Cognitive Systems Research 76 (2022) 78–92

Contents lists available at ScienceDirect

Cognitive Systems Research

journal homepage: www.elsevier.com/locate/cogsys

Human vs. supervised machine learning: Who learns patterns faster?

Niklas Kühl [a,*], Marc Goutier [b], Lucas Baier [a], Clemens Wolff [a], Dominik Martin [a]

[a] Karlsruhe Institute of Technology (KIT) Kaiserstr. 89, 76133 Karlsruhe, Germany
[b] Technical University of Darmstadt (TU Darmstadt) Hochschulstr. 1, 64289 Darmstadt, Germany

ARTICLE INFO

Keywords:
Supervised machine learning
Human learning
Cognitive psychology
Pattern recognition
Small sample size
Experimental study

ABSTRACT

The capabilities of supervised machine learning (SML), especially compared to human abilities, are being discussed in scientific research and in the usage of SML. This study provides an answer to how learning performance differs between humans and machines when there is limited training data. We have designed an experiment in which 44 humans and three different machine learning algorithms identify patterns in labeled training data and have to label instances according to the patterns they find. The results show a high dependency between performance and the underlying patterns of the task. Whereas humans perform relatively similarly across all patterns, machines show large performance differences for the various patterns in our experiment. After seeing 20 instances in the experiment, human performance does not improve anymore, which we relate to theories of cognitive overload. Machines learn slower but can reach the same level or may even outperform humans in 2 of the 4 of used patterns. However, machines need more instances compared to humans for the same results. The performance of machines is comparably lower for the other 2 patterns due to the difficulty of combining input features.

1. Introduction

Supervised machine learning (SML), with its capabilities to support—or even replace—human workers in their daily tasks, is omnipresent in current discussions. While research is investigating the capabilities of SML in a broad range of areas, for example image classification (He, Zhang, Ren, & Sun, 2016) or speech recognition (Hinton et al., 2012), tasks where machine learning models outperform humans are increasing (Grace, Salvatier, Dafoe, Zhang, & Evans, 2018). For instance, in the field of autonomous driving, replacing humans as drivers is supposed to happen sooner or later, although autonomous vehicles are not (yet) able to completely substitute humans in this task (Casner, Hutchins, & Norman, 2016). In the long run, several studies foresee humans being completely substituted by machines in many tasks, including their work processes (Makridakis, 2017; Müller & Bostrom, 2016). But can this development be observed across all tasks? Current examples of (supervised) machine learning models outperforming humans are mainly present in areas where a high amount of training data is available, for example billions of played "Go" games (Chang, Fu, Hu, & Marcus, 2016) or millions of labeled images (Russakovsky et al., 2015).

In real life, however, often only limited "training" data is available (Baier, Jöhren, & Seebacher, 2019)—sometimes just a single instance (Hemmer, Kühl, & Schöffer, 2022; Lee, O'Doherty, & Shimojo, 2015). In this article, we are especially interested in learning patterns by humans and machines in data with only a few instances on a given task. While there is theoretical work in the field of *inductive programming* (Muggleton, 1991; Olsson, 1995), which aims to design techniques capable of capturing patterns with few examples, empirical work in the comparison to human learning is still rare. The research stream of Cognitive Sciences has been investigating learning processes of humans (and recently also machines) (Favela & Martin, 2017), especially with a focus on understanding and mimicking the human brain (Dupoux, 2018; Thagard, 2005). As a subfield, computational cognitive science has been studying the similarities and differences between human and machine learning in the past two decades (Griffiths, Chater, Kemp, Perfors, & Tenenbaum, 2010; Griffiths, Steyvers, & Firl, 2007; Kogler & Pessoa, 2017; Lake, Salakhutdinov, & Tenenbaum, 2015; Lake, Ullman, Tenenbaum, & Gershman, 2017; Lucas, Griffiths, Williams, & Kalish, 2015; Tenenbaum, Kemp, Griffiths, & Goodman, 2011).

However, the investigation of the learning curves, meaning the relation of required training samples and the resulting performance (Perlich, Provost, & Simonoff, 2003) of humans in comparison to SML models, is a topic that has not yet been investigated. This investigation is of major importance, as it needs to be considered whether humans or machines will be performing a task, especially in future, SML-based

* Corresponding author.
E-mail addresses: niklas.kuehl@kit.edu (N. Kühl), goutier@ise.tu-darmstadt.de (M. Goutier), lucas.baier@kit.edu (L. Baier), clemens.wolff@kit.edu (C. Wolff), dominik.martin@kit.edu (D. Martin).

https://doi.org/10.1016/j.cogsys.2022.09.002
Received 6 March 2020; Received in revised form 22 August 2022; Accepted 13 September 2022
Available online 28 September 2022

applications. It gives more insights to the question about which entity learns more efficiently (Hernández-Orallo, 2017b). For instance, in the field of healthcare, data labeling by physicians is extremely costly. From an economic perspective, it might be questionable to use supervised machine learning models in healthcare because the labeling cost can exceed the machine's saving potential (Raghupathi & Raghupathi, 2014). To give first insights into comparing the learning curves of humans and machines with limited training data, we phrase our general research question (GRQ) as follows:

GRQ: How does the learning performance of humans and supervised machine learning models differ with limited training data?

In academia, direct comparisons between humans and supervised machine learning models performing the same task are still rare (Hernández-Orallo, 2017a). Besides the aspect of limited training data, it must be of fundamental interest for researchers to gain a better understanding of which tasks can possibly be undertaken by a supervised machine learning model, as well as the precise conditions that apply (Adler & Schuckers, 2007; Marcus, 2018; Witten, Manzara, & Conklin, 1994). As this work reinforces, there are infinite possibilities of tasks and task characteristics. To provide a starting point for research endeavors in the supervised field, we explore one special scenario: The chosen task for humans as well as machines is identifying patterns with limited training samples (5 to 50 instances). To measure the human performance on this task, we conduct a lab experiment with 44 participants where four different patterns need to be identified. We then apply different supervised machine learning (SML) algorithms on the same patterns and subsequently evaluate and compare the results.

The lab experiment reveals a high dependency between the performance and the used pattern. For humans, the performance is quite similar across the four patterns. They are able to learn the patterns with a smaller number of training instances but their learning curve flattens when the training data exceeds more than 20 training instance. In contrast, SML models show high differences in performance, depending on the underlying pattern. While SML models mirror or outperform the human performance for two patterns, they are struggling to learn the underlying patterns of the other two patterns. Overall, SML models tend to need more training instances to achieve a similar performance compared to humans.

The remainder of this work is structured as follows: In the next Section, we present the necessary fundamentals and related work in the fields of human and machine learning (Section 2). Next, we define the overall task characteristics (Section 3.1) and elaborate on our methodological focus for the experiment design (Section 3.2). In Section 4, we report the isolated results (humans and machines respectively) of the task performance and subsequently introduce the comparison. Finally, we discuss implications (Section 5) and conclude the study (Section 6).

2. Fundamentals and related work

This article assesses and analyzes learning performances of humans and machines. To sketch the foundation for this endeavor, we first give an overview of current research in learning, segmented into learning of humans (Section 2.1), machines (Section 2.2), and research on their comparison (Section 2.3).

2.1. Human learning

Scientific research on human learning started in the second half of the 19th century. In one of the first books about learning, Ebbinghaus (1885) postulated the concept of a learning curve, as the subject group's learning progress flattened over time. Kotovsky and Simon (1973) started analyzing humans learning patterns on a large scale. Human learning is currently divided into three main learning theories: Cognitive psychology, social cognitive theory, and sociocultural theory (Ormrod & Davis, 2004).

For this article's research topic—analyzing human learning with small sample sizes and comparing it to SML—the most relevant research field is cognitive psychology. We leverage phenomena from this area to provide possible explanations for human learning patterns. Cognitive psychology is "the study of how people perceive, learn, remember, and think about information" (Sternberg & Sternberg, 2016, p. 3). The research on cognitive psychology includes studying mental phenomena, such as visual perception, object recognition, attention, memorization, knowledge, speech perception, judgment, and reasoning. To explain such phenomena, cognitive psychology has recourse to neuroscience and its knowledge of brain functioning (Eysenck & Keane, 2015).

In turn, social cognitive theory (Rosenthal & Zimmerman, 1978) includes many ideas from cognitive psychology, but focuses on how humans learn from other human beings through watching and imitating their behavior. The theory suggests that humans can control their own learning. This differs from behaviorism, a now dated theory which led to social cognitive theory and in which learning is solely the result of stimulus–response relationships (Ormrod & Davis, 2004). Learning from others also has the benefit of learning quicker by making fewer mistakes compared to learning from own experiences (Bandura, 1986).

Sociocultural theory stresses the importance of society and culture in learning. Learning a sociocultural tool like a language is not only useful for communication, but also supports humans in their thinking development (Vygotsky, 1964). In contrast with social cognitive theory, humans do not only learn from each other but also work together towards goals that cannot be achieved by individuals. The research focuses on the interaction of children and parents. Children's individual development of capabilities are usually related to interactions with their parents. Additionally, caregivers like parents can broaden a child's problem-solving abilities and stimulate cognitive growth by assisting them to solve more difficult tasks that they would otherwise not be able to accomplish (Vygotsky, 1980).

A few number of studies studied how humans learn from a few instances and have also built computational models to understand human few-shot learning (Lieder, Griffiths, & Goodman, 2012; Vul, Goodman, Griffiths, & Tenenbaum, 2014). However, their aim was to re-engineer the human learning process, while we aim to show empirically how the two entities perform in a direct comparison on limited training data.

2.2. Machine learning

The capabilities of machines have been discussed from various perspectives, including their abilities to capture knowledge (Lieto, Lebiere, & Oltramari, 2018), think (Hoffmann, 2010), feel (O'Regan, 2012; Osuna, Rodríguez, Gutierrez-Garcia, & Castro, 2020; Velik, 2010), be creative (Veale, Gervás, & y Pérez, 2010) and making morally good decisions (Tavani, 2011; Yilmaz, Franco-Watkins, & Kroecker, 2017). The process of how machines obtain their knowledge in the first place is addressed in the area of machine learning. Machine learning describes a set of techniques commonly used to solve a variety of real-world tasks with the help of computer systems that can learn to solve a task instead of being explicitly programmed to do so (Koza, Bennett, Andre, & Keane, 1996). In general, we differentiate between unsupervised, reinforcement, and supervised machine learning (Jordan & Mitchell, 2015).

Unsupervised machine learning comprises methods and algorithms that reveal previously unknown data patterns. Consequently, unsupervised learning tasks do not necessarily have a "correct" solution, because there is no ground truth (Wang, Wang, & Peng, 2009). In the area of reinforcement learning, rewards and punishments allow the model to learn continuously over time with many learning instances. The focus is on a trade-off between an uncharted environment's exploration and the knowledge base's exploitation (Kaelbling, Littman, & Moore, 1996).

In this study, we mainly focus on supervised machine learning, because the most widely used methods are supervised (Jordan & Mitchell, 2015). It therefore seems to be a promising starting point. In respect of supervised machine learning, learning means that a series of examples ("past experience") is used to build knowledge about a given task (Dietterich, 1996). Although statistical methods are used during the learning process, manual adjustment and rule or strategy programming to solve a task are not required. In more detail, (supervised) machine learning techniques aim to build a model by applying an algorithm to a set of known data points to gain insight into an unknown set of data (Hastie, Tibshirani, & Friedman, 2017). Typically, supervised machine learning models rely on large amounts of data to work properly. First techniques, not all directly related to SML, aim to reduce the required amount with different techniques, namely inductive programming (Olsson, 1995; Schmid & Kitzelmann, 2011), genetic programming (Banzhaf, Nordin, Keller, & Francone, 1998), active learning (Settles, 2009), semi-supervised learning (Lin & Cohen, 2010), combinations of both (Rhee, Erdenee, Kyun, Ahmed, & Jin, 2017), external memories (Vinyals, Blundell, Lillicrap, Wierstra, et al., 2016) or one-trial learning (Feng & Sun, 2019).

In terms of a supervised machine learning model's "creation" procedures, the proposed processes vary slightly in their definition of the phases, but generally employ the three main phases: model initiation, performance estimation, and deployment (Hirt, Kühl, & Satzger, 2017). During the model initiation phase, a task is defined, the data is prepared and processed, and a suitable machine learning algorithm is chosen. During performance estimation, various parameter permutations describing the algorithm are validated and a suitable configuration is selected based on its performance when solving a specific task. Lastly, the model is deployed and put into practice to solve a task related to previously unseen data.

2.3. Human vs. machine learning

When it comes to the comparison of human and SML, Hernández-Orallo (2017b) motivates the comparison of natural and artificial intelligence in the first place. The field of Neuroscience (Florez, 2015; Hutto & Kirchhoff, 2015; Rajalingham et al., 2018) aims to understand the learning of humans and the facilitation with machines on a theoretical level. The precise capturing of the related learning curves have been analyzed theoretically and empirically for humans and machine learning techniques separately in different domains, e.g. creativity tests (Olteţeanu, Falomir, & Freksa, 2016), face recognition (Adler & Schuckers, 2007), music prediction (Witten et al., 1994) or cognitive research (Marcus, 2018). In the field of computer vision, multiple comparisons of human and machines have been made (Eckstein, Koehler, Welbourne, & Akbas, 2017; Elsayed et al., 2018; Peterson, Abbott, & Griffiths, 2018; Zhou & Firestone, 2019).

Apart from these specific domains and closer related to our study is the idea to build computer models capable of solving IQ tests (Hernández-Orallo, Martínez-Plumed, Schmid, Siebers, & Dowe, 2016). While not using supervised machine learning, Insa-Cabrera, Dowe, Espana-Cubillo, Hernández-Lloreda, and Hernández-Orallo (2011) aim to compare reinforcement and human learning, however, they only regard small sample size of observations.

Human learning can be compared to machine learning based on various aspects. Dubey, Agrawal, Pathak, Griffiths, and Efros (2018) focus on human priors for playing video games. In their experiment, they use an unknown video game that a human solves quite easily by using its priors on semantics, gravity, and objects. By reversing semantics and masking affordances, the human performance decreases drastically. The machine performance, represented by reinforcement learning algorithms, performs significantly better under the same conditions. Humans' prior knowledge is important when it comes to solving new problems quickly.

Kim, Reif, Wattenberg, and Bengio (2019) does research on psychophysical phenomena—which can be found in human learning—in trained machine learning models. Gestalt phenomena are a part of human visual perception in which humans realize that the whole differs from the sum of its parts (Köhler, 1967). They show that some neural networks are able to show one type of Gestalt phenomena under the proper circumstances.

In hybrid intelligence (Dellermann, Ebel, Söllner, & Leimeister, 2019), humans' complementary strengths, like flexibility and common sense, are combined with those of machines, for example consistency and speed. This sociotechnological ensemble can overcome the current limitations humans and machines have. Another way to combine human and machine abilities is to treat machines as teammates (Burr, Cristianini, & Ladyman, 2018; Seeber et al., 2019; Smart, 2018). This could increase work speed and lead to better decision-making by detecting negative cognitive biases.

In conclusion, a direct comparison of human learning and supervised machine learning for the same task with limited training data availability still remains a research gap and is addressed in this work.

3. Methodology

With the related work at hand, we outline the methodology used in this article. We first set necessary prerequisites (Section 3.1) and then elaborate on the experiment's design (Section 3.2).

3.1. Prerequisites

Before discussing the experiment's design of comparing human and machine learning, we need to set the prerequisites for the task being solved in this experiment. As we require a controllable task with precise benchmarks for performance evaluation, a suitable candidate is supervised machine learning, which we utilize for this article. When it comes to choosing a meaningful task in the area of SML, there are many possible characteristics to describe it. To deduce the possibilities and reason our selection, we look at corresponding task characteristics (Section 3.1.1) and subsequently outline our implementation of the chosen task (Section 3.1.2).

3.1.1. Task characteristics

A learning curve depicts task performance based on experience. In our case, experience is measured by the amount of training data, more precisely by the number of training instances. Task performance is influenced by two main factors: the characteristics of the entity performing the task (humans or machines) and those of the task itself. Depicting a general learning curve for every type of task characteristic exceeds the scope of this article, and we have to limit the scope to an interesting selection of all possible tasks. For our supervised machine learning task, four task characteristics are of importance: input, output, instances, and features.

Input. The input describes the data the task is based on. It can differ by data type (e.g., numeric or binary) and by data representation (e.g., table, picture, or audio).

Output. A task also differs in the demanded output. Two types of output are relevant in this case: classification and regression. A classification determines whether each instance belongs to one of the predetermined classes, whereas the result of a regression is a continuous number.

Instances. The number of instances that are available for the learning process.

Table 1
Overview of the task characteristics of interest and their implementation in this work.

Task characteristic		Attributes	This work
Input	Data type	e.g. numeric data, binary data	Binary data
	Data representation	e.g. table, picture, audio	Picture (humans), table (machines)
Output		Classification, regression	Binary classification
Instances		Number of instances	5 to 50
Features		Number of features	9 features

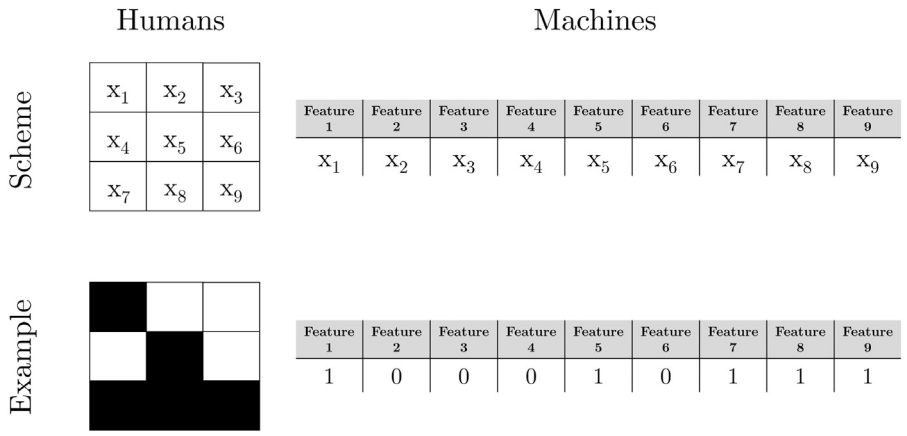


Fig. 1. Schematic representation for humans and machines of an instance with features x_1 to x_9.

Features. The instances of a task are described by a fixed number of distinct features.

To start the research endeavor, we select a task with a binary input, a binary classification as output, a small set of training instances, and a limited number of features. An overview of all task characteristics of interest and their implementation in this work can be found in Table 1. To conclude, we update the general research question to our research question (RQ):

> *RQ: How and when do learning curves differ between humans and supervised machine learning models for small sample sizes, using a binary classification with limited binary features?*

Additionally, we define the following requirements for our task: it should not require prior knowledge and should use a balanced data set (same number of true and false instances) and should be solvable in a reasonable timeframe. The task should be represented in a suitable way for humans and machines, and it should be possible to depict the results in a learning curve.

3.1.2. Implementation of task characteristics

As a last prerequisite, we have to agree on a task that satisfies our set of task characteristics and also complies with the additional requirements defined in Section 3.1.1.

We use two suggestions in the field of intelligence tests as a foundation for our task, namely minimum intelligent signal tests (MISTs) and Raven's progressive matrices (RPMs): MISTs are binary questions that are used to quantify humanness (Łupkowski & Jurowska, 2019; McKinstry, 1997). Compared to other intelligence tests, these questions do not require a complex answer, but only a simple yes or no, which satisfies our limitation on a binary output. However, the input is natural speech and not a set of a few, binary features. RPM (Raven, 2000) is a test of visual geometric objects, designed by a rule. The task is to complete the set of visual geometric objects by selecting an object out of six or eight options. Only one of the selectable objects matches the rule. RPMs have a graphical representation that can be reduced to a set of instances with a few binary features to get standardized instances. However, they lack a binary output.

By combining these two tests, we define the following task: To have the same number of features, we use only 3×3 matrices with 9 elements (= 9 features). Every feature is binary. Accordingly, we have a set of $2^9 = 512$ different matrices. These matrices can be displayed as a picture with elements of black and white (for humans) or as a list of numbers with features of 1 and 0 (for machines). Fig. 1 shows an example of how the same instance is represented for humans and machines respectively. Based on a rule regarding the feature value, we can classify the matrices: Some instances (matrices) fulfill the rule, therefore they are labeled as true, whereas all the other instances do not fulfill the rule and are labeled as false. We define four basic patterns as the four rules that define our classification task (see Fig. 2).

Diagonal. Matrices that fulfill the diagonal rule have at least one diagonal line that is labeled black, either starting in the upper left block and continuing to the lower right block, or stating in the lower left block and ending in the upper right block.

Horizontal. Matrices that fulfill the horizontal rule have at least one horizontal row of only black elements.

Numbers. The numbers rule is satisfied if five elements in total are labeled black.

Symmetry. Symmetry describes axis symmetry, either to the middle column or the middle row of the matrix (see Fig. 2).

3.2. Experiment design

The previously described task has to be adjusted to be conducted by humans (Section 3.2.1) as well as machines (Section 3.2.2) in an experimental setting. In the end, the results should render a learning curve. To generate a learning curve for a specific rule, we define a game with multiple rounds. During the game, the rule does not change. At the start, the player receives access to five labeled instances (training data). We ensure the probability of each instance to be labeled positive is 50% (and accordingly 50% to be labeled negative) to account for imbalances of positive and negative labeled instances in the data set based on the selected rule. Additionally, the player receives five unlabeled instances (test data) that have to be labeled based on the knowledge derived from

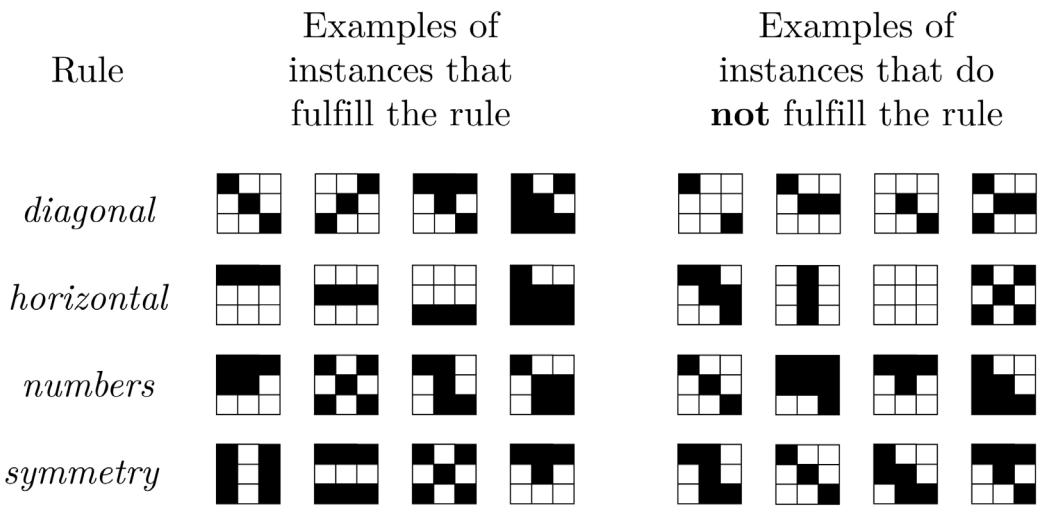


Fig. 2. Exemplary instances for each rule.

the labeled training instances. The probability for each instance to be labeled positive remains 50% as explained before. We then measure the performance on the test data with the *accuracy* metric, which is defined as the number of correctly labeled instances divided by the total number of labeled instances. It takes into account the true positives (TP), true negatives (TN) as well as false positives (FP) and false negatives (FN) taken from the confusion matrix, a table that visualizes the performance of an supervised machine learning algorithm (Stehman, 1997) and is shown in more detail in Table 5 in Appendix. While false positives refer to instances that were predicted to be of the positive class but in reality are negative, the reverse is true for false negatives. Both, false positives and false negatives, have the same effect on reducing the accuracy metric.

$$accuracy = \frac{TP+TN}{TP+TN+FP+FN} = \frac{\sum correctly\ labeled\ instances}{\sum all\ labeled\ instances}$$

As labeling is only a binary decision in our work, an accuracy indicator of "1" is a 100% correct labeling, whereas an accuracy indicator of "0.5" is equivalent to a random guess where labels are randomly assigned. The accuracy of the labeling of the five instances represents the performance in the first round.

Instance. An instance consists of nine elements and a binary label that indicates if the instance fulfills the rule or not.

Round. In every round, humans and machines get five (additional) labeled instances and five new instances to label.

Game. A game has either 10 (humans) or 20 (machines) rounds.

Experiment. An experiment is finished when four games with four different rules are played.

In the second round, the previously labeled instances disappear and five new, unlabeled instances are displayed (new test instances). Five additional labeled instances are shown, leading to a total of 10 labeled instances available for training. The labeling of the five new unlabeled instances in the second round determines the performance in Round 2. Evidently, additional rounds follow the same pattern. This is depicted in Fig. 3. The order of labeled and unlabeled instances is randomized in every game. However, one matrix (instance) will only be part of either the training or testing data, not both. The learning curve is generated based on the performances in each round.

3.2.1. Experiment with humans

The experiment with humans is conducted by studying participants in different sessions. They participate in the experiment individually and without any prior knowledge. In advance, they get a standardized introduction about the general aim of the experiment, the layout of the user interface, and some abstract examples. This introduction is available before and during the experiment in printed form, and they can use scrap paper and a pencil to make notes. The complete instruction is shown in the Appendix.

Every participant has the possibility to play all four rules, leading to four games in total. The total number of rounds per game is limited to 10, which means that the participants will see 50 labeled instances in total and have the opportunity to label 50 instances during one game. After finishing one game, a participant does not receive any feedback about his/her performance. This ensures independent games, as a participant is not influenced regarding the following games. The order of rules is randomized for each participant. Fig. 4 shows an example of GUI of the experiment with humans for the rule *symmetry*.

The experiment is organized and recruited with the software hroot (Bock, Baetge, & Nicklisch, 2014). In total, 44 people participate in two sessions, with 19 people in the first session and 25 in the second one. There are 20 female and 24 male participants, with an average age of 26 years (SD = 9.6). Most experimentees (91%) are currently enrolled at a German university, majoring in 17 different disciplines, mainly Industrial Engineering and Management (10 students) and Computer Science (seven students).

Before each session, the experimentees are given instructions on how the experiment works and about their tasks. These instructions are also available on the screen before every game and in printed form during a game. Every person conducts the experiment individually in soundproof cubicles, using a computer. The time limit to complete the experiment with all four rules is set to one hour. The participants are incentivized by an individual payment (Kvaløy, Nieken, & Schöttner, 2015) which is based on their performance relative to that of all other participants in the same session and which ranges from €16 (best performance) to €7 (worst performance).

3.2.2. Experiment with machines

The experiment with machines is conducted by three different algorithms out of all supervised machine learning algorithms, namely a

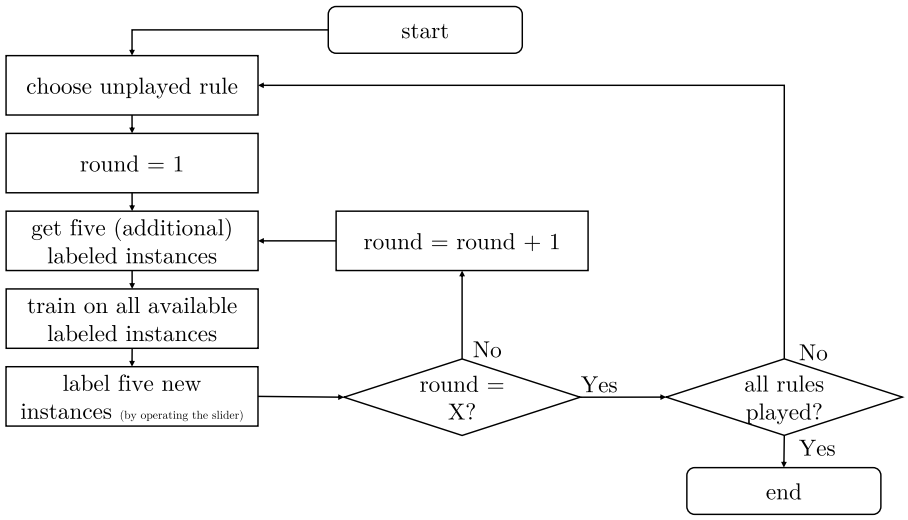


Fig. 3. Procedure of the experiment with X = 10 rounds for humans and X = 20 rounds for machines.

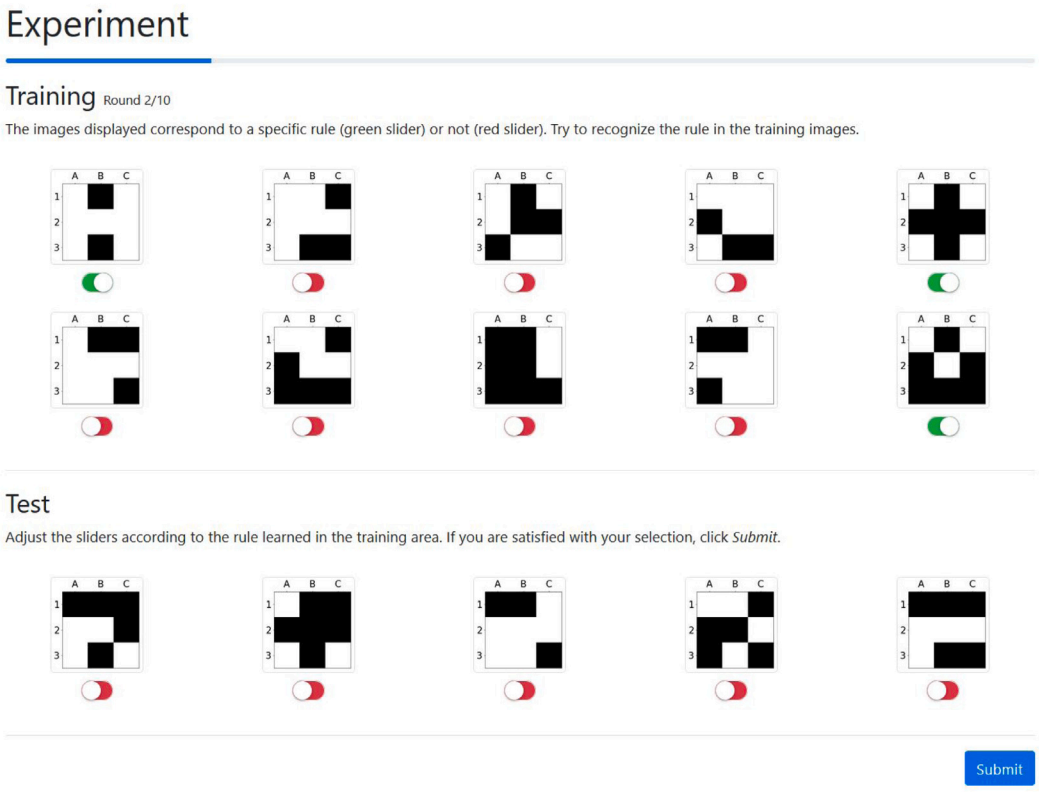


Fig. 4. Screenshot of the experiment with humans in Round 2 for the rule *symmetry*. On the top, 10 training instances are displayed (5 training instances which are still visible from the first round, and an additional 5 instances which were added in the second round). These sliders are not changeable by the participants but depict the labeled "ground truth" of the training. On the bottom part, five unlabeled (test) instances are depicted, which can be changed by the participants. By default, they are all marked as "false".

logistic regression (linear), a decision tree (propositional), and a neural network algorithm (non-linear). We do not perform an excessive parameter tuning for the algorithms, because this requires a large number of instances (which is a limitation of our task). This is consistent with the human experiments, because the study participants have no option of playing the game in advance to gain additional knowledge that facilitates completing the task. However, we discuss the influence of parameter tuning in Section 5.2 with distinct experiments.

To increase comparability, every algorithm is applied to every game with the same number of resulting models as the number of humans who played the game. Since we are not limited concerning resources like time, money, or room availability as in the experiment with humans, we can double the number of rounds to 20, which leads to 100 labeled instances (test instances). While our main focus is on the comparison of the first 10 rounds between human and machine, we are curious about how a machine's performance would develop with

Table 2
ANOVA results for session, rules, instances, and games. Significance is depicted with three different levels of statistical significance (* $\widehat{=}$ $p < 0.05$, ** $\widehat{=}$ $p < 0.01$, *** $\widehat{=}$ $p < 0.001$), represented with one, two, or three stars.

	P-value	Significance
Sessions	0.1769	Not significant
Rules	1.0e−13	***
Instances	6.6e−29	***
Games	0.0002	***

additional samples. The algorithm is instantiated for one game only and gets terminated after every game so that knowledge from previous games is not used.

For implementation, we use the programming language Python (Version 3.9) and the packages *scikit-learn* (Pedregosa et al., 2011) (Version 0.24), *tensorflow* (Abadi et al., 2016) (Version 2.8) and *keras* (Chollet, 2015) (Version 2.8). The code is available at https://github.com/nkukit/humanvssml.

4. Results

After presenting the experiment in the last chapter, we conduct the experiment with humans and the different machine learning algorithms. In this chapter, we evaluate the experiment conducted by humans (Section 4.1) and by the machine learning models (Section 4.2) in detail. In a follow-up step, we compare the human results to those of the machine learning models (Section 4.3).

4.1. Experiment with humans

The experiment was conducted in two sessions, as described in Section 3.2.1. Of the participants, 91% (40 out of 44) finished all four rules within the time limit of one hour. The order of the games is random. We therefore have 42 datasets for the rule *diagonal* and *horizontal* rule and 43 for the rule *numbers* and *symmetry*.

We use analysis of variances (ANOVA) (Girden, 1992) to analyze the dataset in a first step (Table 2). A one-way ANOVA with the two sessions shows no significance in performance, which indicates that there is no statistical significance between the means of the two sessions. We can therefore analyze all sessions together. To determine the influence of the rules and the number of training instances, we use two-way ANOVAs with replication, since we have a set of paired data where one individual has played several games as well as 10 rounds (= 10 data points) within a game. Since two-way ANOVAs with replication require an equal set of paired data per individual, we exclude the four participants who did not play all four games. Rules and instances show a high statistical significance in performance, as expected in our research question. We account for it later by looking at each rule independently and using learning curves that display the performance for each number of training instances separately without aggregation.

The performance of the order of games is also statistically significant. This could indicate something like a learning effect between the games, but analyzing this finding is beyond the scope of this article and can be investigated in future research.

4.2. Experiment with supervised machine learning models

Corresponding to the number of human experimentees who played one rule, we respectively use 42 or 43 individual machine learning model instantiations/runs for each of our three types of machine learning algorithms—regression, decision trees, and neural networks—to play a game for a certain rule. The games are played in the same way that the humans conduct the experiment, seeing five labeled training instances in the first round, and the performance is determined by labeling five instances. The split between training and test instances as well as the order within each group is randomized for each instantiation, similar to the experiment conducted by humans. During each game, only the random seed changes per round. As regression algorithm, we choose a logistic regression (Pedregosa et al., 2011) with an L-BFGS solver (Liu & Nocedal, 1989). The used decision tree algorithm is a *DecicionTree Classifier* (Breiman, Friedman, Olshen, & Stone, 1984; Pedregosa et al., 2011) and a *Multilayer Perceptron* (MLP) (Glorot & Bengio, 2010; Pedregosa et al., 2011) with an L-BFGS solver as our neural network algorithm of choice.[1]

In the following section, we will use the terms of MLP, decision tree, and logistic regression, knowing that we will always compare the aggregate of 42 or 43 individual performances of these machine learning algorithms.

We use *scikit-learn*'s default parameters of all algorithms (Pedregosa et al., 2011). In-line with Sun and Pfahringer (2013), we suspect a possible hyper-parameter optimization will only result in marginal positive effects or introduces bias and low fidelity on small datasets like the one at hand (Janitza & Hornung, 2018; Yu & Zhu, 2020). However, as it would be of interest to see how optimized models perform, we also tested a selection of models in an optimized version within the Discussion. The default parameters used within our experiments are documented in the Appendix.

4.3. Comparison

To compare the results of the three machine learning algorithms with the human performance, we analyze each rule through the learning curves our experiment generated. Fig. 5 depicts the results for the rule *diagonal*. The number of training instances is displayed on the x-axis. The left y-axis belongs to the line charts and shows the average accuracy of all experimentees with the given number of training instances.[2]

The right y-axis belongs to the bar chart. The bar chart shows three different levels of statistical significance (* $\widehat{=}$ $p < 0.05$, ** $\widehat{=}$ $p < 0.01$, *** $\widehat{=}$ $p < 0.001$) between the performance of the machine learning models and that of humans. From 55 training instances onward, there is no corresponding human data and the significance refers to the performance difference between machine learning models and humans with 50 training instances. The statistical significance in performance difference is calculated by a two-sided t-test for unequal variance (Yuen, 1974). As this results in multiple t-tests on the same dataset, we control the false discovery rate (FDR) by the Benjamini–Hochberg procedure (Benjamini & Hochberg, 1995).

Regarding the rule *diagonal* (Fig. 5), the decision tree outperforms all other machine learning models and human participants. In one case, the performance of the decision tree is not significantly better compared to the human within the first 50 training instances. Beginning with 55 training samples, the decision tree performs significantly better ($p < 0.01$) than humans in 50 instances. In contrast, the MLP and the logistic regression show similar accuracy compared to the human and do not improve significantly in later rounds. Therefore, those machine learning models do not outperform humans in 50 training instances significantly.

Fig. 6 displays the results for rule *horizontal*. In contrast to the previous chart, humans significantly outperform the machine learning models in the first 50 training instances. However, the statistical significance in performance decreases with more training instances the

[1] To also include a more modern neural network architecture, we ran experiments with a Convolutional Neural Network (Krizhevsky, Sutskever, & Hinton, 2012) as well, the results of which we present in Table 11 in the Appendix.

[2] Using the accuracy metric allows us to correctly account for false negatives and false positives as mentioned in Section 3.2. Some exemplary confusion matrices are shown in more detail in Tables 6 and 7 in Appendix.

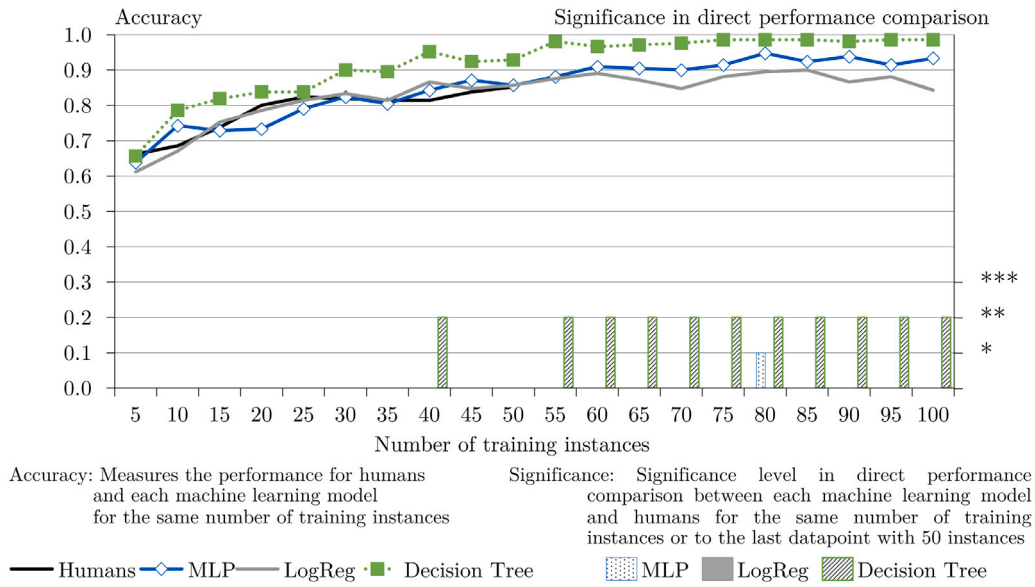


Fig. 5. Performance and significances in performance for the rule *diagonal*.

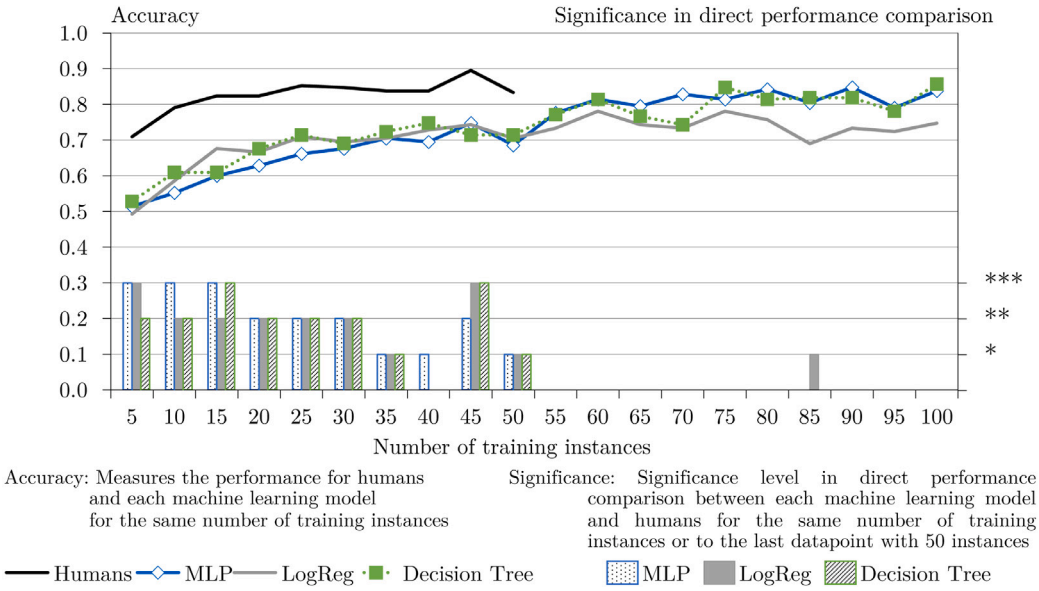


Fig. 6. Performance and significances in performance for the rule *horizontal*.

machine learning models have to learn. Beginning with 55 training instances, the performance of humans with 50 instances and machines with 55 instances does not differ significantly anymore. With the machine learning models, the accuracy is on an equal level and only in the end it seems that the performance of the logistic regression deteriorates a bit.

The rule *numbers* in Fig. 7 shows the highest accuracy of human performance across all four rules. Starting with 15 training instances, the performance is always on or above 90%. The accuracy of the three machine learning models shows no improvement and the accuracy remains at around "0.5" for the entire 100 training instances. The performance difference between humans and machine learning models is therefore significant ($p < 0.001$) for the experiment across all rounds.

The results for rule *symmetry* are depicted in Fig. 8. Similar to the human performance in rule *numbers*, humans outperform the machine learning models. With five training instances, the performance is significantly better (MLP: $p < 0.01$, decision tree and regression: $p < 0.05$). Afterwards, the human performance improves more than the machine performance and the differences become highly significant ($p < 0.001$). However, the human performance reaches its accuracy maximum right below 0.9 after 20 instances and remains on this level, whereas the accuracy of MLP and decision tree slightly improves from round to round. After 50 training instances, the significance level decreases ($p < 0.01$).

5. Discussion

The results of the experiments provide ground for possible interpretations. We discuss possible explanations of the observed values. This is by no means a full explanation of the shape of every learning curve; however, we give insights on how different theories in computer science and in cognitive psychology can explain some of the outlined results. We start with explanations of the human performance (Section 5.1) and end with the machine experiment (Section 5.2).

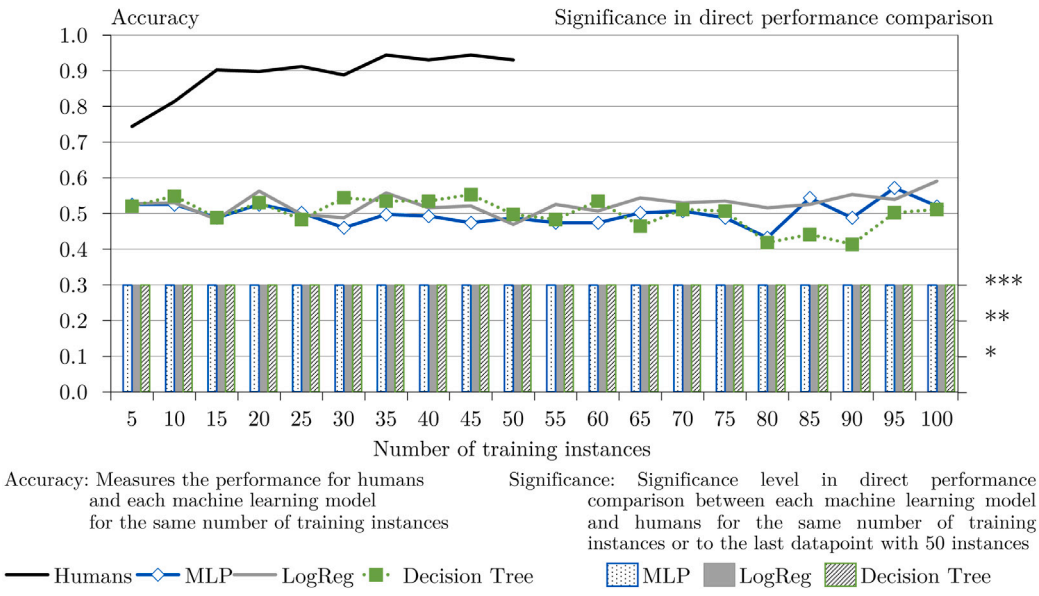


Fig. 7. Performance and significances in performance for the rule *numbers*.

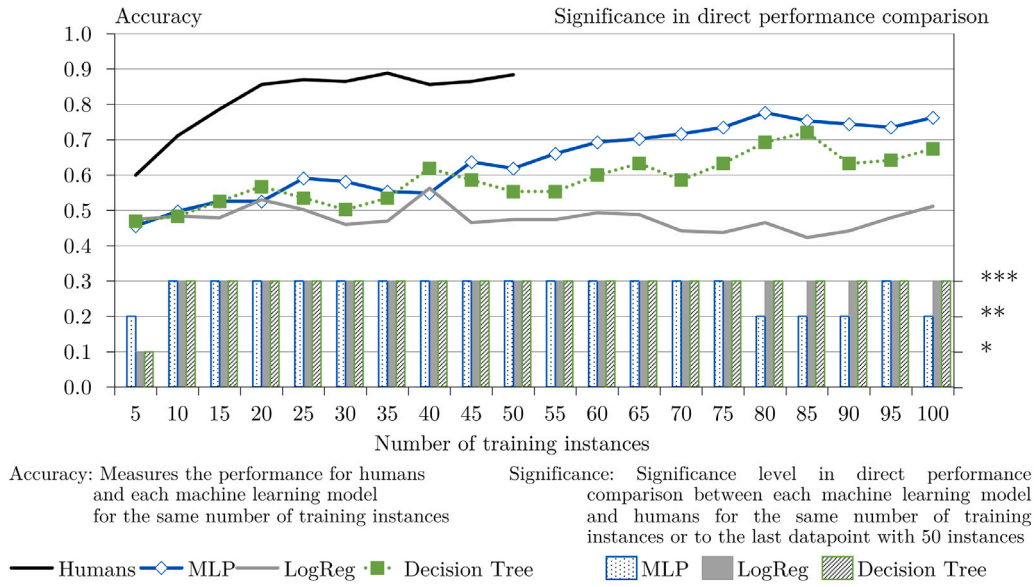


Fig. 8. Performance and significances in performance for the rule *symmetry*.

5.1. Experiment with humans

Given our experiment setup and its results, we employ theories from the area of cognitive psychology, as introduced in Section 2.1, to interpret our results. Other areas of human learning—social cognitive theory and sociocultural theory—are less applicable, as the experiment is performed individually.

The human performance shows two key characteristics across all four rules: High accuracy when labeling the first five instances (no accuracy below 60%, which outperforms the supervised machine learning models in three of the four rules) and only small performance improvements after learning with 20 or more training instances.

An explanation for the first observation is grounded in the concept of *one-shot learning* (Lee et al., 2015). Besides incremental learning, where humans learn step-by-step through trial and error (Thorndike, 1913), a human is also capable of one-shot learning, which is a technique to learn from a single instance. When a child touches a hot stove plate, he/she will immediately learn not to do it again. This single training instance illustrates one-shot learning. With object recognition, one-shot learning enables humans to recognize objects after one instance by relating the newly seen object to prior knowledge (Fei-Fei, Fergus, & Perona, 2006). Although the used patterns in the experiment are highly abstract, the human can still connect them to known shapes.

The second finding can be explained by cognitive load theory (CLT) (Fan, Chen, & Yen, 2010; Shaffer, 2017; Sweller, Van Merrienboer, & Paas, 1998). CLT describes the learning process as the combination of three loads: the intrinsic cognitive load coming from the difficulty and complexity of the learning subject; the extraneous

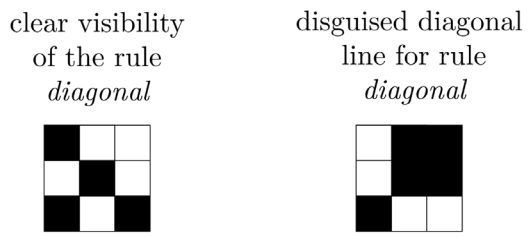


Fig. 9. Two instances that fulfill the rule *diagonal*—one with clear visibility of the diagonal line and the other one with two seemingly separated squares of elements disguising the diagonal line.

cognitive load, which originates in preparation of the learning subject; and the germane cognitive load, which describes humans' needed learning capacity to understand the learning subject (Paas, Renkl, & Sweller, 2003). The working memory, where the cognitive process takes place, is essential for the learning process. The working memory itself is limited (Ayres & Paas, 2009; Van Merrienboer & Sweller, 2005), which everyone can experience when playing the board game "Memory" and then being unable to memorize every card that has been revealed. Our interpretation of the data suggests that 20 training instances are the limit for humans' working memory, with more training instances only leading to cognitive overload (Moreno & Mayer, 1999) and not to improved performance. An additional explanation can be found in the fatigue effect. The longer the human plays the same rule, the more his performance is effected negatively by fatigue (Gonzalez, Best, Healy, Kole, & Bourne, 2011) and counteracts the positive effect of enlarging the data basis by seeing more instances.

Regarding the human performance per rule, the learning curve for the rule *diagonal* is unique. The accuracy in the first round is the second lowest, the maximum performance is the lowest of all four rules, and the machine learning models have similar accuracy numbers or even outperform humans. All these findings indicate that the human performance is particularly bad compared to the other rules. When looking at instances fulfilling the rule *diagonal*, as shown in Fig. 9, one can see that the elements (which form the diagonal line) are not joined on the sides but are only linked via their corners. Similar to an optical illusion (Coren, Ward, Porac, & Fraser, 1978), like the well-known rabbit–duck illustration in which some people see a rabbit and others a duck, the diagonal lines can "disappear" in some instances while seeing other possible rules. Therefore it becomes harder for humans to see the diagonal line as a potential rule.

Humans show the best performance for the rule *numbers*, which implies that this rule benefits human learning the most. Cognitive fit theory (CFT) (Vessey, 1991) indicates a link between the task and the chosen type of presentation that leads to superior task performance—the finding of the maximum value in a numerical dataset is completed quicker by humans when plotting the data as a graph compared to a data table. In our experiment, the matrix representation of the data favors counting the number of true elements as well as comparing the counts between training instances.

Despite the mentioned theories explaining many findings from the results in our experiments with humans, we have to keep the statistical significance in performance between the different games in mind. In future research, we must analyze the results of every game individually and we may find further theories that explain other aspects of human performance.

5.2. Experiment with machines

Within the experiments performed with machines, we discuss two areas. First, we analyze the influence of hyperparameter tuning on the results. Second, we interpret the experimental results and reline them with insights from existing literature.

Table 3
Search space of the MLP classifier for the rule *symmetry*.

Parameter	Value range
Amount of hidden layers	{1;2;3}
Neurons per layer	100
Solver	{lbfgs;adam}
Activation	{relu;tanh}
Learning rate type	{constant;adaptive}
Initial learning rate	{0.001;0.01;0.1}

5.2.1. Hyperparameter tuning

We conducted our experiment with machines initially with the default parameters of the respective SML algorithms. However, we believe it is of increased interest to analyze the relationship between the parameters chosen and the resulting performance, i.e., does parameter tuning make the model significantly better, even though we only work with few data instances.

To address this aspect, we proceed as follows: We first split our available data (per rule) into two equal subsets: one for optimization of the hyperparameters and one for the conduction of the experiment with the optimized parameters. We require this split to not overfit the model during optimization (Cawley & Talbot, 2010). Within the optimization step we conduct a Grid Search within a 15-fold cross-validation, which allows us to test different permutations and still receive reliable results. The best-performing parameters from this optimization procedure are then used to train a new classifier, which is then applied to the second subset of data for performance evaluation. As an exemplary candidate, we choose the MLP classifier for optimization with the search spaced shown in Table 3. The steps of dividing, optimizing and experimenting is repeated 42 times to match the previous 42 runs and 42 human participants per rule. The results are then averaged.

As the results in Table 4 show, we see similar but marginal effects for all rules *diagonal*, *horizontal* and *symmetry*. Within the 50 training instances, we see no clear trend of either improving or decreasing performance, with alternating signs in front of the figures. Overall, the effect is marginal with an average improvement of up to 3% points. We have no indication that the optimized MLP classifiers reach a considerable higher level of abstraction, i.e., capturing the pattern of the task in a better way.

Although the effects are only marginal in most cases, it is interesting that we achieved better performances of our models in comparison to the default instantiations (MLP_{opt} vs. MLP_{def}) for all 4 rules, in particular for larger number of training instances. These performances can be seen as an upperbound of the MLP in our case, which is interesting to further investigate. For the rule *diagonal*, $MLP_{opt} = 92.86$, while $MLP_{def} = 85.71$, both exceeding the human performance—which it also did prior to optimization (and was still exceeded by the decision tree classifier). In case of the *horizontal* rule, $MLP_{opt} = 76.67$, whereas the default model was only able to reach $MLP_{def} = 68.57$. However, the optimized model is still not able to achieve as good performance as the humans. The rule *numbers* was the most challenging one for the machine learning models—optimization did not help and the classifier is still guessing ($MLP_{opt} = 54.88$ in comparison to $MLP_{def} = 48.84$). Humans perform significantly better. Finally, we see a slight improvement in the maximum score for the rule *symmetry* with $MLP_{opt} = 65.12$, while $MLP_{def} = 61.86$. However, even the optimized model is far from the human performance.

5.2.2. Experimental interpretations

There are two general findings regarding the machine learning models across all four models: The performance after five training instances is similar or lower compared to the human performance and

Table 4
Comparison (in percentage points = %p) between the optimized MLP classifier (MLP_{opt}) and the default MLP classifier (MLP_{def}) for different rules. Comparison/improvement is calculated by $MLP_{opt}[\%] - MLP_{def}[\%]$.

Rule	Number of training instances										Average improvement
	5	10	15	20	25	30	35	40	45	50	
Diagonal	+2%p	−9%p	+2%p	+5%p	+1%p	−3%p	+5%p	+8%p	+2%p	+7%p	**+2%p**
Horizontal	0	+1%p	0	+5%p	0	−3%p	+9%p	+9%p	+6%p	+8%p	**+3%p**
Numbers	−1%p	0	+4%p	−2%p	−2%p	+6%p	+1%p	+5%p	+2%p	+6%p	**+2%p**
Symmetry	+8%p	−1%p	−5%p	−2%p	−7%p	−3%p	+11%p	+10%p	−6%p	+3%p	**+1%p**

the machine learning models' performance correlates negatively with the complexity[3] of the individual rules.

The first finding relates to the one-shot learning (Lee et al., 2015) we discussed in Section 5.1. In contrast to humans, all three machine learning models can only do incremental learning. The chosen machine learning models require a certain amount of training instances to perform properly. However, there are special machine learning algorithms designed for one-shot learning, and this is an interesting topic for future work. The Bayesian one-shot algorithm (Fei-Fei et al., 2006) is an example of a machine learning algorithm that is able to learn via a single instance.

Regarding the second finding, the complexity of each rule can be defined by the number of basic components. For example, the rule *diagonal* consists of two basic components—either a diagonal line starting in the upper left corner and continuing to the lower right corner, or starting in the lower left corner and ending in the upper right corner. The rule itself leads to the best accuracy numbers for all rules, even outperforming humans. The rule *horizontal* is the combination of three different basic components—either a line in the first, second, or third row. The performance with one additional basic component is slightly lower and gets outperformed for the same number of training instances. The learning curve for the rule *symmetry* shows the third best performance. Because we use an uneven number for the rows and columns of our matrix, the axis on symmetry lies on three elements, which leads to six elements of our matrix being used for the rule. To fulfill the rule *symmetry*, one has to check pairwise whether two elements on opposite sides have the same value, which leads to three pairwise comparisons. By using two different symmetry axis, horizontal and vertical, this rule consists of six basic components. The rule *numbers* can also not simply be broken down into a number of basic components, which may lead to the worst performance of the machine learning models across all four rules.

Analyzing the machine learning model performances for each rule individually, the decision tree model shows remarkable accuracy for the rule *diagonal*, even outperforming humans. On the one hand, this relates to the comparatively bad performance of humans discussed in Section 5.1. On the other hand, the rule *diagonal* is unique: The feature x_5, referring to the central cell of the matrix, is true—irrespective of the direction of the diagonal line. This circumstance is easy to detect via a decision tree and is a good indication of whether the instance follows the rule or not.

The rules *numbers* and *symmetry* require the combination of several features and either counting (*numbers*) or comparing features, disregarding their binary status (*symmetry*). A logarithmic regression only looks at each feature individually and fails to detect both rules correctly. In machine learning, the process of feature engineering is utilized frequently—the machine learning model is trained with additional, often human-generated features that are a combination of other (original) features (Yu et al., 2010). For example, without feature engineering, a decision tree is not able to find the feature count that is necessary for the rule *numbers*. Furthermore, it has problems with detecting differences and ratios, which are essential for the rule *symmetry* (Heaton, 2016). This may explain the poor performance of the decision tree for the rule *symmetry* and its failure to learn the rule *numbers*. In contrast, a neural network like the used MLP can generate complex features like counting by its layer structure (Heaton, 2016). The better performance of the MLP compared to the other machine learning models is visible for rule *symmetry*. However, the MLP also fails to learn the rule *numbers* without feature engineering. This may be up to the low number of training instances or an unsuitable default configuration of layers and neurons for the rule *numbers*.

In accordance with Vapnik, Levin, and Cun (1994), in future work an analysis could be undertaken on the number of examples needed (depending on the class of learner and the class of pattern).

6. Conclusion

This article provides first insights on how learning performance differs between humans and SML models by comparing three different types when there is limited training data. The results of our experiment show a high dependency between performance and the underlying rules of the task. Whereas humans perform relatively similarly across all rules, SML models show big differences between the various patterns. On average, humans seem to learn more out of a small number of instances compared to machines. Interestingly, we can observe large differences in the learning curves of our SML models for the different rules we applied in our experiment. In half of the rules we employ, SML models reach the same level or even outperform humans. In the other half, SML models struggle to learn the respective patterns, as those require a deeper understanding that could be gained by a more complex combination of input features, referring to feature engineering. After 20 training instances, humans' performance does not improve anymore in our experiment—arguably due to cognitive overload. Machines learn slower and need more training instances compared to humans.

Our experiment design comes with several limitations: The number of experiment participants could be increased and lead to more, statistically significant results. In addition, we chose three different supervised machine learning algorithms out of hundreds of possible algorithms and parameter combinations. Our selection can only provide a hint of how SML performs in general—future work needs to discover more insights on the relationship between algorithm/parameters chosen and results achieved. The task characteristics have been selected out of a whole set of possibilities. In future, other task combinations need to be used to answer the question on how learning performance differs between humans and machines in general.

This work shows that further research on the application of supervised machine learning is needed. It is crucial for the application of SML, e.g. as part of autonomous agents, to gain a reliable understanding of tasks and their characteristics that are suitable for automation with SML models. From a business perspective, more research is required on the cost–benefit ratio of replacing human tasks with SML models. This may come with lower task performance, but provides the benefit of automation. We also stress the need for special supervised machine learning algorithms for limited training data, apart from inductive and genetic programming. Continuing the outlined road map of task characteristics and looking at other task characteristic combinations in the future, the single results of each combination will form a more general understanding of the differences between human learning and SML.

[3] In this case, complexity can be understood as the number of constraints/number of basic components needed to describe the pattern; the more basic components, the more complex the problem.

Table 5
Confusion matrix.

		Prediction outcome p	Prediction outcome n	total
actual value	p′	True Positive	False Negative	P′
	n′	False Positive	True Negative	N′
total		P	N	

Table 6
Exemplary confusion matrix for the best performing classifier decision tree for rule *diagonal* after 50 labeled instances.

		Prediction outcome p	Prediction outcome n	total
actual value	p′	22	2	24
	n′	7	19	26
total		29	21	

Table 7
Exemplary confusion matrix of worst performing classifier logistic regression for rule *symmetry* after 50 labeled instances.

		Prediction outcome p	Prediction outcome n	total
actual value	p′	11	14	25
	n′	10	15	25
total		21	29	

Table 8
Parameters of logistic regression.

Parameter	Value
penalty	I2
dual	false
tol	1e−4
C	1.0
fit_intercept	true
intercept_scaling	1
class_weight	none
random_state	none
solver	lbfgs
max_iter	100
multi_class	auto
verbose	0
warm_start	false
n_jobs	none
l1_ratio	none

Table 9
Parameters of decision tree classifier.

Parameter	Value
criterion	gini
splitter	best
max_depth	none
min_samples_split	2
min_samples_leaf	1
min_weight_fraction_leaf	0.0
max_features	none
random_state	none
max_leaf_nodes	none
min_impurity_decrease	0.0
min_impurity_split	0
class_weight	none
ccp_alpha	0.0

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

All data and code is available.

Appendix

Confusion matrices

See Tables 5–7.

Algorithm parameters

See Tables 8–10.

CNN results

See Table 11.

Instructions for participants

The experiment was conducted at the KD^{2}Lab[4] in German language. The participants were motivated according to Kvaløy et al. (2015). The following instructions are translated into English for this paper.

[4] https://www.kd2lab.kit.edu/.

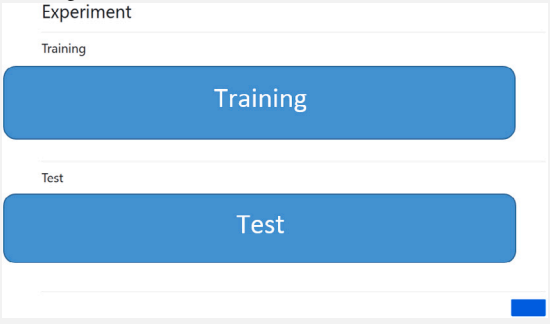
Table 10
Parameters of multi-layer perceptron classifier.

Parameter	Value
hidden_layer_sizes	1 hidden layer with 100 neurons
activation	relu
solver	lbfgs
alpha	0.0001
batch_size	auto
learning_rate	constant
learning_rate_init	0.001
power_t	0.5
max_iter	200
shuffle	true
random_state	none
tol	1e−4
verbose	false
warm_start	false
momentum	0.9
nesterovs_momentum	true
early_stopping	false
validation_fraction	0.1
beta_1	0.9
beta_2	0.999
epsilon	1e−8
n_iter_no_change	10
max_fun	15 000

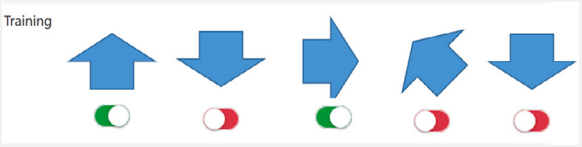
Table 11
Results of Convolutional Neural Network on different rules.

Rule	Number of training instances										Average
	5	10	15	20	25	30	35	40	45	50	
Diagonal	0.52	0.51	0.53	0.43	0.44	0.52	0.53	0.51	0.50	0.50	0.50
Horizontal	0.52	0.46	0.51	0.50	0.48	0.59	0.57	0.50	0.51	0.48	0.50
Numbers	0.55	0.51	0.50	0.47	0.49	0.54	0.51	0.51	0.48	0.50	0.51
Symmetry	0.45	0.55	0.48	0.50	0.50	0.56	0.53	0.50	0.50	0.53	0.50

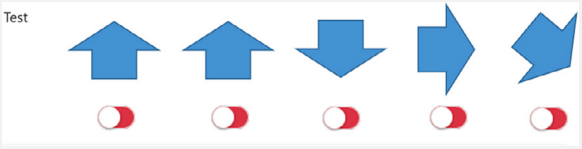
Dear participants,
Welcome to this study in the area of cognitive intelligence. In this study, you will be shown pictures in which you are to recognize rules and reproduce them correctly. With your help, we can better understand thought processes of humans, especially in comparison to artificial intelligence.
The experiment consists of two parts, an upper section (= training) and a lower section (= test).



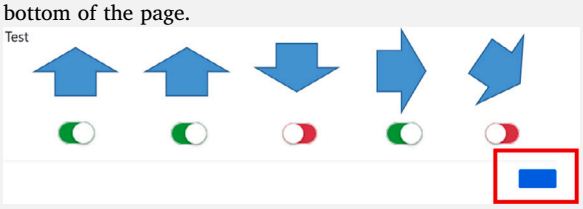
In the training area, five images are displayed at the beginning, each with an invariable slider. This slider indicates whether the respective image conforms to a certain rule (green slider) or not (red slider).



In this example, the images are represented by abstract arrows. The rule used is "point up or point right". This rule is fulfilled by images 1 and 3, for this reason the sliders are displayed in green. The other images do not satisfy the rule used and consequently their sliders are red.
In the test area, you will be shown another five images with variable sliders. Your task is to adjust the sliders according to the rule learned in the training area.



In this example, image 1, 2 and 4 fulfill the rule "point up or point right". For this reason, the sliders of image 1, 2 and 4 must be changed by clicking on green. Image 3 and 5 do not fulfill the rule, therefore no change of the sliders is necessary there.
If you are satisfied with your sliders in the test area, confirm your assignment by clicking the button **"Submit"** at the bottom of the page.



This will end your first round and start the second round. The **rule used** will remain **existing** in the second round and you will still be able to see and use the old images from the first round in the training area. In addition, in the second round you will receive five more training images and five new images in the test area. You also finish the second round by clicking "Submit". The further rounds proceed analogously; you will receive five more training images and five new images in the test area per round.
After the tenth round, the game is over and you will see a "Game finished" on your screen. From there you will be redirected back to the introduction page. By reconfirming that you understand the rules of this experiment, the second game begins, but with a **different rule** for the images than before.
You will play **four games of 10 rounds each** in this experiment. The better you perform on the test area assignment compared to the other study participants during the four games, the higher your individual reward will be.

Helpful hints:

- The order of the images does not matter for the rule. Also no rotation of the pictures is necessary to recognize the rule.
- For one game we have planned about 15 min—so you have enough time to think about your assignment.
- Especially at the beginning of the game with few training pictures it can be difficult or impossible to find the rule you are looking for. Do not despair! Guess or click "Submit" without changing the sliders to get new images. In the following pass you will receive five more images, which will make it easier for you to find the rule you are looking for.
- You will receive the images by the random principle. As a result, in the first passes you may see few or no pictures created according to the rule.
- During the game you may think you have found a rule, but it turns out to be wrong as the game progresses. Be sure to check in each run to see if you need to adjust your rule if necessary.

Once you have finished all four games, please remain seated in your booth. Do not close your program and stay on the welcome page. Only when all participants have finished the experiment, the payout can take place. To do so, please have the document you were given filled out with your **demographic data**.
If any technical problems occur during the experiment, please step out of the booth and come to us. Please avoid using the "forward" or "back" arrow of the browser during the duration of the experiment.
The use of cell phones is strictly prohibited.
We wish you good luck with the following experiment.
Thank you very much for your support! You are making an important contribution to current research!

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., et al. (2016). {TensorFlow}: A system for {Large – Scale} machine learning. In *12th USENIX symposium on operating systems design and implementation (OSDI 16)* (pp. 265–283).

Adler, A., & Schuckers, M. E. (2007). Comparing human and automatic face recognition performance. *IEEE Transactions on Systems, Man and Cybernetics, Part B (Cybernetics)*, *37*(5), 1248–1255.

Ayres, P., & Paas, F. (2009). Interdisciplinary perspectives inspiring a new generation of cognitive load research. *Educational Psychology Review*, *21*(1), 1–9.

Baier, L., Jöhren, F., & Seebacher, S. (2019). Challenges in the deployment and operation of machine learning in practice. In *Proceedings of the 27th European conference on information systems*.

Bandura, A. (1986). *Social foundations of thought and action*. Englewood Cliffs, NJ.

Banzhaf, W., Nordin, P., Keller, R. E., & Francone, F. D. (1998). *Genetic programming*. Springer.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Statistical Methodology*, *57*(1), 289–300.

Bock, O., Baetge, I., & Nicklisch, A. (2014). Hroot: hamburg registration and organization online tool. *European Economic Review*, *71*, 117–120.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *The Wadsworth statistics/probability series, Classification and regression trees*. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.

Burr, C., Cristianini, N., & Ladyman, J. (2018). An analysis of the interaction between intelligent software agents and human users. *Minds and Machines*, *28*(4), 735–774.

Casner, S. M., Hutchins, E. L., & Norman, D. (2016). The challenges of partially automated driving. *Communications of the ACM*.

Cawley, G. C., & Talbot, N. L. (2010). On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, *11*, 2079–2107.

Chang, H. S., Fu, M. C., Hu, J., & Marcus, S. I. (2016). Google deep mind's alphaGo. *OR/MS Today*, *43*(5), 24–29.

Chollet, F. (2015). Keras. https://github.com/fchollet/keras.

Coren, S., Ward, L. M., Porac, C., & Fraser, R. (1978). The effect of optical blur on visual-geometric illusions. *Bulletin of the Psychonomic Society*, *11*(6), 390–392.

Dellermann, D., Ebel, P., Söllner, M., & Leimeister, J. M. (2019). Hybrid intelligence. *Business & Information Systems Engineering*, 1–7.

Dietterich, T. (1996). Machine learning. *ACM Computing Surveys*, *28*(4es).

Dubey, R., Agrawal, P., Pathak, D., Griffiths, T., & Efros, A. (2018). Investigating human priors for playing video games. In *International conference on machine learning* (pp. 1348–1356).

Dupoux, E. (2018). Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, *173*, 43–59.

Ebbinghaus, H. (1885). *Über das gedächtnis: Untersuchungen zur experimentellen psychologie*. Duncker & Humblot.

Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, *27*(18), 2827–2832.

Elsayed, G., Shankar, S., Cheung, B., Papernot, N., Kurakin, A., Goodfellow, I., et al. (2018). Adversarial examples that fool both computer vision and time-limited humans. In *Advances in neural information processing systems* (pp. 3910–3920).

Eysenck, M. W., & Keane, M. T. (2015). *Cognitive psychology: A student's handbook*. Psychology Press.

Fan, X., Chen, P.-C., & Yen, J. (2010). Learning HMM-based cognitive load models for supporting human-agent teamwork. *Cognitive Systems Research*, *11*(1), 108–119.

Favela, L. H., & Martin, J. (2017). "Cognition" and dynamical cognitive science. *Minds and Machines*, *27*(2), 331–355.

Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *28*(4), 594–611.

Feng, N., & Sun, B. (2019). On simulating one-trial learning using morphological neural networks. *Cognitive Systems Research*, *53*, 61–70.

Florez, J. F. M. (2015). *Michael S. Gazzaniga, George R. Mangun (Hrsg.): The cognitive neurosciences, 5th edition*. Springer.

Girden, E. R. (1992). *ANOVA: Repeated measures. 84*. Sage.

Glorot, X., & Bengio, Y. (2010). Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 249–256).

Gonzalez, C., Best, B., Healy, A. F., Kole, J. A., & Bourne, L. E., Jr. (2011). A cognitive modeling account of simultaneous learning and fatigue effects. *Cognitive Systems Research*, *12*(1), 19–32.

Grace, K., Salvatier, J., Dafoe, A., Zhang, B., & Evans, O. (2018). When will AI exceed human performance? Evidence from AI experts. *Journal of Artificial Intelligence Research*, *62*, 729–754.

Griffiths, T. L., Chater, N., Kemp, C., Perfors, A., & Tenenbaum, J. B. (2010). Probabilistic models of cognition: Exploring representations and inductive biases. *Trends in Cognitive Sciences*, *14*(8), 357–364.

Griffiths, T. L., Steyvers, M., & Firl, A. (2007). Google and the mind: Predicting fluency with PageRank. *Psychological Science*, *18*(12), 1069–1076.

Hastie, T., Tibshirani, R., & Friedman, J. (2017). *The elements of statistical learning: Data mining, inference and prediction, Vol. 9*. Springer.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Heaton, J. (2016). An empirical analysis of feature engineering for predictive modeling. In *SoutheastCon 2016* (pp. 1–6). IEEE.

Hemmer, P., Kühl, N., & Schöffer, J. (2022). Deal: deep evidential active learning for image classification. In *Deep Learning Applications, Volume 3* (pp. 171–192). Springer.

Hernández-Orallo, J. (2017a). Evaluation in artificial intelligence: from task-oriented to ability-oriented measurement. *Artificial Intelligence Review*, *48*(3), 397–447.

Hernández-Orallo, J. (2017b). *The measure of all minds: Evaluating natural and artificial intelligence*. Cambridge University Press.

Hernández-Orallo, J., Martínez-Plumed, F., Schmid, U., Siebers, M., & Dowe, D. L. (2016). Computer models solving intelligence test problems: Progress and implications. *Artificial Intelligence*, *230*, 74–107.

Hinton, G., Deng, L., Yu, D., Dahl, G., Mohamed, A.-r., Jaitly, N., et al. (2012). Deep neural networks for acoustic modeling in speech recognition. *IEEE Signal Processing Magazine*, *29*.

Hirt, R., Kühl, N. J., & Satzger, G. (2017). An end-to-end process model for supervised machine learning classification: from problem to deployment in information systems. In *Designing the digital transformation: DESRIST 2017 research in progress proceedings of the 12th international conference on design science research in information systems and technology. Karlsruhe, Germany. 30 May-1 Jun.* (pp. 55–63). Karlsruher Institut für Technologie (KIT).

Hoffmann, A. (2010). Can machines think? An old question reformulated. *Minds and Machines*, *20*(2), 203–212.

Hutto, D. D., & Kirchhoff, M. D. (2015). Looking beyond the brain: Social neuroscience meets narrative practice. *Cognitive Systems Research*, *34*, 5–17.

Insa-Cabrera, J., Dowe, D. L., Espana-Cubillo, S., Hernández-Lloreda, M. V., & Hernández-Orallo, J. (2011). Comparing humans and AI agents. In *International conference on artificial general intelligence* (pp. 122–132). Springer.

Janitza, S., & Hornung, R. (2018). On the overestimation of random forest's out-of-bag error. *PLoS One*, *13*(8), Article e0201904.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science*, *349*(6245), 255–260.

Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, *4*, 237–285.

Kim, B., Reif, E., Wattenberg, M., & Bengio, S. (2019). Do neural networks show gestalt phenomena? An exploration of the law of closure. arXiv preprint arXiv:1903.01069.

Kogler, J. E., & Pessoa, O. F. (2017). Celebration of twenty years promoting cognitive science. *Cognitive Systems Research*, *100*(43), 125–127.

Köhler, W. (1967). Gestalt psychology. *Psychological Research*, *31*(1), XVIII–XXX.

Kotovsky, K., & Simon, H. A. (1973). Empirical tests of a theory of human acquisition of concepts for sequential patterns. *Cognitive Psychology*, *4*(3), 399–424.

Koza, J. R., Bennett, F. H., Andre, D., & Keane, M. A. (1996). Automated design of both the topology and sizing of analog electrical circuits using genetic programming. In *Artificial intelligence in design '96* (pp. 151–170). Springer.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*.

Kvaløy, O., Nieken, P., & Schöttner, A. (2015). Hidden benefits of reward: A field experiment on motivation and monetary incentives. *European Economic Review*, *76*, 188–199.

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, *350*(6266), 1332–1338.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*.

Lee, S. W., O'Doherty, J. P., & Shimojo, S. (2015). Neural computations mediating one-shot learning in the human brain. *PLoS Biology*, *13*(4).

Lieder, F., Griffiths, T., & Goodman, N. (2012). Burn-in, bias, and the rationality of anchoring. In *Advances in neural information processing systems* (pp. 2690–2798).

Lieto, A., Lebiere, C., & Oltramari, A. (2018). The knowledge level in cognitive architectures: Current limitations and possible developments. *Cognitive Systems Research*, *48*, 39–55.

Lin, F., & Cohen, W. W. (2010). Semi-supervised classification of network data using very few labels. In *2010 international conference on advances in social networks analysis and mining* (pp. 192–199). IEEE.

Liu, D. C., & Nocedal, J. (1989). On the limited memory BFGS method for large scale optimization. *Mathematical Programming*, *45*(1–3), 503–528.

Lucas, C. G., Griffiths, T. L., Williams, J. J., & Kalish, M. L. (2015). A rational model of function learning. *Psychonomic Bulletin & Review*, *22*(5), 1193–1215.

Łupkowski, P., & Jurowska, P. (2019). The minimum intelligent signal test (MIST) as an alternative to the turing test. *Diametros*, *16*(59), 35–47.

Makridakis, S. (2017). The forthcoming artificial intelligence (AI) revolution: its impact on society and firms. *Futures*, *90*, 46–60.

Marcus, G. F. (2018). *The algebraic mind: Integrating connectionism and cognitive science*. MIT Press.

McKinstry, C. (1997). Minimum intelligent signal test: an objective turing test. *Canadian Artificial Intelligence*, 17–18.

Moreno, R., & Mayer, R. E. (1999). Visual presentations in multimedia learning: conditions that overload visual working memory. In *International conference on advances in visual information systems* (pp. 798–805). Springer.

Muggleton, S. (1991). Inductive logic programming. *New Generation Computing*, *8*(4), 295–318.

Müller, V. C., & Bostrom, N. (2016). Future progress in artificial intelligence: a survey of expert opinion. In *Fundamental issues of artificial intelligence* (pp. 555–572). Springer.

Olsson, R. (1995). Inductive functional programming using incremental program transformation. *Artificial Intelligence*, *74*(1), 55–81.

Olteţeanu, A.-M., Falomir, Z., & Freksa, C. (2016). Artificial cognitive systems that can answer human creativity tests: An approach and two case studies. *IEEE Transactions on Cognitive and Developmental Systems*, *10*(2), 469–475.

O'Regan, J. K. (2012). How to build a robot that is conscious and feels. *Minds and Machines*, *22*(2), 117–136.

Ormrod, J. E., & Davis, K. M. (2004). *Human learning*. Merrill London.

Osuna, E., Rodríguez, L.-F., Gutierrez-Garcia, J. O., & Castro, L. A. (2020). Development of computational models of emotions: A software engineering perspective. *Cognitive Systems Research*, *60*, 1–19.

Paas, F., Renkl, A., & Sweller, J. (2003). Cognitive load theory and instructional design: recent developments. *Educational Psychologist*, *38*(1), 1–4.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Perlich, C., Provost, F., & Simonoff, J. S. (2003). Tree induction vs. logistic regression: a learning-curve analysis. *Journal of Machine Learning Research*, *4*(Jun), 211–255.

Peterson, J. C., Abbott, J. T., & Griffiths, T. L. (2018). Evaluating (and improving) the correspondence between deep neural networks and human representations. *Cognitive Science*, *42*(8), 2648–2669.

Raghupathi, W., & Raghupathi, V. (2014). Big data analytics in healthcare: promise and potential. *Health Information Science and Systems*, *2*(1), 3.

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*(33), 7255–7269.

Raven, J. (2000). The raven's progressive matrices: change and stability over culture and time. *Cognitive Psychology*, *41*(1), 1–48.

Rhee, P. K., Erdenee, E., Kyun, S. D., Ahmed, M. U., & Jin, S. (2017). Active and semi-supervised learning for object detection with imperfect data. *Cognitive Systems Research*, *45*, 109–123.

Rosenthal, T. L., & Zimmerman, B. J. (1978). *Social learning and cognition*. Academic Press.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., et al. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Schmid, U., & Kitzelmann, E. (2011). Inductive rule learning on the knowledge level. *Cognitive Systems Research*, *12*(3–4), 237–248.

Seeber, I., Bittner, E., Briggs, R. O., de Vreede, T., de Vreede, G.-J., Elkins, A., et al. (2019). Machines as teammates: a research agenda on AI in team collaboration. *Information & Management*.

Settles, B. (2009). *Active learning literature survey*: *Technical Report*, University of Wisconsin-Madison Department of Computer Sciences.

Shaffer, R. (2017). Cognitive load and issue engagement in congressional discourse. *Cognitive Systems Research*, *44*, 89–99.

Smart, P. R. (2018). Human-extended machine cognition. *Cognitive Systems Research*, *49*, 9–23.

Stehman, S. V. (1997). Selecting and interpreting measures of thematic classification accuracy. *Remote Sensing of Environment*, *62*(1), 77–89.

Sternberg, R. J., & Sternberg, K. (2016). *Cognitive psychology*. Nelson Education.

Sun, Q., & Pfahringer, B. (2013). Pairwise meta-rules for better meta-learning-based algorithm ranking. *Machine Learning*, *93*(1), 141–161.

Sweller, J., Van Merrienboer, J. J., & Paas, F. G. (1998). Cognitive architecture and instructional design. *Educational Psychology Review*, *10*(3), 251–296.

Tavani, H. T. (2011). Can we develop artificial agents capable of making good moral decisions? *Minds and Machines*, *21*(3), 465–474.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Thagard, P. (2005). *Mind: Introduction to cognitive science, Vol. 17*. Cambridge, MA: MIT press.

Thorndike, E. L. (1913). *The psychology of learning, Vol. 2*. Teachers College, Columbia University.

Van Merrienboer, J. J., & Sweller, J. (2005). Cognitive load theory and complex learning: recent developments and future directions. *Educational Psychology Review*, *17*(2), 147–177.

Vapnik, V., Levin, E., & Cun, Y. L. (1994). Measuring the VC-dimension of a learning machine. *Neural Computation*, *6*(5), 851–876.

Veale, T., Gervás, P., & y Pérez, R. P. (2010). Computational creativity: A continuing journey. *Minds and Machines*, *20*(4), 483–487.

Velik, R. (2010). Why machines cannot feel. *Minds and Machines*, *20*(1), 1–18.

Vessey, I. (1991). Cognitive fit: a theory-based analysis of the graphs versus tables literature. *Decision Sciences*, *22*(2), 219–240.

Vinyals, O., Blundell, C., Lillicrap, T., Wierstra, D., et al. (2016). Matching networks for one shot learning. In *Advances in neural information processing systems* (pp. 3630–3638).

Vul, E., Goodman, N., Griffiths, T. L., & Tenenbaum, J. B. (2014). One and done? Optimal decisions from very few samples. *Cognitive Science*, *38*(4), 599–637.

Vygotsky, L. S. (1964). Thought and language. *Annals of Dyslexia*, *14*(1), 97–98.

Vygotsky, L. S. (1980). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, K., Wang, B., & Peng, L. (2009). CVAP: validation for cluster analyses. *Data Science Journal*.

Witten, I. H., Manzara, L. C., & Conklin, D. (1994). Comparing human and computational models of music prediction. *Computer Music Journal*, *18*(1), 70–80.

Yilmaz, L., Franco-Watkins, A., & Kroecker, T. S. (2017). Computational models of ethical decision-making: A coherence-driven reflective equilibrium model. *Cognitive Systems Research*, *46*, 61–74.

Yu, H.-F., Lo, H.-Y., Hsieh, H.-P., Lou, J.-K., McKenzie, T. G., Chou, J.-W., et al. (2010). Feature engineering and classifier ensemble for KDD cup 2010. In *KDD Cup* (pp. 1–16).

Yu, T., & Zhu, H. (2020). Hyper-parameter optimization: A review of algorithms and applications. arXiv preprint arXiv:2003.05689.

Yuen, K. K. (1974). The two-sample trimmed t for unequal population variances. *Biometrika*, *61*(1), 165–170.

Zhou, Z., & Firestone, C. (2019). Humans can decipher adversarial images. *Nature Communications*, *10*(1), 1–9.