Progress in Artificial Intelligence (2019) 8:475–485
https://doi.org/10.1007/s13748-019-00192-0

REGULAR PAPER

Learning machines in Internet-delivered psychological treatment

Magnus Boman[1] · Fehmi Ben Abdesslem[2] · Erik Forsell[3] · Daniel Gillblad[2] · Olof Görnerup[2] · Nils Isacsson[3] · Magnus Sahlgren[2] · Viktor Kaldo[3,4]

Received: 12 February 2019 / Accepted: 15 May 2019 / Published online: 29 May 2019

Abstract
A learning machine, in the form of a gating network that governs a finite number of different machine learning methods, is described at the conceptual level with examples of concrete prediction subtasks. A historical data set with data from over 5000 patients in Internet-based psychological treatment will be used to equip healthcare staff with decision support for questions pertaining to ongoing and future cases in clinical care for depression, social anxiety, and panic disorder. The organizational knowledge graph is used to inform the weight adjustment of the gating network and for routing subtasks to the different methods employed locally for prediction. The result is an operational model for assisting therapists in their clinical work, about to be subjected to validation in a clinical trial.

Keywords Learning machine · Machine learning · Ensemble learning · Gating network · Internet-based psychological treatment

1 Introduction

Machine learning is here employed to help answer questions concerning prediction of outcome and engagement in psychological treatment. The purpose is to learn how to successfully assist therapists in their daily work in delivering treatment, using machine learning, and resting on a unique historical data set involving more than 5000 patients. As the interesting questions as well as the data at hand are complex, a myriad of machine learning methods and algorithms are put to use for analyses of historical as well as new data. The data set goes back more than 10 years and is in a sense complete: relatively few details about the treatment and the entities constituting its environment are beyond reach, because the care is Internet-delivered self-help material and all communication between patient and therapist is recorded as text messages. Each machine learning method employed helps identify and amplify signals of bias, but due to the different nature of data points (e.g., standardized questionnaires, long texts, logs of system use) no machine learning method can be used to analyze *every* weak signal well enough to address the task of predicting future patient behavior and the clinical outcome of treatment. For this reason, a number of machine learning methods are used in tandem, with their signal analyses fused and unified to produce decision support for the clinician. The output of one method may be the input to another method in this model. Because new data are added over time, the fusion and unification procedures may adapt dynamically. This fact does not prompt any changes to the employment of individual machine learning methods, but instead requires efficient meta-learning. A learning machine is therefore employed, which takes into account the dynamic success rate of the individual methods and their relevant combinations. The object here is to describe that learning machine, a form of super learner, at the conceptual level. We start by describing this machine and the data at hand. We then describe related research and the computational model and how it relates to the knowledge graph of the clinic. Some of our initial experiments are then described and discussed, before conclusions and some pointers to next steps.

✉ Magnus Boman
mab@kth.se

[1] KTH/EECS/SCS/MCS, Electrum 229, 16440 Kista, Sweden

[2] RISE, Box 1263, 16429 Kista, Sweden

[3] Department of Clinical Neuroscience, Centre for Psychiatry Research, Karolinska Institutet and Stockholm Health Care Services, Stockholm County Council, Sweden

[4] Department of Psychology, Faculty of Health and Life Sciences, Linnaeus University, Växjö, Sweden

1.1 Learning machines

A learning machine can be defined as an autonomous self-regulating open reasoning system that actively learns in a decentralized manner, over multiple domains [6], or from one abstract model of its world to another. The basic definition of machine learning is as follows. “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E” [19, 2]. A learning machine, by contrast, not only learns, but also meta-learns, since it has to adjust its behavior according to the individual machine learning modules that it continually receives and assesses output from. This leads to: A computer program learns to learn if its performance of each task in T improves with experience E and also with the number of tasks (cf. [29]). Feedback allows it to self-correct its models and its processes of learning. Autonomy yields the possibility of sustained autonomous learning [14]. Openness indicates that new machine learning modules could be added at any time to the learning machine. That its learning is active means that it can pursue learning goals without explicitly being told to do so, guided by self-testing in accordance with the principles of operant conditioning [27].

Important aspects of learning machine “education” were mapped out by Turing [31]. The machine trains and dynamically adjusts its perception, after which it can move into reasoning, which requires further education, and then to interaction. The machine can then interact with humans, e.g., by presenting a prediction of an outcome, or by explaining its reasoning steps. The behavior of a learning machine is ideally interpretable to authorized observers, i.e., they should be able to follow its reasoning and not accept it as merely an educated black box. In some cases, less than optimal performance (cf. [21]) could be accepted to provide interpretability, e.g., when explicating a model by reference to its education from training data [17].

The learning machine has internal states, and rules for switching between them, like an ordinary Turing machine. Intuitively, each patient in treatment will pass through different states of a much simpler representation, in the form of finite automata: one automaton per patient. An end state can correspond to labels indicating, e.g., level of engagement with the treatment program, or successful response to it. To learn under which circumstances state-switching in the learning machine controlling this array of automata occurs is an important component of meta-learning. For a self-supervised Turing machine to be able to study its own tape operations and state changes, self-observation must be possible. This separates the learning machine from the much simpler machine learning modules that it incorporates, as finite automata are too simple models for managing introspection. The learning machine will not monotonically improve its performance over time and over tasks, but will have test modes in which experimental reasoning and inference will be self-evaluated and possibly lead to revisions of state-switching rules, e.g., probability or utility values.

To generalize from one domain to another, as in, e.g., switching between groups of patients having been referred to different Internet treatment programs, learning machines must adopt their learning to the settings in which they are situated. Such multitask learning is realized by means of inductive bias created from training inputs of related tasks [8]. This allows for perception, reasoning, and interaction about things pertaining to more than one domain [13]. In return, the reasoning may be directed by meta-rules that help steer cross-domain inference. Such steering involves ethics, norms, rules, laws, and all other forms of constraints. The interplay between perception, reasoning, and interaction makes reductionist approaches unsuitable for learning machines. Instead, a systemic methodology in which the machine can learn more than the sum of its learning from each task performed, should be adopted for its design. Given the generality of a learning machine, very few currently exist, and for very specialized tasks, or subject to meta-level constraints on data or its distributions [21]. Our aim is therefore to provide a constructive example of what a learning machine can be and how it can be applied to a case of practical significance. The hypothesis is that a learning machine can provide robust decision support to clinicians. This hypothesis will be evaluated in 2020 in a large and pre-planned randomized controlled trial. The confirmatory role of that evaluation will determine the importance of the contribution from the clinical side, thereby assessing the value of the exploratory search for regularities in the data.

1.2 Data

The data analyzed by the learning machine in the here presented case study are unique in that it represents virtually all of the information passed between patients and therapists in a particular mental health care setting. Because care is Internet based, there are few subtle pieces of information missed out on. Given such a special data set, why is it reasonable to think that learning machines can contribute to the task of drawing important conclusions about patients? Deductive inference from data can be valuable for clinical decision support (in spite of the fact that deductive methods only make explicit what is already there) because patient data are multimodal. Methods for multimodal fusion let different modalities be represented in one database (or knowledge base, or expert system component, or some other external explicit representation), allowing for new propositions to be deduced that could not be deduced from either of its constituents. Inferring unknowns from knowns, where the latter come in the form of data, is the general challenge for drawing useful inductive conclusions. To generalize from what has been observed in order to predict future data with a minimum of error is done through classification or regression. Precisely which results pattern matching methods can achieve has been mapped out in statistical learning theory [32].

Data are not the sole resource for machine learning, as in most clinical applications there is expert knowledge available on, e.g., diagnoses, treatments, substances, and prevention. In the ideal situation, all such knowledge is assembled into a knowledge graph, indicating how pieces of knowledge are interrelated. Even if knowledge elicitation can be costly in that clinical staff is a scarce resource, it is generally worth the effort, as the knowledge graph is so useful. Knowledge fusion therefore sometimes involves knowledge from experts being integrated into learning machines, e.g., for bootstrapping training by finding reasonable initial weights instead of just starting from random values.

2 Methodology

2.1 Application case

A unique opportunity for testing learning machines recently arose at the Internet Psychiatry Clinic at *Psykiatri Sydväst* in Stockholm. Since 2008, the clinic has offered Internet-based Cognitive Behavioral Therapy (ICBT) for depression, social anxiety, and panic disorder with documented significant treatment effects [30]. The clinic would in the future like to predict, as early as possible, if the current treatment will have a positive effect on the primary problems the treatment aims to reduce, e.g., on symptoms of social anxiety. Over time, the goal of introducing a learning machine for ICBT decision support will become an important part of an adaptive treatment strategy that has already been tested for one of the treatments [11], but could be further improved.

At the clinic, 10-week ICBT interventions with diagnose-specific content based on established CBT techniques are divided into step-wise modules, with therapist support chiefly via asynchronous messages on a secure platform. The therapists normally see their patients face to face neither during treatment nor during intake and follow-up. The ICBT platform collects large amounts of mostly self-reported data for each patient before, during, and after treatment. Both established self-reporting questionnaires and more anamnestic and open-ended questions are used. Patients are largely self-referred and apply for ICBT via a public eHealth platform. A screening is then completed, including standard symptom scales for depression, social anxiety, and panic disorder, which also constitutes the primary outcome for each respective diagnose-specific ICBT program. Also included are scales for insomnia, general quality of life and functioning, use of health care, work ability, sick leave, and domestic functioning, problematic use of alcohol and drugs, and ADHD symptoms. Many of these measurements are then repeated at treatment start as well as post-treatment. Structured data from the psychiatric assessment are summarized, registered, and quantified in the platform together with some administrative data. During treatment, weekly questionnaires of primary symptoms, depression, and suicidal ideation are filled out, and data from interactive worksheets, patients' reports on their use of therapeutic techniques, and messages between therapist and patient are collected. Moreover, data on the patients' behavior within the platform, for example number of logins and number of clicks, and accessed treatment modules, are also recorded for possible future use.

Covering relevant parts of the clinic, a conceptual model in the form of an entity–relationship (ER) diagram [10] was built to represent the ICBT information flow (Fig. 1). This abstract model describes all events (e.g., a self-evaluation questionnaire being submitted) that result in bursts of information being added, and it thus constitutes a dynamic model useful to suggesting types of data susceptible to analytics [5]. The model has been consolidated with ICBT experts and serves as an ontological frame for discussing data and processes in a structured and unambiguous way. It constitutes an important step toward a finished knowledge graph for the clinic.

2.2 Related research

Machine learning has only recently been introduced in psychiatric research, for example in suicidality prediction, using fMRI to distinguishing CBT responders from non-responders, and measuring brain resting-state network connectivity to predict responders in electroconvulsive therapy [23]. Automated analyses of free speech have been able to predict later development of psychosis [2] and to estimate long-term severity of depression [16]. Another study used mostly patient self-reports to identify those likely to respond to a specific anti-depressant and validated the results in an external sample [9]. Machine learning methods can thus successfully learn to predict prospective events that are clinically relevant, and this can enable the identification of patients who are likely to respond to a particularly well-standardized intervention. The predictive power depends on the number of available predictors and their level of independence, data quality, sophistication of the learning algorithms, and how well they match the nature of data. In the design of a learning machine for ICBT, each classifier and each component of predictive analytics that is deemed relevant to test by the domain experts is a potentially valuable input signal.

The learning parameter space and the combinatorial space of processing multiple input signals in order to produce a maximally useful output signal are both hard to characterize mathematically, making empirical exploratory work impor-

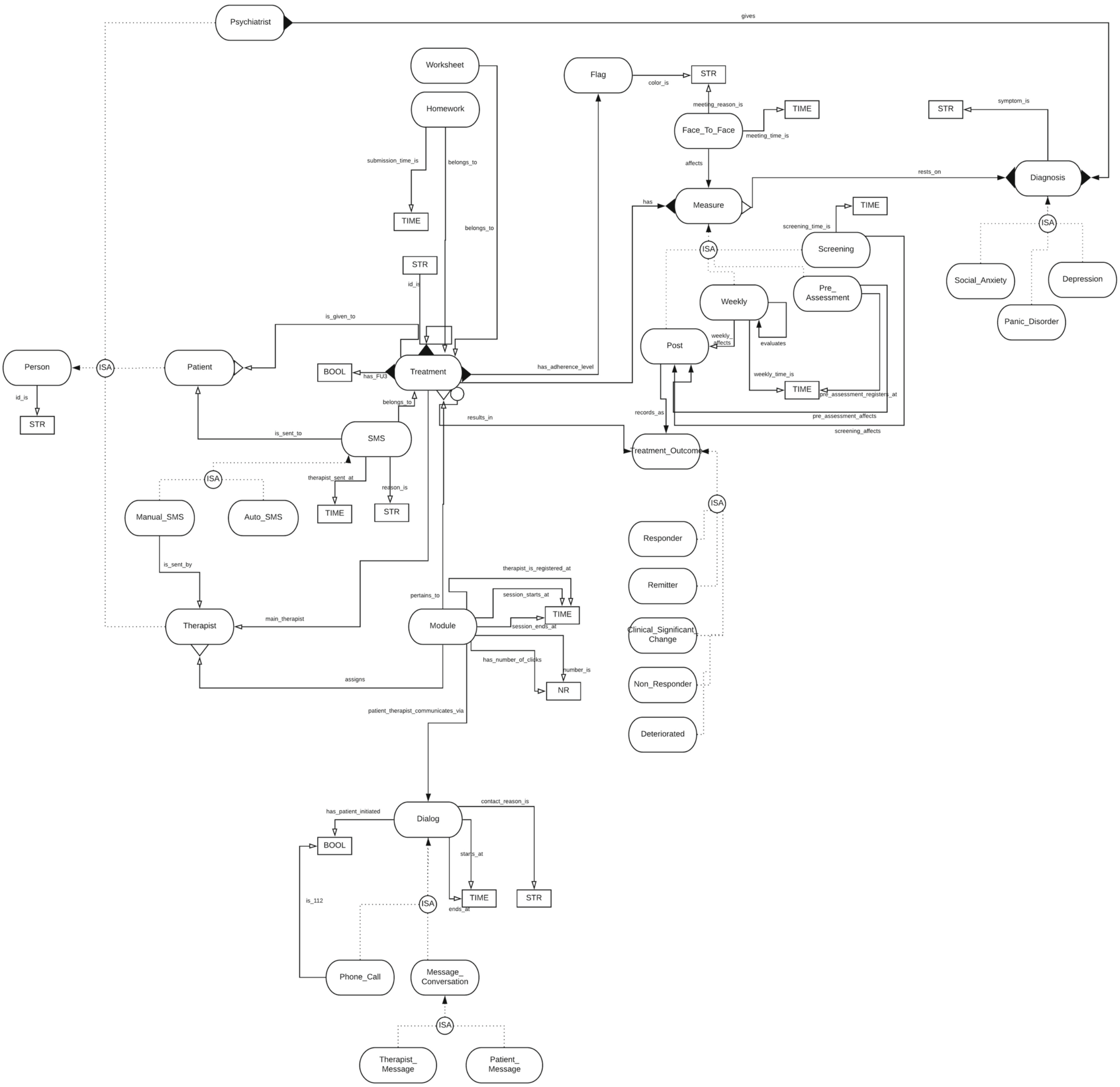

Fig. 1 A conceptual model depicting the information flow in treatment management. In this graphical version of the model, events and rules are not shown, in the name of clarity. They do constitute important types of information in the organization knowledge graph, the details of which are still under discussion. Ellipses represent non-lexical objects, and rectangles represent lexical objects. Non-lexical objects are entity types or relation types, while lexical objects are data types. Links represent attributes and have directions and mappings (total or partial, one or many). Inheritance hierarchies are represented by ISA-links, as in "Patient is a Person"

tant as a complement to formal arguments. Grounded in a clinical application, such progress can be slow and costly, as improving clinical outcome and counting costs produce uncertain results [7]. Learning machines that perform well on a task do so in part because the domain-specific bias obtained from performing earlier tasks has been coded efficiently. In the case analyzed here, the relatively large amount of patients will allow a learning machine to (with some still unknown rate of success) solve the meta-level learning problem [24]. The learning machine learns features with the help of expert bias (cf. [1]), and the general idea used to fuse and unify knowledge here was first proposed by Jacobs and Hinton in 1988 to overcome problems with interference between subtasks when using backpropagation [15, 79]:

> If one knows in advance that a set of training cases may be naturally divided into subsets that correspond to distinct subtasks, interference can be reduced by using

> a system composed of several different "expert" networks plus a gating network that decides which of the experts should be used for each training case. …The idea behind such a system is that the gating network allocates a new case to one or a few experts, and, if the output is incorrect, the weight changes are localized to these experts (and the gating network).

2.3 Model

In the perception–reasoning–interaction loop, the machine will store perceived input signals and recall them as part of its reasoning. The directly visible actions of the machine are restricted to its interaction; answering a specified set of questions, such as *Will this patient receive enough symptom reduction?*, sending out private messages to the therapist, and similar. In the future, an example of a question-driven activity (pertaining to a future in which the adaptive behavior platform is in place in full) would be a therapist asking the machine the equivalent of the human to human question *Will the patient 2535 complete the planned exposure?* An example of autonomous reasoning leading to a message being sent from the learning machine to a therapist would be *Please check status of Patient 2535 with respect to possible dropout*.

Previous research done at the Internet Psychiatry Clinic showed that simple regression using change scores on patients' primary symptom measure 4 weeks into treatment can explain between 34 and 43% of the variance in outcome at the end of treatment [26]. This indicates that moving away from baseline only predictions to instead monitor progress and predict outcomes after treatment has started can yield large information gains. Inspired by those findings, and by previous research on continuous monitoring of psychological treatment [18], the concept of an adaptive treatment strategy has been tested in a randomized controlled trial at the clinic [11]. In this trial, patients undergoing ICBT for insomnia were classified as at risk of failure (i.e., not benefiting from treatment) or not 4 weeks into treatment. The classifier was a multi-step algorithm using patient rating as well as clinician ratings in a simple spreadsheet that made calculations based on coefficients from previous predictor studies and rules based on clinical experience and hunches put in by the researchers (i.e., a rather rudimentary procedure from a statistical and computational point of view). Those classified as at risk were then randomized either to continue treatment as normal or to get extra attention and support from their therapist. Out of those not classified as at risk, 23% ended treatment with a poor outcome, whereas 64% of those classified as at risk who did not get extra help ended treatment with a poor outcome, indicating a clinically meaningful accuracy of the classifier. Those classified as at risk who did get extra help only had poor outcomes in 37% of cases, indicating that the predicted failures can be avoided in many cases. The trial by Forsell et al. [11] shows that waiting a few weeks into treatment and then making even a relatively simple prediction can have a large clinical impact if that information is clearly stated and then acted upon by the therapist. These studies demonstrate the potential of predictions using data from both baseline and the early weeks in treatment. However, apart from less than perfect predictions, Schibbye et al. [26] leave a lot of available information unused, and Forsell et al. use an algorithm that is partially manual and often requires input and effort from the therapist. Therefore, there is still room for improvement that could potentially be partially covered with a learning machine.

A finite number of different machine learning methods are employed for basic tasks of classification and prediction, and the learning machine is thus using ensemble learning [22] for efficient fusion and unification of the individual methods. To classify patients into those that responded well to the program and those that did not, for instance, a score function from the data into a set of expert classifications and assessments is defined. The function is typically parameterized and linear, and its value range is sometimes referred to as the ground truth since it is against these labels any machine learning method will be validated initially. A loss function that measures the discrepancy between this ground truth and the predictions given the trained parameter set is defined next. Because the loss function is defined over a high-dimensional space, even a linear classifier weight matrix will have tens of thousands of elements. To explore this feature space, it is possible to fix points (e.g., via a random weight assignment) and then travel along lines (dim 1) or planes (dim 2), computing loss along the way. The final step is to optimize by minimizing the loss. The overarching representation in the here presented model will be a distributed probabilistic continuous neural network with iterative parameter tuning [3]. The methods for basic tasks are thus not one time only, but classification and prediction are done repeatedly over time, as new data are made available to the learning machine. As its score function is not linear, a model that can handle nonlinearity must be chosen, e.g., a recurrent convolutional neural network.

Since most of the data can be represented as time series, a finite mixture model [20] is also employed to detect various kinds of latent behavior change in the patients over the course of the ICBT program. In historical data, a hidden patient behavior pattern is assumed to exist. A finite automaton describes the different health or intervention states a patient may be in, observed as well as latent states, and the states can be identified with the help of machine learning methods (e.g., an HMM, see [25]). The transition probabilities between states are in the model computed for each patient. This individualized approach still allows for reasoning about classes of patients, formed, e.g., from mapping behaviors onto class membership. An example class label is *Patients in the depression program under 30 years of age that only* *reply to therapists messages late at night*. Another example is comorbidity: *Patients in the depression program that suffer from irritable bowel syndrome*.

The model is naïve in the sense that it is defined to capture all relevant properties of the data. One way of explaining why it needs to be systemic in order to learn how to learn is to say that it codes for a family of models that all have to be able to grow in complexity and predictive power as the data grow. Therefore, a prior distribution must be developed that can encompass this family, producing a nonparametric model [12]. This Bayesian approach to modeling will employ distributed representations, just like in the networks employed in the model, since these are suitable for overlapping population clusters like the patient classes just mentioned. The clusters themselves are inductively inferred from the mixture model representation, and will form a hierarchy (cf., e.g., [28]). What becomes known about other patients in the same classes can then benefit an individual patient. The model is generative in the sense that each set of individual patient data is a random mixture over latent patient class descriptions, similar to parameterized topic models based on latent Dirichlet allocation [4].

3 Experiments

First steps have been taken in the particular tasks relevant to the learning to learn general problem. To illustrate how the expert discussions around features tie in with exploratory machine learning experiments (i.e., running various machine learning modules), a few simple examples are given in this section. These examples should not be seen as an exhaustive or perfect set of machine learning modules, but are included here to illustrate how the knowledge graph (cf. Fig. 1) can inform which modules to give priority to. Supervised machine learning methods can be used, because labels for success and failure have been appended to the data. A success label is given for either a remitter (post-treatment symptom score below the clinical cutoff for the symptom scale) or a responder (post-treatment score 50% or less of what the pre-treatment score was).

The first experiment concerns the prediction of treatment outcomes from communication patterns (the dialog entity type in Fig. 1, which is detailed in the underlying full knowledge graph still under construction). It is here hypothesized that patterns of communication between patients and therapists carry information relevant for predicting the outcome of treatments. By communication pattern is meant the form of communication rather than its contents. More specifically, the ordering by which messages are sent between a patient (denoted p) and therapist (denoted t) is analyzed. For example, the sequence ptp corresponds to the therapist sending the patient a message (this also includes automatic messages, e.g., concerning homework assignments), the patient responding, followed by another message from the therapist. (Hence, a symbol in the sequence represents the message recipient.) The conjecture is that such a sequence can provide hints about the engagement in, and later outcome of, a treatment. For instance, one may suspect that a ppppp sequence (the therapist sends five messages to a patient without he or she responding) indicates that a treatment will be unsuccessful.

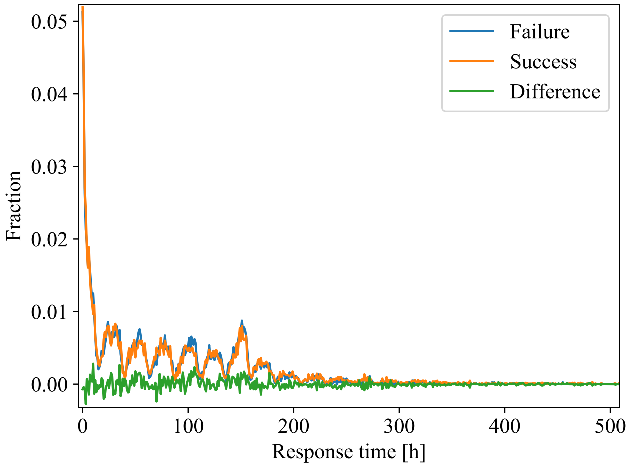

Fig. 2 Densities (i.e., normalized counts) of patient response times for successful and unsuccessful treatments, and the difference between these densities

In an initial test, a multilayer perceptron was trained to predict the outcome of a treatment, given by the labels. Random forest and logistic regression were also tested, with very similar results. The feature inputs to the classifiers were the n-gram counts for a sequence, where an n-gram is a subsequence with n characters (i.e., the four possible 2-gram counts are: pp, pt, tp, tt). Using these features, the perceptron was able to predict outcomes, albeit with a modest accuracy of 60% (diminished to 58% if using data only from the first 3 weeks of messaging). This can be compared to a baseline of 51% when randomly guessing the outcome based on the frequency of failures. Despite the modest accuracy of the perceptron, it may be concluded that the communication sequences at least have a certain predictive power. This can be pursued further, by combining with another weak early signal or by focusing on a subset of messages (e.g., excluding autogenerated messages sent to the patients). An example already tested concerns time: response time to messages and time of day when a message was sent. The latter produces a signal comparable to the n-gram one, while the former is too weak to be useful (Figs. 2, 3). There could be many explanations for this. Not every message prompts a reply, for example.

In a second set of experiments, the patients' texts are analyzed from a *stylometric* perspective in order to investigate the possibility to characterize the various disorders based on the patients' textual behavior. The analysis is stylometric

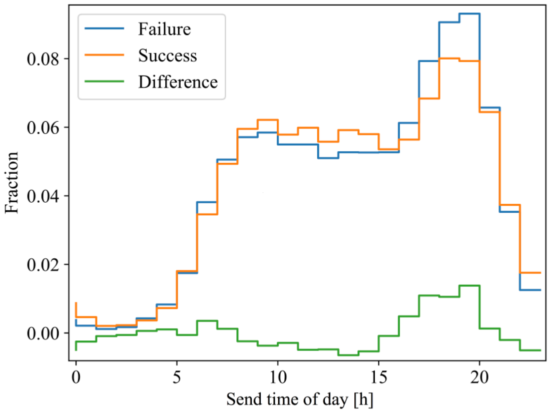

Fig. 3 Densities of the times of day patients sent messages, separated by successful and unsuccessful treatments, and the difference between these densities

because the type of textual features analyzed are primarily stylistic rather than topical. The underlying assumption is that the writing style of the patients at least to some extent reflect their psychological states, and that a stylometric analysis therefore could provide some initial clues to the progress made by the patients. Stylometry is predominantly used in authorship attribution and other forensic linguistic applications, and includes a range of analyses from very basic text statistics such as text length and vocabulary size to more linguistically informed categorization. The latter type of stylometry is normally implemented as a simple lexical matching approach in which terms from a set of predefined lexica are matched in the target texts. Here, the same naïve approach is used to manually construct a set of seed lexica for the following categories:

- *Stress* the patient explicitly writes that she is feeling stress (i.e., terms such as "have time" and "stress").
- *Future* the patient writes about the future (using terms such as "will do").
- *Dedication* the patient expresses dedication to the treatment (using terms such as "dedicated" and "prioritize").
- *Cognition* the patient uses terms referring to cognitive processes (terms such as "comprehend" and "understand").
- *Clicking* the patient explicitly writes about having problems with the Web-based interface (using terms such as "click" and "submit")
- *1st person* the patient uses first person singular pronouns (e.g., "I," "me," and "mine").

Next, the frequency of occurrence of terms in the various lexica in text from the patients from three different treatments: social anxiety, panic disorder, and depression is counted. As can be seen in Fig. 4, there are some (potentially) interesting differences between these three disorders. As

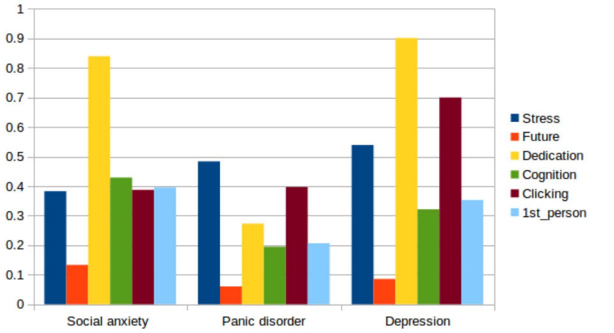

Fig. 4 Average stance values for the patient messages in three different treatments

one example, patients suffering from panic disorder express a much lower degree of dedication, and also less use of first person pronouns and terms referring to cognitive processes compared to patients suffering from social anxiety and depression. Another interesting difference is the fact that patients in the depression treatment express more problems with the Web interface than the other patients.

Figures 5 and 6 show the average message length and the vocabulary richness (computed as the ratio between the number of word types and the number of word tokens in the data) for the three treatments under consideration. Another interesting difference is the that panic disorder had the low-

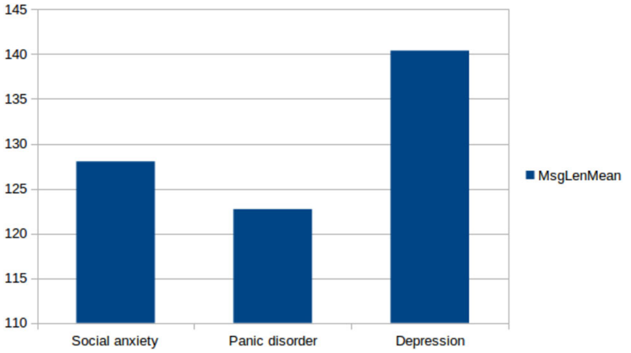

Fig. 5 Average message length for patient messages in three different treatments

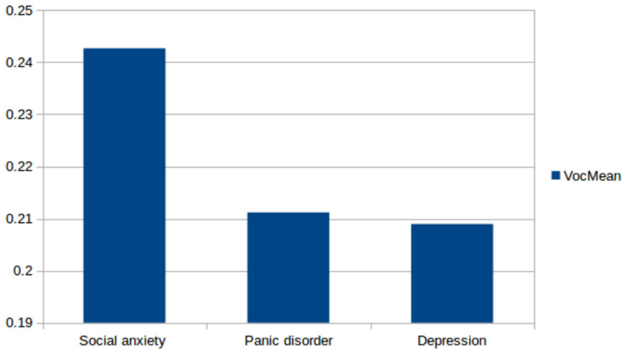

Fig. 6 Type–token ratio over the patient messages in three different treatments

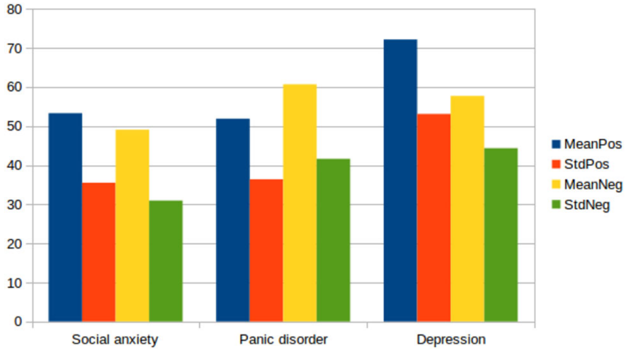

Fig. 7 Sentiment values (average and standard deviation) for patient messages in three different treatments

est average message length. However, from these data, it is unclear as to why (it could be a function of how the treatment encourages verbalization, e.g., in terms of homework). Likewise that individuals in social anxiety have the largest could indicate the amount of verbalization encouraged in some of the homework (e.g., cognitive restructuring).

A simple sentiment analysis using the same type of lexical matching approach as above was also completed. Sentiment analysis attempts to gauge the general attitudinal polarity of text by measuring whether the text predominantly expresses positivity or negativity. Two lexica are used: one consisting of positive terms and one consisting of negative terms. Figure 7 shows the mean and standard deviation for both positivity and negativity over the three treatments. It is interesting to note that depression has the highest presence of positive terms, while panic disorder has the highest occurrence of negative terms.

The above type of treatment-level characterization *may* provide some insights into the differences in textual behavior between patients in the different treatments. As previously stated, because the treatments are different this is likely to be a function of the different treatment modules and there would be confounders. Therefore, it could be more informative and useful to consider each individual patient, and his or her development over time. As an example of how this type of stylometric analysis can be used individually and temporally, four graphs that show the difference between positivity and negativity over time for four different patients are included (Fig. 8). If a message from the patient contains predominantly positive terms, the value for that time step will be positive; if the message contains predominantly negative terms, it will be negative. Based on this simple analysis, a rough idea of how patients express themselves throughout the treatment can be obtained, bearing in mind that the treatment encourages different types of verbal behavior throughout. One may speculate that if the time series ends with a positive derivative, as in the third and fourth example, the patient leaves the treatment with more positive sentiment (probably influenced by the last module encouraging to look back at progress and

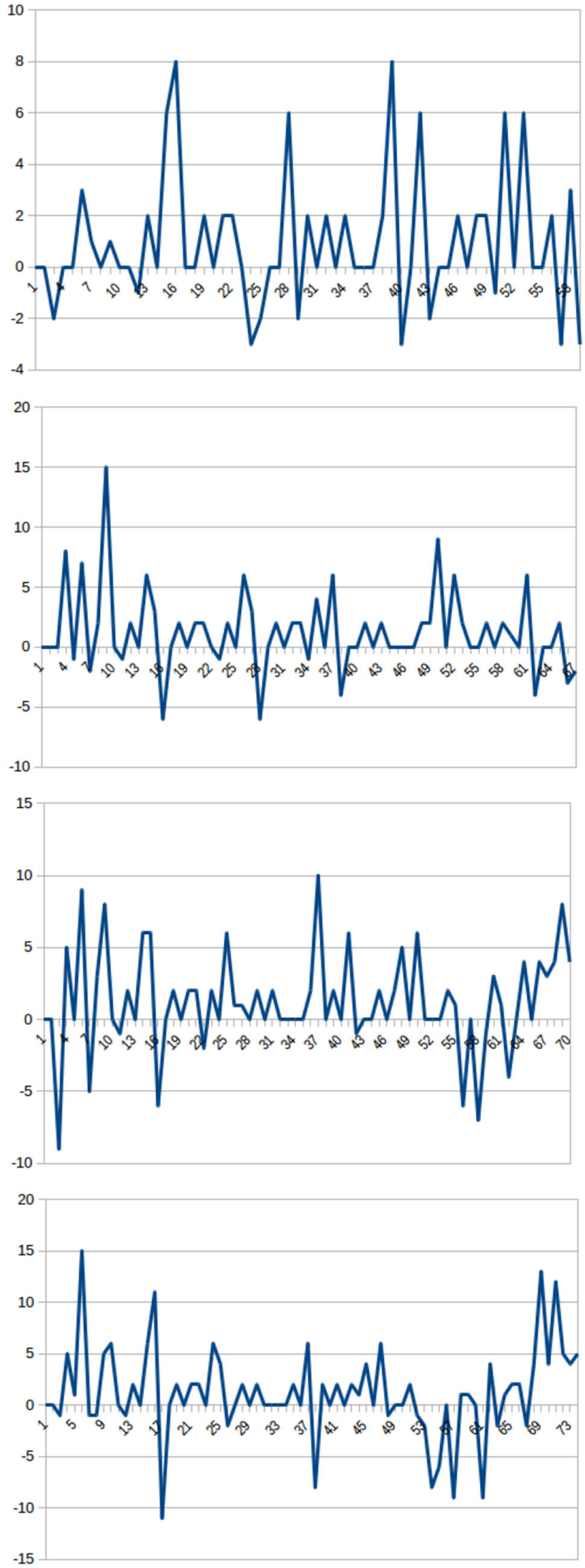

Fig. 8 Positivity/negativity difference for four different patients

reinforce the improvements made during the treatment) than if the derivative is negative toward the end, as in examples one and two.

The type of analysis exemplified above can be used to characterize the textual behavior of patients suffering from the various disorders, which may lead to further insights. Another possible use of this type of stylometric analysis is as a feature extraction step for further applications of machine learning (e.g., prediction or clustering), or as input to the more general learning machine.

4 Discussion

In order to make clear how the experiments briefly described in the previous section contribute to the long-term goal of a learning machine over time becoming an ever more useful tool providing decision support to clinicians, the model in its canonical form must be relativized to the experiments. If one pictures a gating CNN with backpropagation, the purpose of each module built for experimentation with the data becomes a bias module that helps the machine adjust its weights. But with many such modules, it would be foolish to think their bias would contribute independently of each other. Instead, one module will often serve as an amplifier of a signal from a different module. In the example of n-gram message passing analysis, for instance, the sequences of t's and p's have limited predictive power. In the previous section, there were related weak signals, such as the time of day, but it is by no means obvious how to fuse the two signals. For example, plotting distributions for pre- and post-score ratios per times of day messages were sent gives a signal too noisy to be of value (Fig. 9).

But what if n-gram analyses were paired with verbosity analyses? One such module could investigate the length of patient messages in the first 3 weeks, under the assumption that patients likely to churn will either write very long or very short messages. Intuitively, the short messages would indicate too little time devoted to the program, while the long messages would hold explanations for delays and involved reasoning around why assignments could not be completed on time. In this simple example, there would most likely be no scientific results prompting the investigation. Instead, the reasoning behind amplifying the first modules signal would be data driven and could build on a hunch by experts or come from a project data scientist brainstorming session, for example. Its viability is easy to test, since the data are there. The test itself is multi-step in that the new module is validated first, then its tandem performance with the n-gram message passing, and finally, the improvement in performance of the learning machine can be tested. Only after full tests can the decision of introducing the new module be assessed: is its introduction worth the cost of overhead, as measured, e.g., in increased computational complexity and loss of some degree of transparency? There are also methods like feature stacking that enables analyses like n-gram and verbosity in the same model, hence creating a new module [33].

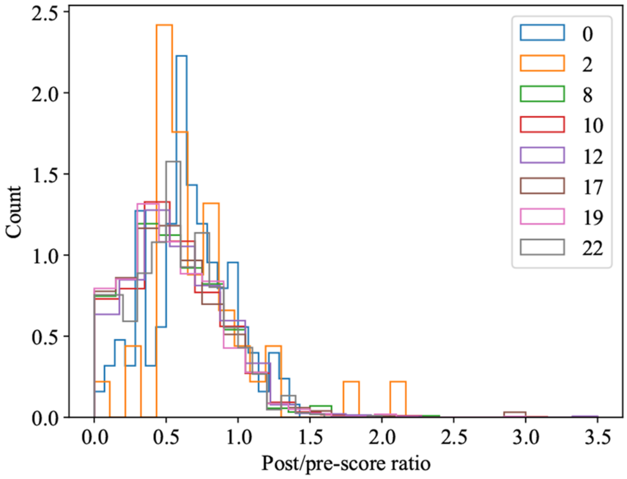

Fig. 9 Histograms of post/pre-score ratios per different send times of day

A choice has to be made concerning the information presented to the therapist once the learning machine has made its prediction and the therapist is supposed to act on the prediction. The primary outcome measures in this field are usually continuous symptom scales, and there is a lot to be gained from using continuous data for predictions. The main clinical outcomes are still success and failure, or the subcategories remitter, responder etc., which in this case are all dichotomizations of the continuous outcome measure for an individual. The question then becomes whether to show the predicted score with, e.g., 95% confidence intervals, or the dichotomized label that is based on the predicted score to the therapist. One might assume that showing the score is better since it carries more information. However, it also leaves more room for bias from the therapist.

Previous research [18] shows that therapists are both poor at making predictions and specifically biased toward optimism (i.e., thinking that their own patients will likely do well). If presented only with a predicted post-treatment score, they might be prone to underestimate the severity of poor but not catastrophic outcomes and not fully act on them. This could be solved by having clear decision rules that apply to the predicted scores, but that would simply mean categorizing the scores into decision categories (0–5 do this, 5–15 do this, etc.). That, however, leaves the therapist with the potential of biased effort if a patient is only just within a decision bracket or almost in a bracket above, again allowing them to make up their own minds perhaps more than desired. Lambert also pointed out that optimism is probably good for therapy, and that it is not necessarily desirable to make therapists cynical even if it is more realistic [18]. To counteract therapists' optimism and biased predictions more fully, it might be desirable to simply label a patient as at risk of being unsuccessful and have a single and clear procedure guiding the subsequent actions of the therapist. These questions will be addressed in a collaborative project where clinicians, clinical researchers, and computer scientists will design the user interface of the decision support tool, along with clinical routines on how to act on different information given by the tool.

Another factor to consider in a learning machine is the learning rate. Since its input signals are heterogeneous and multimodal, the corresponding modules will have varying effect on the learning rate. If, for instance, stochastic gradient descent with momentum is used, then rarely used features will be updated in accordance with their frequency and not according to some homogeneous learning rate schedule. As in the previous example, learning rate should also be subject to multi-step testing in exploratory studies.

5 Conclusion

A conceptual description of a dynamic albeit narrow domain learning machine was given, with some indications of straightforward generalizations to related likewise narrow domains. The machine learns to learn how to provide decision support to therapists in an Internet-delivered psychotherapy program. A number of input signals for machine education were exemplified. The work is ongoing, and the most important future step is to finalize it for employment at the clinic and then critically evaluate its usefulness and general quality in a randomized clinical trial. The development of a learning machine for the clinic is an ambitious undertaking, and the underlying research projects run for 3 years. Implementation and empirical results will be presented in a forthcoming paper. Mathematical aspects of the model will be covered in a series of more formal papers on representability, learnability, and generalizations to even wider handling of uncertainty, such as using intervals for imprecise probabilities and utilities.

References

1. Baxter, J.: Theoretical models of learning to learn. In: Thrun, S., Pratt, L. (eds.) Learning to learn, pp. 71–94. Springer (1998)
2. Bedi, G., Carrillo, F., Cecchi, G.A., Slezak, D.F., Sigman, M., Mota, N.B., Ribeiro, S., Javitt, D.C., Copelli, M., Corcoran, C.M.: Automated analysis of free speech predicts psychosis onset in high-risk youths. NPJ Schizophr. **1**, 15030 (2015)
3. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A neural probabilistic language model. J. Mach. Learn. Res. **3**(Feb), 1137–1155 (2003)
4. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**(Jan), 993–1022 (2003)
5. Boman, M., Bubenko Jr., J.A., Johannesson, P., Wangler, B.: Conceptual Modelling. Prentice-Hall Inc, Upper Saddle River (1997)
6. Boman, M., Sahlgren, M., Görnerup, O., Gillblad, D.: Learning machines. In: AAAI Spring Symposium Series (2018)
7. Bremer, P.: Understanding dynamical systems in high-dimensional parameter spaces. Tech. rep., Lawrence Livermore National Lab.(LLNL), Livermore, CA (United States) (2018)
8. Caruana, R.: Multitask learning. Mach. Learn. **28**(1), 41–75 (1997). https://doi.org/10.1023/A:1007379606734
9. Chekroud, A.M., Zotti, R.J., Shehzad, Z., Gueorguieva, R., Johnson, M.K., Trivedi, M.H., Cannon, T.D., Krystal, J.H., Corlett, P.R.: Cross-trial prediction of treatment outcome in depression: a machine learning approach. Lancet Psychiatry **3**(3), 243–250 (2016)
10. Chen, P.P.: The entity-relationship model–toward a unified view of data. ACM Trans. Database Syst. **1**(1), 9–36 (1976)
11. Forsell, E., Jernelöv, S., Blom, K., Kraepelien, M., Svanborg, C., Andersson, G., Lindefors, N., Kaldo, V.: Proof of concept for an adaptive treatment strategy to prevent failures in internet-delivered CBT: a single-blind randomized clinical trial with insomnia patients. Am. J. Psychiatry **176**(4), 315–323 (2019). https://doi.org/10.1176/appi.ajp.2018.18060699
12. Ghahramani, Z.: Probabilistic machine learning and artificial intelligence. Nature **521**(7553), 452 (2015)
13. Graves, A., Wayne, G., Reynolds, M., Harley, T., Danihelka, I., Grabska-Barwinska, A., Colmenarejo, S.G., Grefenstette, E., Ramalho, T., Agapiou, J., Badia, A.P., Hermann, K.M., Zwols, Y., Ostrovski, G., Cain, A., King, H., Summerfield, C., Blunsom, P., Kavukcuoglu, K., Hassabis, D.: Hybrid computing using a neural network with dynamic external memory. Nature **538**(7626), 471–476 (2016). https://doi.org/10.1038/nature20101
14. Houeland, T.G., Aamodt, A.: A learning system based on lazy metareasoning. Prog. Artif. Intell. **7**(2), 129–146 (2018)
15. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Comput. **3**(1), 79–87 (1991)
16. Kessler, R.C., van Loo, H.M., Wardenaar, K.J., Bossarte, R.M., Brenner, L.A., Cai, T., Ebert, D.D., Hwang, I., Li, J., de Jonge, P., et al.: Testing a machine-learning algorithm to predict the persistence and severity of major depressive disorder from baseline self-reports. Mol. Psychiatry **21**(10), 1366 (2016)
17. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6–11 August 2017, pp. 1885–1894 (2017)
18. Lambert, M.J.: Progress feedback and the oq-system: the past and the future. Psychotherapy **52**(4), 381 (2015)
19. Mitchell, T.M.: Machine Learning, 1st edn. McGraw-Hill Inc, New York (1997)
20. Muthén, B., Shedden, K.: Finite mixture modeling with mixture outcomes using the EM algorithm. Biometrics **55**(2), 463–469 (1999)
21. Naimi, A.I., Balzer, L.B.: Stacked generalization: an introduction to super learning. Eur. J. Epidemiol. **33**(5), 459–464 (2018)
22. Opitz, D., Maclin, R.: Popular ensemble methods: an empirical study. J. Artif. Intell. Res. **11**, 169–198 (1999)
23. Passos, I.C., Mwangi, B., Kapczinski, F.: Big data analytics and machine learning: 2015 and beyond. Lancet Psychiatry **3**(1), 13–15 (2016)
24. Rendell, L.A., Sheshu, R., Tcheng, D.K.: Layered concept-learning and dynamically variable bias management. In: IJCAI, pp. 308–314 (1987)

25. Rothenbuehler, P., Runge, J., Garcin, F., Faltings, B.: Hidden markov models for churn prediction. In: SAI Intelligent Systems Conference (IntelliSys), 2015, pp. 723–730. IEEE (2015)
26. Schibbye, P., Ghaderi, A., Ljótsson, B., Hedman, E., Lindefors, N., Rück, C., Kaldo, V.: Using early change to predict outcome in cognitive behaviour therapy: exploring timeframe, calculation method, and differences of disorder-specific versus general measures. PLoS ONE **9**(6), e100,614 (2014)
27. Skinner, B.F.: The Behavior of Organisms: An Experimental Analysis. BF Skinner Foundation (1938/1990)
28. Teh, Y.W.: A hierarchical bayesian language model based on pitman-yor processes. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics, ACL-44, pp. 985–992. Association for Computational Linguistics, Stroudsburg, PA, USA (2006). https://doi.org/10.3115/1220175.1220299
29. Thrun, S., Pratt, L. (eds.): Learning to Learn. Kluwer Academic Publishers, Norwell (1998)
30. Titov, N., et al.: ICBT in routine care: a descriptive analysis of successful clinics in five countries. Internet Interv. **13**, 108–115 (2018)
31. Turing, A.: Intelligent machinery, a heretical theory. In: Ince, D. (ed.) Collected Works of A. M. Turing Volume 1: Mechanical Intelligence. North Holland, Amsterdam (1948)
32. Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer, New York (1995)
33. Wolpert, D.H.: Stacked generalization. Neural Netw. **5**(2), 241–259 (1992)

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.