in this case, therapists – who are asked to make a decision that represents the action they would take if they faced the same scenario from a real patient. Because the scenario set is ideally balanced, comparatively small datasets can create reliable models (roughly 300 choice scenarios). The resulting choice models, which explicitly capture implicit expert trade-offs, are inherently transparent and explainable: for all new decision cases fed into the model, the model’s prediction can be traced to a set of criteria weights that determine how the output was calculated.

Enabling a Human Expert Feedback Loop

Incremental learning: updating models through expert feedback

To make sure that the developed system works for the therapist, a feedback collection system was set up through the Deeploy platform. This feedback can then be used to retrain the model if it is not performing according to the expected standards of the practitioners, enabling incremental learning. Incremental learning, also known as online learning or lifelong learning, is a machine learning paradigm where a model is continuously updated and improved as new data becomes available. In traditional batch learning, a model is trained on a fixed dataset, and any updates or improvements require retraining the model on the entire dataset, which can be computationally expensive and time-consuming. Incremental learning, on the other hand, allows a model to adapt and learn from new data without the need for retraining from scratch.

To solve the problem, we investigated how to effectively make use of expert feedback on model prediction and explanation as a source for incremental learning. Feedback on model performance can either be domain-level feedback, which captures high-level conceptual feedback, or observation-level feedback, which captures how a model should behave on specific datapoints[5]. The feedback in turn can translate into different model updates, for example updates to the dataset (by using feedback as annotation), adjustment of parameters, or updates to the loss function[5].



Adapted from Chen et al., 2023

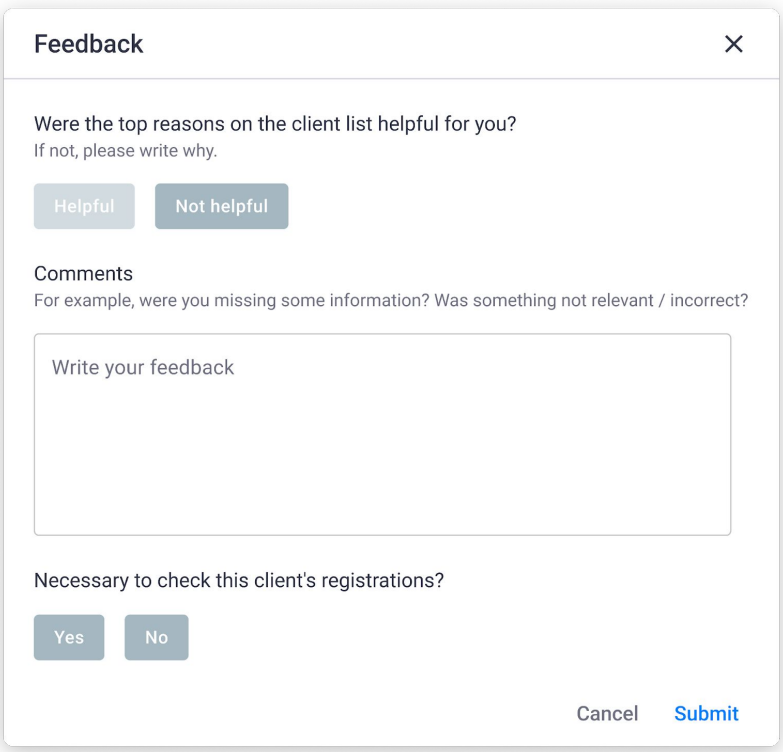
In this initiative, observation level feedback was provided by human experts. This was done not just for model recommendations, but also for the model explanations. The goals of XAI is to support a complete and sound understanding of the model[6], and to foster trust in the model[7]. Evaluations on explanations are especially important in this project as the users of the decision-support system (the therapists) are considered experts and the model itself is developed with the help of domain-experts. Hence, if the experts do not think the explanations were helpful it might indicate a need for better explanation methods or that the model reasoning does not align with expert opinion. This could be an indication that the model needs to be retrained.

Enabling a Human Expert Feedback Loop

Consumption and retrieval of feedback

The feedback loops were realized by connecting the NiceDay interface with API endpoints of Deeploy. For each recommendation, the experts can give feedback on whether they would follow the model's recommendation or not, and whether the explanation (top three features contributing to the recommendation) was helpful, using the integrated feedback screen in the NiceDay app. If the expert chose “not helpful”, they had the option to give additional feedback in free form text.

To use the collected feedback, Deeploy supports the retrieval of all model predictions and evaluations through the Deeploy API. To obtain the prediction logs, the model developers can choose a time frame indicating what period they want to collect the prediction logs from. They can also choose to only obtain logs for predictions that have been given feedback on. The retrieval of prediction logs is easily done through the Deeploy Python Client.



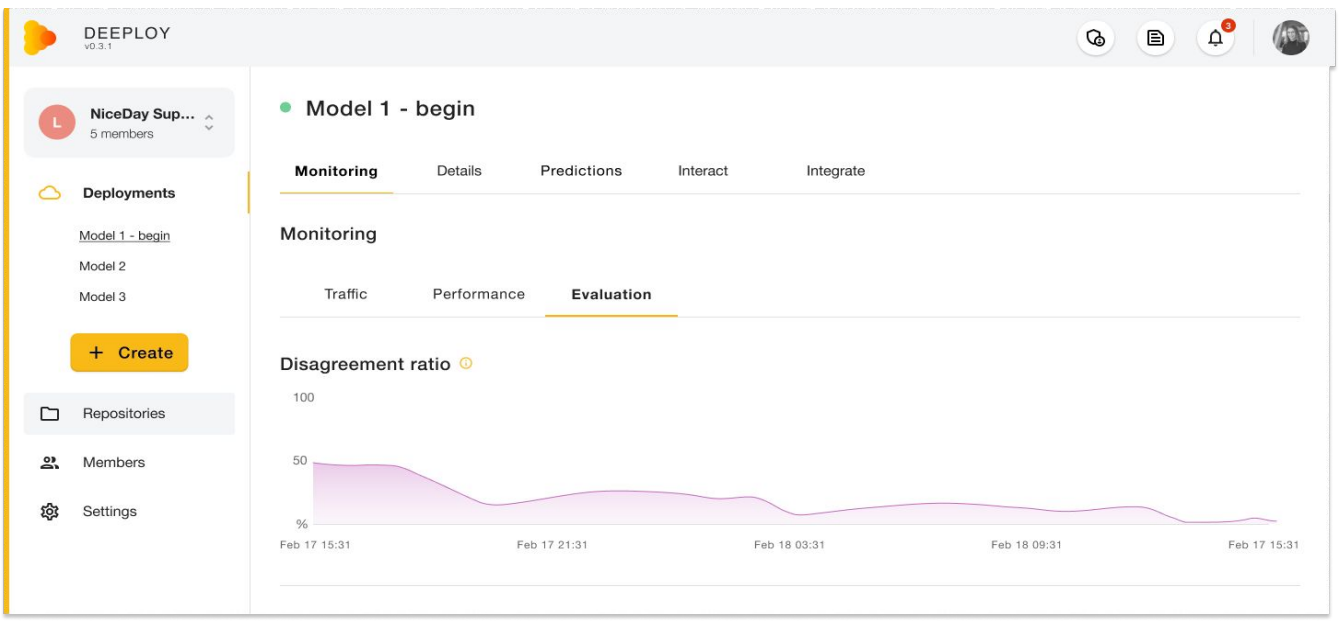
Challenges in incremental learning based on human feedback

Using human feedback for incremental learning presents several challenges. Firstly, measuring the performance of a new model based on user and expert feedback can be difficult, as it's often challenging to define what counts as an impactful update[5]. Secondly, determining who qualifies as an expert and has the authority to influence the model is another critical consideration[8]. Thirdly, balancing stability and flexibility is crucial; a model must be stable with its training data but flexible enough to adapt to new information without being overly rigid or volatile[9]. This is called the stability-flexibility dilemma and means that model updates need to happen at the right time, based on an aggregation of expert feedback that shows what direction model updates should take. Hence, feedback needs to be easily digestible so that the model owner can easily understand the feedback given and the status of approval or disapproval of model predictions and explanations.

Enabling a Human Expert Feedback Loop

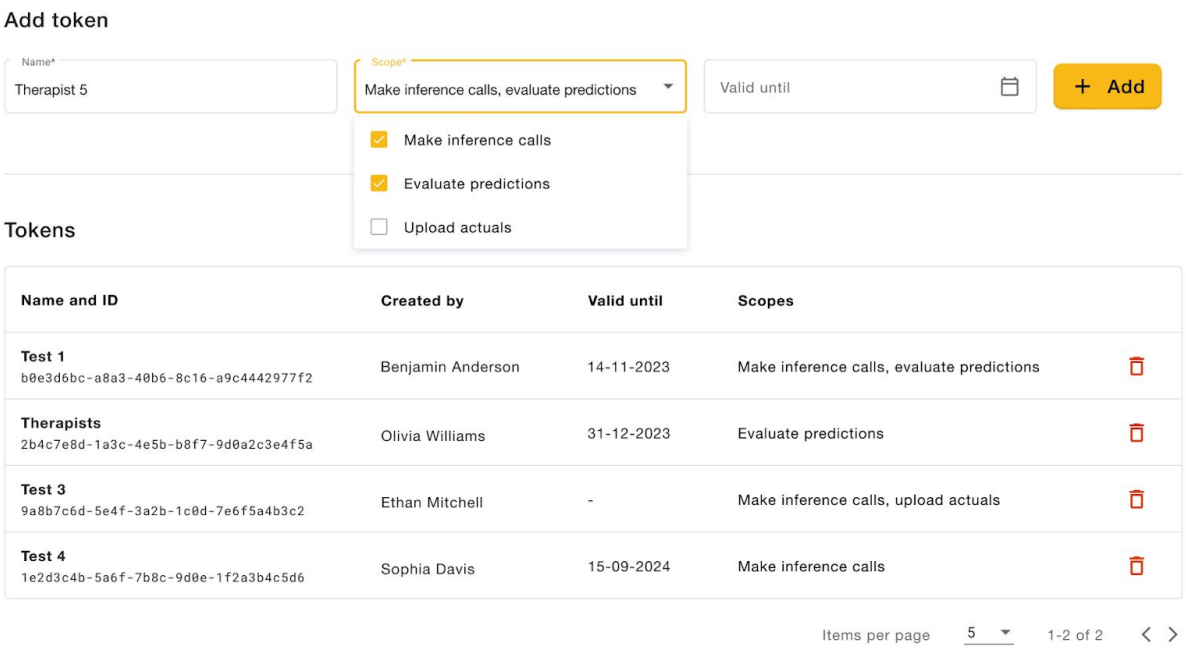
Enabling an effective human feedback loop in Deeploy

Deeploy as a model management platform already comes with a few solutions that can help answer the challenges of incremental learning with human feedback. For measuring the performance of an updated model, model owners can assess how human experts approve of the model by referring to the disagreement ratio as visualized in the Deeploy platform (below). The disagreement ratio describes how many of the evaluated instances have a negative evaluation, meaning how many of the evaluated predictions the user thought were incorrect. In the Deeploy platform, the user can filter what time period they would like to compare, meaning they can compare the disagreement ratio before and after the model update. This can inform whether the update to the model was impactful. It is also possible to set an alarm for the disagreement ratio, meaning model owners get notified when the disagreement ratio gets too high, which can inform when to perform a model update. This makes balancing the stability and flexibility of the model easier.



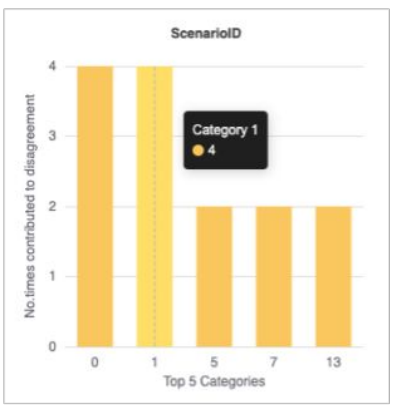
To ensure that only feedback from trusted experts are used, Deeploy supports the creation of different tokens for predictions and feedback. Users can have the authority to make predictions but not to make evaluations for example. This way it becomes easy to manage who gets a say about the model performance. If the model owner wants to collect feedback both from expert users and regular users, they can create tokens with evaluation authority for both, but add additional information in the token description, such as creating standardized tags like 'ExpertUser' or 'RegularUser' that can then be sorted on when performing feedback analysis. A figure showing how tokens look like in the Deeploy platform can be seen in the figure on the next page.

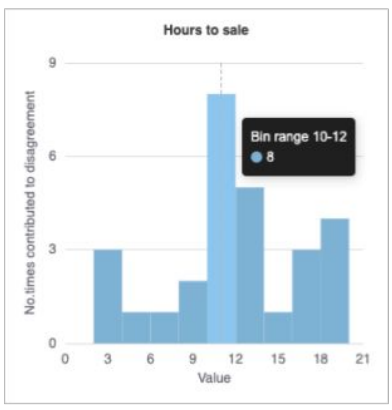
Enabling a Human Expert Feedback Loop

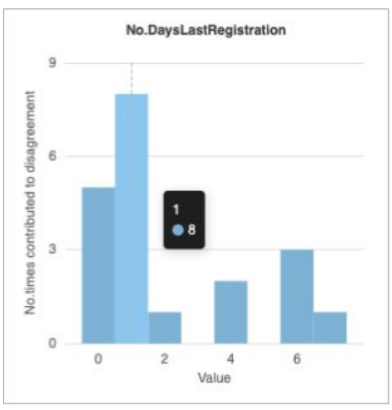


Feature level visualization of feedback

During this initiative, Deeploy also developed a prototype to be able to visualize more clearly what values of each feature were involved in the evaluation of the prediction or explanation. This means that the model developers can reason for each variable about what values of the variable most often seem to contribute to an agreement or disagreement. For example, if for a continuous variable, it is always values in a certain bin that are involved when the experts disagree with the model, it could indicate that the model is not capturing this part of the data sufficiently. This visualization can help guide the data scientist to what actions to take next, for example perform further feature engineering or data collection, or choose a more complex model. Hence, this new feature can help guide incremental learning by giving data scientists some initial information that they can use to investigate further to find the specific problem area for the task in question, as shown below.



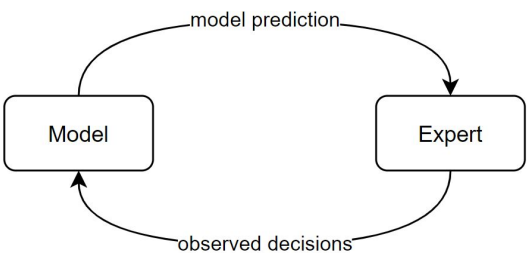




Quantifying the Human-in-the-Loop

Human-in-the-loop (HITL) AI systems are a class of artificial intelligence systems that involve human intervention or oversight in various stages of the AI's operation, such as data collection, model training, decision-making, or quality control. These systems harness the complementary strengths of both humans and AI, leveraging human expertise and judgement to improve the performance, safety, and ethical considerations of AI applications. For example, social media platforms like Facebook and Twitter use HITL AI systems to filter and moderate user-generated content. AI algorithms initially flag potentially inappropriate content, and human moderators make final decisions on whether to remove or allow it. HITL systems aim to strike a balance between the capabilities of AI and the nuanced judgement of humans, making them crucial in many applications where both efficiency and reliability are required.

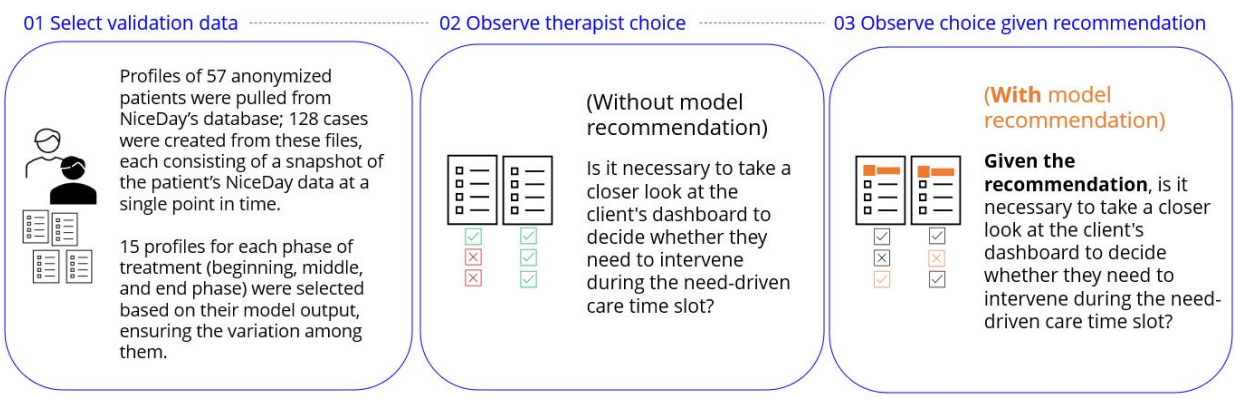
In the application of HITL decision support systems, a two-way interaction between decision models and users occurs. Experts as users receive a recommendation based on a model; the extent to which the experts follow the model prediction is determined by their dependency on the model. After decisions are made, these observations serve as new data that can be used to recalibrate models powering the recommendation engine.



In this project, Councyl set out to measure the impact of including human experts in the recommendation system – that is, to quantify the human-in-the-loop. We approach this from two perspectives:

1) To what extent do experts depend on the recommendation engine in making their decisions?

2) How does the experts' decision influence recalibration of models as a part of incremental learning?



To conduct this analysis, after performing the model validation tasks (steps 1-2), therapists reviewed **the same decision cases** included in the validation (step 3), but this time with the benefit of the model's recommendation, and again indicated their choice.

Quantifying the Human-in-the-Loop

Expert dependency on recommendation engine (1/2)

Experts' dependency on the model's prediction is defined by two parameters: a) Experts' constant dependency, and b) experts' variable dependency (e.g., the more confident the model is, the higher the dependency).

Two types of estimations were performed: overall and phase-specific dependency. For overall dependency, the parameters were estimated based on the whole dataset (N=270). Three variations are applied: (i) estimation of a and b simultaneously, (ii) estimation of b while a is kept at zero, and (iii) estimation of a while b is kept at zero. For phase-specific dependency, the parameters were estimated based on the dataset from each phase with the same set of estimation variations (Table 1).

Table 1: Parameter estimates for a and b

Type	Variation	*a*	(p-value)	*b*	(p-value)
1. Overall (N = 270)	1.i	0.067	(0.754)	0.237	(0.606)
	1.ii			0.349	(0.313)
	1.iii	0.164	(0.327)		
2. Phase-specific: Beginning (N = 90)	2.i	-0.025	(0.943)	0.518	(0.536)
	2.ii			0.476	(0.397)
	2.iii	0.178	(0.478)		
3. Phase-specific: Middle (N = 90)	3.i	0.078	(0.831)	0.098	(0.879)
	3.ii			0.218	(0.550)
	3.iii	0.126	(0.543)		
4. Phase-specific: End (N = 90)	4.i	0.102	(0.795)	0.329	(0.754)
	4.ii			0.551	(0.446)
	4.iii	0.204	(0.453)		

To interpret the parameters, take variation 1.a as an example. Three scenarios are presented. Each scenario is described with hypothetical initial experts' decision probability and model prediction. Then, the final experts' decision probability is calculated using the formula as defined in the approach.

The expert's final decision probability P_{final} is assumed to be a weighted average between their initial decision probability $P_{initial}$ and the model prediction P_{BAIT} as defined[10] by $P_{final} = \lambda P_{BAIT} + (1- \lambda) P_{initial}$ where the weight λ is the expert's dependency on the model as defined by $\lambda = a + 2b P_{BAIT} - 0.5$.

Quantifying the Human-in-the-Loop

Expert dependency on recommendation engine (2/2)

Experts' dependency on the model is observed, however, with all p-values above 0.05, the observed dependency **is not statistically significant**. This data was collected in the first instances of experts using the recommendation system; more data, in addition to increasing the possibility of finding a dependency of significance, may also increase dependency as experts become accustomed to using the system.

If the parameters were statistically significant with estimates as observed in variation 1.a (non-phase-specific, includes a and b), the model would only influence the expert to change their decision in a scenario 3 (where the expert is doubtful in a certain decision and the model prediction is confident in the opposite decision).

Table 2: Impact of 3 expert dependency scenarios

Scenario 1: **Confident expert, doubtful model prediction**	**Scenario 2:** **Confident expert, confident model prediction**	**Scenario 3:** **Doubtful expert, confident model prediction**
Given initial experts' decision probability $P_{initial}$ of 0% and model prediction P_{BAIT} of 50% the final experts' decision probability P_{final} would be 3.5%.	Given initial experts' decision probability $P_{initial}$ of 0% and model prediction P_{BAIT} of 100% the final experts' decision probability P_{final} would be 30.4%.	Given initial experts' decision probability $P_{initial}$ of 45% and model output P_{BAIT} of 100% the final experts' decision probability P_{final} would be 62%.
When the expert is confident in a negative answer, and the model prediction is doubtful, there is very little change (increase) in the expert's decision probability.	When the expert is confident in a negative decision, and the model is confident in predicting a positive decision, there is a change (increase) in the expert's decision probability, but not enough to change it to above 50% (the threshold for a 'positive' recommendation used for validating model predictions).	When the expert is doubtful in a negative decision, and the model is confident in predicting a positive decision, there is a change (increase) in the expert's decision probability, enough to change it to above the 50% threshold (in other word, to change a "no" choice to a "yes").

The approach described here is suitable to observe experts' reliance on recommendation engines, with interpretable results. Notably, periodically rerunning this analysis is recommended to ensure experts have adequate opportunity to become accustomed to the recommendation system, such that true or developing dependency can be observed. This is related both to experts' dependency being likely to change over time, as well as ensuring adequate data is available to observe any significant effect.

Quantifying the Human-in-the-Loop

Experts' influence on model recalibration (1/2)

A new decision model was created for each phase of treatment. The training data for each model consists of observations collected in the choice experiment (comprising two-thirds of training data) and the 270 observations from real (anonymised) patient cases, specifically the experts' decisions with the model predictions (comprising the remaining one-third of training data). The dataset for beginning, middle, and end phase models consists of 300, 279, and 279 observations respectively.

The initial and newly trained models were then compared in terms of the inclusion of significant decision criteria and model performance. The parameter size/importance of the criteria cannot be compared, because they are estimated on different datasets. The model performance is measured in terms of hit rate, sensitivity, and specificity. To measure the performance of the initial model, the real-life observations from the relevant phase are used as the test dataset. To measure the performance of the new model, due to the small amount of data, a K-fold validation (with K of 100) was performed.

Inclusion of criteria

The biggest difference is observed in the beginning-phase models (above). With four criteria removed in the new model, this may be due the combination of relatively low effects thereof and much less variation in the real-life data – especially because the criteria in the choice experiment data is systematically varied. For example, in criteria *'Completion of the diary in the past 7 days', (Voltooiing van het dagboek in de afgelopen 7 dagen)* 80% of the observations have a value of *'Extremely low score (below 4.5 points) on one or more of the 5 aspects'*.

Quantifying the Human-in-the-Loop

Experts' influence on model recalibration (2/2)

Table 3. Changes in model criteria with retraining		
Beginning-phase model	*Middle-phase model*	*End-phase model*
Three out of seven criteria in the initial model remain significant in the new model. Four other criteria of the initial model are no longer significant in the new model. One additional criteria has become significant.	Five out of seven criteria in the initial model remain significant in the new model. Two other criteria of the initial model are no longer significant in the new model. One additional criteria has become significant.	Four out of five criteria in the initial model remain significant in the new model. One criteria of the initial model is no longer significant in the new model No additional criteria become significant.

Model performance

The performance of all models are presented in Table 4. The performance depends on the threshold used to convert the continuous model prediction (0%-100%) to binary prediction (yes/no), which is displayed in the first row.

Table 4. Model performance of initial and new models for all treatment phases						
Phase	**Beginning (threshold: 50%)**		**Middle (threshold: 70%)**		**End (threshold: 40%)**	
Model	Initial	New (std)	Initial	New (std)	Initial	New (std)
Accuracy[1]	70%	68% (6%)	64%	60% (6%)	72%	73% (6%)
Sensitivity[2]	67%	69% (16%)	63%	60% (6%)	78%	75% (10%)
Specificity[3]	73%	70% (20%)	69%	61% (11%)	64%	72% (10%)

[1] Accuracy : (TP + TN) / (TP + TN + FP + FN)
[2] Sensitivity : TP / (TP + FN)
[3] Specificity : TN / (TN + FP)
(TP: true positive, TN: true negative, FP: false positive, FN: false negative)

For all phases, the performance of the initial and new models are relatively similar, with the performance of the initial model falling within the range of new model performance plus/minus the standard deviation.

Application

In incremental learning, adding new observations to retrain models may change the performance of the decision support, and it can become better or worse. This may be especially true for models initially trained with datasets that are **more balanced** than real decision cases.

By **quantifying** the precise change that occurs when a model is incrementally improved using data from real cases – including those where a human decision maker overrides a recommendation – model developers can make informed design choices suited to their specific decision support needs and environment. For example, is an incremental learning approach appropriate, or is traditional (batch) retraining more suitable? Is ideally balanced choice data preferable to historical data, or can a combination of both strike the right balance?

Results in Practice

NiceDay therapists have started to use the eXplainable AI system in Personalized Mental Health care recently, in a testing environment. With the prioritised client dashboard, therapists can quickly have the overview of their clients and see who needs to be checked-in with.

These names are fictional clients

Moreover, because therapists are able to provide feedback on the XAI prioritisation directly in the dashboard, the models and explainers are incrementally improved.

"The XAI system helps you, and you help the XAI."

Wouter Schippers
NiceDay Psychologist

NiceDay is continuing to improve the XAI recommendation system. This includes gaining therapist feedback on the current iteration for incremental improvement, as well as expanding the scope of the XAI system.

About the partners

NICEDAY

NiceDay is a leading (online) mental healthcare platform in the Netherlands, aiming to strengthen the mental wellbeing of as many people as possible.

NiceDay introduced the NiceDay way - a new and innovative way of working based on evidence-based practices in psychopathology such as CBT, experience sampling methodology (ESM) and feedback informed treatment. Therapists make use of the NiceDay platform's tools and features that facilitate this way of working

DEEPLOY

Deeploy has developed a novel, best-of-breed MLOps platform for serving, compliance, monitoring and explaining AI models and decisions, including human interaction with models and a feedback loop in order to control models and improve on expert feedback. By providing both explainability and feedback loops, Deeploy helps experts stay in control and gain understanding of their AI-enabled tools, without compromising on transparency, control, and compliance.

COUNCYL

Councyl makes a decision management platform that enables organisations to build, serve, and manage transparent decision models based on expert choice – rather than historical data. With Councyl, organisations make in-house expertise digitally available in a straightforward and auditable way. Our aim is to elevate the human element to provide control & accountability to our customers' most sensitive decision processes.

Acknowledgements

This work was supported by RVO funding under project number: MIT-AI-22-03040414.

Citations

1. Chorus, C. G., Arentze, T. A., & Timmermans, H. J. (2008). A random regret-minimization model of travel choice. Transportation Research Part B: Methodological, 42(1), 1-18.

2. van Cranenburgh, S., Guevara, C. A., & Chorus, C. G. (2015). New insights on random regret minimization models. Transportation Research Part A: Policy and Practice, 74, 91-109.

3. Chorus, C. G., Pudāne, B., Mouter, N., & Campbell, D. (2018). Taboo trade-off aversion: a discrete choice model and empirical analysis. Journal of choice modelling, 27, 37-49.

4. Chorus, C., van Cranenburgh, S., Daniel, A. M., Sandorf, E. D., Sobhani, A., & Szép, T. (2021). Obfuscation maximization-based decision-making: Theory, methodology and first empirical evidence. Mathematical Social Sciences, 109, 28-44.

5. Chen, V., Bhatt, U., Heidari, H., Weller, A., & Talwalkar, A. (2022). Perspectives on incorporating expert feedback into model updates. Patterns, 4.

6. Kocielnik, R., Amershi, S., & Bennett, P.N. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

7. Cheng, H.F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F.M., & Zhu, H. (2019). Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders. Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems.

8. Park, H., Megahed, A., Yin, P., Ong, Y.J., Mahajan, P.D., & Guo, P. (2023). Incorporating Experts' Judgment into Machine Learning Models. Expert Syst. Appl., 228, 120118.

9. Lin, G., Chu, H., & Lai, H. (2021). Towards Better Plasticity-Stability Trade-off in Incremental Learning: A Simple Linear Connector. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 89-98.

10. Wijnands, M. (2022). Behavioural Artificial Intelligence Technology: Designing a Bayesian Approach and Investigating the Feedback Loop [Master's thesis, Erasmus University Rotterdam].