KI - Künstliche Intelligenz (2020) 34:199–207
https://doi.org/10.1007/s13218-020-00638-x

DISCUSSION

How and What Can Humans Learn from Being in the Loop?

Invoking Contradiction Learning as a Measure to Make Humans Smarter

Benjamin M. Abdel-Karim[1] · Nicolas Pfeuffer[1] · Gernot Rohde[1,2] · Oliver Hinz[1]

Received: 16 September 2019 / Accepted: 13 January 2020 / Published online: 22 January 2020

Abstract
This article discusses the counterpart of interactive machine learning, i.e., human learning while being in the loop in a human-machine collaboration. For such cases we propose the use of a Contradiction Matrix to assess the overlap and the contradictions of human and machine predictions. We show in a small-scaled user study with experts in the area of pneumology (1) that machine-learning based systems can classify X-rays with respect to diseases with a meaningful accuracy, (2) humans partly use contradictions to reconsider their initial diagnosis, and (3) that this leads to a higher overlap between human and machine diagnoses at the end of the collaboration situation. We argue that disclosure of information on diagnosis uncertainty can be beneficial to make the human expert reconsider her or his initial assessment which may ultimately result in a deliberate agreement. In the light of the observations from our project, it becomes apparent that collaborative learning in such a human-in-the-loop scenario could lead to mutual benefits for both human learning and interactive machine learning. Bearing the differences in reasoning and learning processes of humans and intelligent systems in mind, we argue that interdisciplinary research teams have the best chances at tackling this undertaking and generating valuable insights.

Keywords Machine teaching · Machine learning · Experts · Feedback loop

1 Introduction

Machine learning (ML) and especially deep learning (DL) have seen a dramatic resurgence in the past decade, largely driven by increases in computational power and the availability of large datasets that enable a faster training and ultimately more accurate models [38]. Although in many use cases, ML models have been reported to efficiently automate various tasks like automated object recognition [12], it appears that we have only yet begun to harness the benefits of these technological advancements. The victory of the AlphaGo algorithm over the world champion of the highly complex board game Go (estimated 2^{10170} board positions) stands as a beacon example of this inferential power [28].

At the time where AlphaGo defeated the world champion, Silver et al. [28] had already submitted a paper to the journal "Nature" featuring an even more powerful version of AlphaGo, named AlphaGo Zero. By relying only on its reinforcement learning strategy and experiences from matches vs. AlphaGo only, AlphaGo Zero became by a magnitude more powerful and innovative than its predecessor, without the need of directly relying on human knowledge [29]. AlphaGo Zero's winning strategies were previously unknown and therefore difficult to identify, understand and combat for humans and also for AlphaGo. Although this victory of AlphaGo and the advancements of AlphaGo Zero were regarded as major milestones in artificial intelligence (AI) research itself [28], it could also constitute the advent of an equally important new pattern for human-machine collaboration. Fan Hui, former European champion of Go who had also lost to AlphaGo, mentioned the following in an interview with nature news: "[...] The problem is humans

✉ Oliver Hinz
ohinz@wiwi.uni-frankfurt.de

Benjamin M. Abdel-Karim
abdel-karim@wiwi.uni-frankfurt.de

[1] Chair of Information Systems and Information Management, Goethe University Frankfurt, Theodor-W.-Adorno-Platz 4, 60323 Frankfurt am Main, Germany

[2] University Hospital Frankfurt, Theodor-Stern-Kai 7, 60590 Frankfurt am Main, Germany

sometimes make very big mistakes, because we are human. [...] So I lose, I study the game, and maybe I change my game. I think it's a good thing for the future."[1]

This statement implies, that by making mistakes, we may be able to learn from them with the aid of machines and therefore improve our own intelligence. The concept of learning from mistakes may be nothing new to us, yet it differs drastically between cultural regions [18]. In schools within the western hemisphere, the process of learning from errors appears to be mostly limited to the recognition of errors by the teacher, disregardment of the error and drawing attention to proven methods [32]. This approach may lead us to apply strategies that allow us to avoid errors at all and stick to proven methods which may hinder the emergence of new and innovative problem solutions. In comparison to this learning model, it was found that a different approach of dealing with errors in the learning process resulted in much higher scores of Japanese students in international math competitions.

It appeared that—because students are encouraged to recognize, analyze and confront themselves with the errors—they tend to find the solutions to the errors themselves and they develop stronger capabilities in inference and problem solution [18, 32].

In effect, this very much resembles the ideas represented in the domain of ML, where an algorithm trains itself, calculates its own error rate in the task it is set to perform and continuously tries to optimize its own capabilities. While new patterns of mass-training algorithms such as human-in-the-loop or interactive machine learning [1] have evolved that utilize human judgment to show algorithms where their errors lie, we appear to increasingly accept that algorithms need to make errors to become better.

These insights suggest that, instead of just harnessing the power of ML solutions as decision support systems (DSS) [16] or automation mechanisms and instead of continuously empowering these algorithms, it is time that we accept our errors and make use of those algorithms as teachers that may see things differently than we do [18]. It may indeed be possible, that this approach is not only limited to domains like the board games Go or chess, such that we may be able to enhance our knowledge with the help of machines in a plethora of domains. However, areas of conflict arise for the operationalization of this stream of thought, due to the way humans deal with conflicting information which is delivered by machines.

For one, there is the danger of an automation bias, which may lead to uncritical acceptance of possibly erroneous machine behavior [20]. A conjunction of learned helplessness resulting from automation bias, as well as automation bias itself has led to fatal decision making in the area of aviation in the past, resulting in the deaths of hundreds[2]. In this regard, humans tend to trust too much in the capabilities of the technology and instead of learning to improve their own skills, they simply rely on the technology. Secondly, humans also tend to be overconfident or anchored to their own opinion and may be reluctant to change their judgments and beliefs, because they either strongly believe in the ex-ante available information that they are able to recall using their own knowledge [7] or they are highly confident of their own expertise and therefore are not willing to use conflicting information to adapt their own assessment [20].

Thirdly, because we have learned certain strategies or procedures to deal with problems, we tend to stick to these strategies [14]. Analogical reasoning leads us to compare new problems to already known problems and therefore look for already known solutions, which makes us inert to find new ways of solving a problem [17, 27]. Inertia has been shown to drive individuals to use known systems instead of alternative and better systems [24]. Aside from the individual level, Nijssen et al. [22] even found evidence that knowledge inertia lead organizations to be less innovative. Some aspects of inertia, such as knowledge and procedural inertia, appear to be a hindrance to successfully learning novel solution methods [14].

These and related factors create a tension around the idea of us being taught by intelligent machines on the basis of our errors. Nevertheless, we believe that mankind needs to grasp the current opportunity of man-computer symbiosis [15], such that not only algorithms become more intelligent, but also humans may widen their horizons of reasoning and understanding with the help of the new technology. In this discussion we will start to answer the following question with respect to these conflicting pieces of information: How can we benefit from being a human-in-the-loop and how can we learn from our mistakes by the help of intelligent machines?

2 Human-in-the-Loop and Contradiction Learning

First, it is important to note that the concept proposed by this paper makes only sense for situations where humans work with ML-based (and especially ML-based) systems and where there is still some level of inaccuracy on side of the user and on the side of the system. This applies certainly for recent IML systems—the focus of this Special Issue—but can also be extended to decision support systems that use

[1] https://www.nature.com/news/go-players-react-to-computer-defeat-1.19255

[2] https://bit.ly/2HvO6KY.

ML, but the concept cannot be extended to decision automation in general.

If ML-based systems and users agree on an outcome, the likelihood that the system or the human learns, is low. However, if there is a disagreement, this discrepancy may be useful in two ways: the user can overrule and teach the system in case of IML, but this situation can also lead to a reconsideration of the own judgment on the side of the user. In the long-run, this can lead to the detection of unknown patterns on the side of the individual user which would make her or him smarter and could also lead—ultimately—to new discoveries that would advance our understanding of the world.

If we think of the AlphaGo Zero example from the introduction, this aspect of the human-computer symbiosis becomes apparent: AlphaGo Zero tried out strategies that had not been pursued by humans in their short history of 2500 years of Go game playing and that thus had not been taught by adepts to their disciples. These strategies did not belong to the human "handbook of successful Go playing". AlphaGo Zero played millions of games and tried out very new and innovative strategies and found strategies that can be deemed to be very successful and ultimately "pulverized" even the best players and systems that had been trained on the base of human vs. human Go matches. Experts and professional Go players can now use AlphaGo Zero to learn these new strategies which thus is a nice instantiation of the general concept that this article suggests.

One may argue that games like Chess and Go are in principal competitive and specifically AlphaGo does not explain which particular move in the player's strategy was especially important for the outcome, yet the idea of being taught by machines may greatly differ to ML-based system usage in practice. Nevertheless, the principle concept of machines teaching us new knowledge based on our own errors is independent on the contextual setting (e.g., competitive vs. cooperative) and may directly be related to us making decisions on our knowledge and the machine making decisions based on ML-trained knowledge which may supersede our own knowledge in certain problem domains (e.g., either through pattern recognition on large amounts of data, or through reinforcement learning which enables the search for efficient and potentially novel strategies in an increasingly high-dimensional problem space). Therefore, we may in principle in many situations be able to learn from machines that reveal our own errors and which may have led us to wrong or inefficient decisions. This, in effect, may even enable us to improve our own decision making capabilities.

Theoretically, the aforementioned concept is linked to what we, in our research context, refer to as "Contradiction Error Learning". In psychology, an increasing number of research works is dedicated to the research on learning from errors [18]. While traditional literature on learning suggests that errors must be avoided by all means during learning [4, 30], scholars like Metcalfe [18] postulate that the western world obstructs its own learning possibilities by the traditional paradigm of error avoidance which we are taught from our early lives on. The power of learning from errors is underlined by studies that investigate the superior performance of Japanese students' versus American students' mathematical capabilities. Stevenson and Stigler [32] for example found that while American students are taught to avoid errors and to simply follow established ways, Japanese students are encouraged to learn from their errors, which enables them to find more creative and possibly novel ways of solving problems.

As literature on learning from errors states, feedback on the errors a learner makes is obligatory [13, 19] and must be understood and accepted by the learner; only then he or she is able to process the information that the contradiction carries and hence can improve his or her skills on this basis [3]. While receiving constructive and corrective feedback on making errors is necessary for a learning effect to occur, the question is if the teacher has to be necessarily human. Since tutoring systems have shown that we are very well able to learn from systems (e.g., [2, 35]), it appears plausible to implement intelligent machines that help us in overcoming erroneous decisions, such that we become more competent and skilled.

However, we must also acknowledge that people tend to be too uncritical with machine reasoning and therefore are prone to automation bias [21], too uncritical with their own knowledge and thus fall victim to an overconfidence bias [7] or too inert to their existing knowledge to learn from their errors [14]. It remains therefore an open question—and lastly an empirical question—whether ML systems can be used to make humans, i.e., their users, smarter or can even be used to make totally new discoveries.

3 Testing the Concept in Real Life: First Insights

In order to motivate our fellow researchers to explore the question *if humans are able to learn from their errors with the aid of ML-based DSS*, we want to provide the first steps towards an answer with this discussion and an explorative experiment. To take the first steps in giving an answer to the open question and to improve our understanding on the human-in-the-loop concept, we conduct a two-step sequential experiment with a medical expert in the area of pneumonology and observe his interaction with an ML-based DSS with the aid of a think-aloud protocol. Through our discussion including a small-scale, explorative experiment we aim to (1) provide first indicators for an error-learning process induced by ML-based DSS and (2) we aim to motivate fellow researchers to thorougly investigate the research phenomenon of error-learning with the aid of ML-based DSS. To provide a better understanding of our methodology, we first present the components of our experiment in more detail in the following.

3.1 Development of the ML-Based System

For the purpose of this study, we train a Convolutional Neural Network (CNN) with data from the Chest X-ray database from the National Institutes of Health[3] that comprises 112,000 frontal view X-ray images of 30,805 unique patients. These cases have already been labeled with respect to the most common lung diseases. This dataset contains 14 common thorax disease categories. Therefore this dataset is an extension of the eight common disease patterns listed in [37]. For our experiment, we decided to use diseases which a medical expert should clearly recognize on X-ray images, especially against the background of the clinic daily business. A medical expert helped us to create this sub sample. Based on the professional expertise, we have selected nine diseases (see Table 1).

For data preprocessing we decided to remove X-ray images with insufficient quality from the training set and the validation set to retain an unbiased and clean data set for the best possible training effect. We identified X-rays with such insufficient quality, i.e., blur that can occur when the patient moves during the recording, X-rays that contain medical devices, which partially cover the lung, and photographed X-rays, which differ in pixel structure from the actual X-rays. Due to the fact that an automation of this filtering process is difficult, we have selected the appropriate X-rays manually.

Recent research has investigated advantages and disadvantages of this approach (for example [8]). We decided to train the classifier for the most common lung diseases [37]. The raw data contains sets of X-ray images that correspond to the same shot but from multiple angles, or multiple shots of similar conditions on the same person. We manually selected unique and high-quality X-ray images from the raw data set. We made sure that a patient is not included several times with different diseases. The selection of unique and high-quality X-rays serves two purposes: (1) we aimed to ascertain comparability in the performance between human and machine. (2) In addition, previous work showed clearly that insufficient quality of input data has a negative influence on the accuracy of the models. In addition, we only use X-rays with a distinctive diagnosis. This means that patients with multiple diseases at the same time are excluded. We use 90–100 images for each class and apply image augmentation. For this purpose, we have extended the individual classes accordingly, as shown for example in the work of Esteva et al. [8], such that the number of images corresponds to the distribution of the different diseases. Classes with rotated X-rays are determined by a random selection of existing images and this selected image is rotated randomly between −15° and 15°. Therefore, we apply an augmentation factor of approximately 44. Based on our data preprocessing we use 40,000 images for training. As the ground truth for the X-ray diagnosis we use the labels for each image provided in the Chest X-ray database. Against this background, our approach of data preparation follows the recommendations of [8].

For the training procedure, we selected the AlexNet architecture. This decision is the result of many pre-tests with different CNN models such as the Inception V3. AlexNet is based on the work of [12] and is a well-known model in the research area of CNN. This CNN consists of eight weighted layers, consisting of five convolutional layers followed by three fully connected layers [12]. We feed images into the network with a 277×277 pixels resolution. All layers use the same global learning rate of 0.001 and a decay factor of 16 [33]. Additionally, we use the RMSProp optimizer with a momentum of 0.9 [8]. Batches are sampled using a fixed batch size of 16 images in line with [11]. In line with Irvin et al. [11] and Esteva et al. [8] we utilize a validation set for hyperparameter tuning and assess the performance of our model on a distinct test set. The procedure of using an additional test set is considered state of the art and is intended to ensure the generalizability of the trained ML model. The test set contains 200 unique images from 200 patients randomly sampled from the full preprocessed data set. In this step of our work, we put a lot of effort in ensuring that no patient X-ray image overlaps with data from our training procedure. Therefore, the test set contains unique X-rays that are not contained in the training set and validation set [8, 11].

3.2 Experimental Setup

After we have successfully trained the system, we take the human into the loop. Human-in-the-Loop in the context of machine learning is defined as a procedure that integrates human capabilities in the entire machine learning process [26]. The idea behind our experiment is that instead of just harnessing the power of ML solutions as decision support systems (DSS) or automation mechanisms and instead of continuously empowering these algorithms by human input, we may utilize those algorithms as teachers that may see things differently than we do [18].

We invite a medical expert to identify the diseases based on randomly selected X-ray images (which is definitely not an easy question because usually more than this visual information—which is often provided with low resolution—is

[3] A current version containing 14 disease labels can be found under https://nihcc.app.box.com/v/ChestXray-NIHCC. We downloaded the data in September 2018.

available from e.g., a radiologist). The experiment is conducted in two steps: In step one, we ask the the medical expert to make a diagnosis based on an X-ray image that is presented to him within a UI of an application. In each diagnosis decision, the medical expert is able to choose one of eight potential diagnoses. In addition, the medical expert must provide his degree of certainty with respect to the diagnosis. In step two, we confront the expert with the results of our ML-based system for each previously diagnosed X-ray image. The results of the ML-based system is presented in the adapted application UI from step one by providing the ML-based system's diagnosis and the level of decision certainty to the medical expert. Both experimental steps are performed with the same 20 X-ray images which contain instances of all eight possible disease labels. The X-ray images are mainly taken from the test set, with the exception of the instances of fibrosis which we took from the validation set, because our algorithm did not perform well for fibrosis images from the test set. It must be noted that—aside from assuring that the X-ray images were of high quality, such that the expert would be able to assess them—we especially selected X-ray images where our ML-based system achieved high accuracy. This was done such that the expert would not immediately disregard the information of the system, such that we could observe at all if the expert would even consider learning from the system. Furthermore, we opted for the number of 20 images to meet the practical requirements (i.e., tight time frames due to hospital work) and not to generate too much cognitive stress for the subjects which could have led to biased results.

We use a think-aloud protocol to better understand the cognitive process and content of a problem-solving strategy while the expert reflects his decisions. This is a frequently used technique which has been successfully applied in the domains of Human-Computer Interaction and Information Systems [36]. We log the verbalization and write it down accordingly [5], which primarily serves the knowledge extraction to make expert knowledge accessible. In the context of computer science and IS research, the usage of think-aloud protocols in interface design are effective to capture both, a user's approach to a task and why problems occur when users interact with computer systems [34]. Figure 1 shows the experimental setup.

4 Results

4.1 ML-Based System Results

In this article we provide very first insights with respect to the proposed approach. Table 1 presents the model performance which is in line with other works.

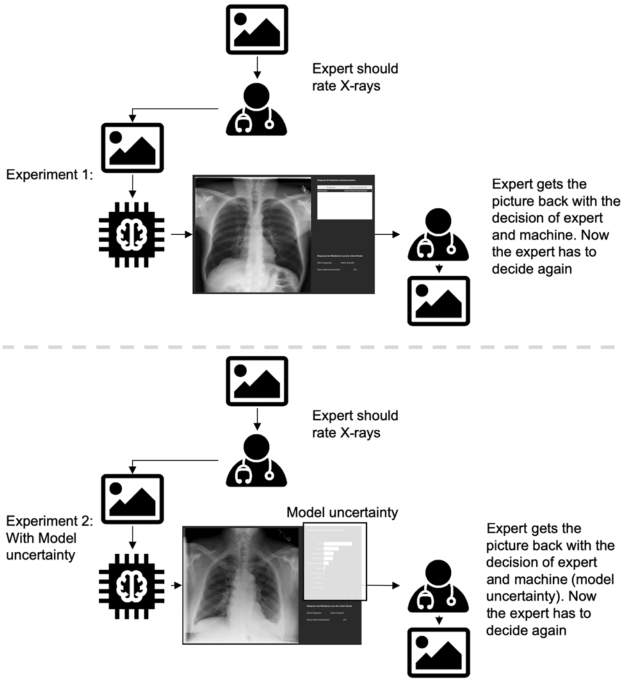
Expert should rate X-rays
Experiment 1:
Expert gets the picture back with the decision of expert and machine. Now the expert has to decide again
Expert should rate X-rays
Experiment 2: With Model uncertainty
Model uncertainty
Expert gets the picture back with the decision of expert and machine (model uncertainty). Now the expert has to decide again

Fig. 1 Experimental setup

Table 1 Results of our model in comparison with previous works bases on the classification accuracy

Diagnosis	Our	[11]	[37]	[39]	[25]
Atelectasis	0.923	0.976	0.716	0.772	0.8094
Cardiomegaly	0.452	0.647	0.807	0.904	0.9248
Effusion	0.586	N\A	0.784	0.859	0.8638
Fibrosis	0.000	N\A	0.769	0.767	0.8047
Infiltration	0.800	N\A	0.609	0.695	0.7345
Mass	0.610	N\A	0.706	0.792	0.8676
Nodule	0.701	N\A	0.671	0.717	0.7802
Pneumothorax	0.610	N\A	0.806	0.841	0.8887
Pleural thick	0.690	0.993	0.708	0.765	0.8062

Table 1 shows that the performance of our trained model is in line with results reported in previous literature. A more detailed analysis of the data shows that our model is able to recognize the different diseases with unequal precision. It is remarkable that Atelectasis is recognized quite well (as with [37]). However, Fibrosis is obviously difficult for the model to recognize, which is consistent with previous research. In line with previous research [11, 37] we agree that it is not advisable to use such ML systems in the current stage in business practice without complementing the results with human expert knowledge. Building truly large-scale, fully-automated medical diagnosis systems with high precision seems to remain a difficult task [37].

Table 2 Contradiction matrix

	Human	
	Correct	Incorrect
ML-based system		
Correct	How often were both, the human user and the system, accurate?	How often was the system accurate, but the human user not? →Contradiction learning, human user should learn
Incorrect	How often was the human user accurate, but the system not? → System should be corrected (IML)	How often were both, the human user and system, inaccurate?

Table 3 Initial Contradiction Matrix in the first experiment

	Human		∑
	Correct	Incorrect	
ML			
Correct	0.10	0.50	0.60
Incorrect	0.05	0.35	0.40
∑	0.15	0.85	

4.2 Quantitative Experiment Results

Nevertheless, the data we gathered from our analyses and observations are sufficient to take the first steps in illuminating, whether humans are able to learn from ML-based DSS. Whenever systems and humans assessing the same material (e.g., in IML), we propose to present a matrix that we call a “Contradiction Matrix” which helps to understand the learning effect that might occur in the particular setting. Table 2 shows the structure of this matrix which is based on the concept of the confusion matrix [31] and is adapted for the application in IML.

For our particular case and in the very first step of the experiment (the ML system and the medical expert evaluate the X-rays independently of each other and without any information on uncertainty) the Contradiction Matrix looks as depicted in Table 3.

Table 3 shows that in 10% of all cases, both, the human user and our system came up with the correct diagnosis at the same time. In 50% of all cases, our system was accurate, but the human user was not. In this case, we would like the human user to learn. But in 5% of the cases, the user was accurate yet our system was not. In this situation, ML systems should be corrected by the means of IML measures. The overall accuracy of the ML system in the first step (Fig. 1) is 60% while the overall accuracy of the human expert is only 15%.

After confronting the human expert with the system’s diagnosis in the second step of the experiment, we arrive at the following results: the human changes his decision in 35% of all cases after being presented the results of the ML system. However, in 20% of the cases, the medical expert wants to tell the system some important features that the system probably did not recognize.

Table 4 Contradiction Matrix in the first experiment after confrontation

	Human		∑
	Correct	Incorrect	
ML			
Correct	0.20	0.40	0.60
Incorrect	0.05	0.35	0.40
∑	0.35	0.75	

Table 5 Initial Contradiction Matrix in second experiment

	Human		∑
	Correct	Incorrect	
ML			
Correct	0.05	0.55	0.60
Incorrect	0.15	0.25	0.40
∑	0.20	0.80	

After correction, the accuracy of the user rises up to 25% (previously 15%). Table 4 provide the details.

In the first experiment, we focused on supporting the decision making process and in the second experiment we present additional information on uncertainty to the medical experts. In the second experiment we selected new X-ray images and now provide information on system uncertainties in the form of the probabilities of the possible diseases to the human expert. Table 5 shows the results of the initial diagnoses.

Table 3 shows that in 5% of all cases the human user and the system provide an accurate diagnosis at the same time. In contrast to that, we observe that in 50% of all cases, our system is accurate, but the human user is not. In these cases, human users may enhance their knowledge by learning from the contradictions. Moreover, in 15% of all cases, the human user is accurate but the system not. In these situation, ML systems should ideally be corrected by the means of IML measures. In 25% of all cases, the human user and

Table 6 Contradiction Matrix in the second experiment after confrontation

	Human		∑
	Correct	Incorrect	
ML			
Correct	0.25	0.35	0.60
Incorrect	0.15	0.25	0.40
∑	0.40	0.60	

the system are inaccurate. The overall accuracy of the ML system in the second experiment is also 60%. The overall accuracy of the human expert in the second experiment is 20%.

After confronting the human expert with the system's diagnoses [now including information on system (un-)certainty, i.e., diagnosis accuracy], the human expert changed his decision in 60% of all cases. However, in 25% of all cases, the medical expert wants to point the system to some important features that the system probably did not recognize. Table 6 shows the results after the confrontation.

After the confrontation with the contradicting information of the ML system, the human accuracy rises up to 40% (previously 20%). Chest X-rays are commonly used as diagnosis measure and therefore X-rays are critical for screening, diagnosis, and management of many life threatening diseases [11]. However, our results from the first experiment show that the evaluation of X-rays by medical experts for lung diseases without additional strong expertise in radiology is not a trivial task. In this regard, the medical expert indicates that our ML system provides an added value for medical experts in terms of diagnostic quality.

The think-aloud protocol reveals for example that it was a big challenge for the medical expert to arrive at a distinctive diagnosis based on a single X-ray image. In the first experiment, where solely the ML-based system decision was presented, the human accuracy improved from 15 to 25%. In the second experiment, with additional information on uncertainty within the system decision, the diagnostic quality improved from 20 to 40%.

4.2.1 Qualitative Experiment Results

In addition to quantitative measures, the think-aloud protocol provides qualitative indications. For example, the physician verbalizes that the system helps to reconsider his diagnosis:

> Expert: "The system makes me aware of something here. I haven't seen this before."

Moreover, our analyses of the think-aloud protocol show that the presentation of ML uncertainty measures generates additional trust, which ultimately results in better diagnoses by the medical expert. For example, the expert says that the system sensitizes to a disease. This enables him to reevaluate the X-ray with these clues.

> Expert: "With the reference to the Atelectasis I have already seen these indicators here, so that I would follow the recommendation."

With regard to the two experiments (with and without information on system uncertainty), the expert recommends the presentation of system uncertainty information which he deems to be very useful.

> Expert: "By presenting the probabilities, I can better evaluate the system decision."

This statement confirms previous literature where medical experts prefer to have better explanations to help them interpret the system's confidence, to verify that the medical condition fits the suggestion, to better understand the decision model, and to make differential diagnoses [6].

The medical expert further sees the limits of existing systems and recommends new applications as the following excerpt from the think-aloud protocol shows:

> Expert: "But you have to consider that in practice AI would be used in such a way that the AI has already suggested a diagnosis to you."

From these statements we conclude that the medical expert has—in general—a positive attitude towards such systems, as long as he retains the power of decision to intervene in critical cases in a regulated manner.

5 Concluding Thoughts

In this article we discuss the counterpart of Interactive Machine Learning, i.e., the learning on the human side. Whenever humans and ML-based systems need to work together, there is the possibility of a joint growth. Each side can stimulate learning on their counterpart. IML offers the possibility to interactively improve the learning process on the machine's side, while contradiction learning can be helpful for learning on the human side.

First, our small study in the medical domain of pneumonology shows that humans and systems tend to make errors and are—at this point of time—far from being perfect. Second, this imperfectness suggests that a human-machine-collaboration resulting in mutual benefit can in theory result in better decisions.

Our experiments show that the human expert is willing to adapt his diagnosis if the system presents a contradictory diagnosis. The willingness to reconsider can however be greatly improved by providing additional information on the uncertainty of the system's diagnosis. This piece of information would also be essential in a discussion between two human experts and could also be seen as a measure to increase the grade of anthropomorphism in the interplay (see [23]).

This article also suggests that in the domain of IML and contradiction learning the presentation of a Contradiction Matrix can be useful to assess the accuracy of human experts and the system, the overlap and the contradiction between these two players. Although our observations mainly stem from the experimental context of medicine, we conclude that departing from the traditional human-in-the-loop paradigm towards a mutual learning approach may be beneficial to humans in many different domains. Our observations may additionally only be seen as first indicators for error-learning with ML-based systems in specific situations. Many questions of different nature arise at this point that warrant further research from various perspectives.

First, knowledge, problems and decisions within different domains may be of different nature, structure and complexity [9]. Therefore, we consider it necessary that the research phenomenon should be examined domain-specifically, to be able to derive dedicated design implications for specific domains, as well as to unravel nuanced yet decisive factors for theory development.

Second, even within a domain, similar problems may vary drastically due to different contextual conditions [9]. For example in medicine, standard procedures in a diagnostic process may quickly change with a fast and drastic deterioration of a patient's condition. Therefore, the examination of the phenomenon of learning in the same decision-making problem in different contexts may already give rise to different, yet complementary findings.

Third, previous research suggests that novices use explanations and advice from intelligent systems differently than experts [10]. Although we have specifically examined the phenomenon with experts in mind, it would be highly interesting to understand potential differences in error-learning between novices and experts and if propositions of prior research hold.

Four, especially critical questions for system design arise in regard of our own biases that lead us to uncritically accept or willingly reject machine made recommendations [14, 20, 24]. Cognitive biases and related factors should thus receive substantial attention during investigations of error-learning with ML-based systems.

While the above mentioned points only constitute a fraction of the potentially interesting research possibilities, we are confident that this discussion can serve as a motivational basis for a further investigation of this important phenomenon. Overall we propose that we need to find ways and technical measures to enable human decision makers to identify and correct erroneous system predictions while also enabling humans to gain alternative and novel insights with the aid of ML-based systems. This process of error learning may be structured as follows: (1) we make a decision based on our own knowledge. (2) We receive a machine prediction that is inconsistent with our own knowledge, but leads to more efficient and accurate decisions. (3) We recognize potential errors in our own decision making. (4) With the help of the contradictory information (and explanations) from the machine, we are finally able to incorporate alternative insights to improve our own (cognitive) skills. Investigating the presented research phenomenon is a challenge that certainly needs to be addressed by truly interdisciplinary research teams which should consist of domain experts (e.g., medicine, management, law), AI experts and potentially additional researchers from social science (like sociology or psychology) and HCI. We believe that the investigation of the research phenomenon of error-learning with ML-based systems will not only lead to an extension of the theory and a better understanding of the phenomenon itself, but will also lay the foundations for a man-machine symbiosis that is truly human-centered [15].

Acknowledgements Open Access funding provided by Projekt DEAL. We thank the University Hospital Frankfurt for the support during the research investigation.

References

1. Amershi S, Cakmak M, Knox WB, Kulesza T (2014) Power to the people: the role of humans in interactive machine learning. Assoc Adv Artif Intell 35(4):105–120
2. Anderson JR, Boyle CF, Reiser BJ (1985) Intelligent tutoring systems. Science 228(4698):456–462
3. Anderson RC, Kulhavy RW, Andre T (1972) Conditions under which feedback facilitates learning from programmed lessons. J Educ Psychol 63(3):186
4. Barnes JM, Underwood BJ (1959) "fate" of first-list associations in transfer theory. J Exp Psychol 58(2):97
5. Bera P, Burton-Jones A, Wand Y (2011) Guidelines for designing visual ontologies to support knowledge identification. MIS Q 34(4):883–908
6. Bussone A, Stumpf S, O'Sullivan D (2015) The role of explanations on trust and reliance in clinical decision support systems. Healthcare informatics (ICHI). In: IEEE International Conference on healthcate informatics. Dallas, pp 160–169

7. Epley N, Gilovich T (2006) The anchoring-and-adjustment heuristic: why the adjustments are insufficient. Psychol Sci 17(4):311–318
8. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S (2017) Dermatologist-level classification of skin cancer with deep neural networks. Nature 542(2):115–127
9. Fournier-Viger P, Nkambou R, Nguifo EM (2010) Building intelligent tutoring systems for ill-defined domains. In: Nkambou R, Bourdeau J, Mizoguchi R (eds) Advances in intelligent tutoring systems. Studies in computational intelligence, vol 308. Springer, Berlin, Heidelberg, pp 81–101
10. Gregor S, Benbasat I (1999) Explanations from intelligent systems: theoretical foundations and implications for practice. MISQ 23(4):497–530
11. Irvin J, Rajpurkar P, Ko M, Yu Y, Ciurea-Ilcus S, Chute C, Marklund H, Haghgoo B, Ball R, Shpanskaya K, Seekins J, Mong DA, Halabi SS, Sandberg JK, Jones R, Larson DB, Langlotz CP, Patel BN, Lungren MP, Ng AY (2019) Chexpert: a large chest radiograph dataset with uncertainty labels and expert comparison. In: Association for the advancement of artificial intelligence, pp 1–9
12. Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. Springer, Berlin, pp 1097–1105
13. Lhyle KG, Kulhavy RW (1987) Feedback processing and error correction. J Educ Psychol 79(3):320
14. hsien Liao S (2002) Problem solving and knowledge inertia. Expert Syst Appl 22(1):21–31
15. Licklider JCR (1960) Man-computer symbiosis. IRE Trans Hum Factors Electron 1:4–11
16. Maedche A, Legner C, Benlian A, Berger B, Gimpel H, Hess T, Hinz O, Morana S, Söllner M (2019) Ai-based digital assistants. Bus Inf Syst Eng 61(4):1–10
17. Marchant G (1989) Analogical reasoning and hypothesis generation in auditing. Account Rev 64(3):500–513
18. Metcalfe J (2017) Learning from errors. Annu Rev Psychol 68(2017):456–489
19. Metcalfe J, Kornell N (2007) Principles of cognitive science in education: the effects of generation, errors, and feedback. Psychon Bul Rev 14(2):225–229
20. Moore DA, Healy PJ (2008) The trouble with overconfidence. Psychol Rev 115(2):502–517
21. Mosier KL, Skitka LJ, Heers S, Burdick M (1997) Automation bias: decision making and performance in high-tech cockpits. Int J Aviat Psychol 8(1):47–63
22. Nijssen EJ, Hillebr B, Vermeulen PA, Kemp RG (2006) Exploring product and service innovation similarities and differences. Int J Res Mark 23(1):241–251
23. Pfeuffer N, Benlian A, Gimpel H, Hinz O (2019) Anthropomorphic information systems. Bus Inf Syst Eng 61(4):523–533
24. Polites GL, Karahanna E (2012) Shackled to the status quo: the inhibiting effects of incumbent system habit, switching costs, and inertia on new system acceptance. MIS Q 36(1):21–42
25. Rajpurkar P, Irvin J, Zhu K, Yang B, Mehta H, Duan T, Ding Daisy, Bagul Aarti, Langlotz C, Shpanskaya K, Lungren MP, Ng AY (2017) Chexnet: radiologist-level pneumonia detection on chest X-rays with deep learning. Comput Vis Pattern Recognit 11(1):1–7
26. Shih PC (2018) Beyond human-in-the-loop: empowering end-users with transparent machine learning. In: Jianlong Z, Fang C (eds) Human and machine learning: visible, explainable, trustworthy and transparent. Springer, Berlin, pp 37–57
27. Sillic M (2019) Critical impact of organizational and individual inertia in explaining non-compliant security behavior in the shadow it context. Comput Secur 80(1):108–119
28. Silver D, Huang A, Maddison CJ, Guez A, Sifre L, van den Driessche G, Schrittwieser J, Antonoglou I, Panneershelvam V, Lanctot M, Dieleman S, Grewe D, Nham J, Kalchbrenner N, Sutskever I, Lillicrap T, Leach M, Kavukcuoglu K, Graepel T, Hassabis D (2016) Mastering the game of go with deep neural networks and tree search. Nature 484(529):484–489
29. Silver D, Schrittwieser J, Simonyan K, Antonoglou I, Huang A, Guez A, Hubert T, Baker L, Lai M, Bolton A, Chen Y, Lillicrap T, Hui F, Sifre L, van den Driessche G, Graepel T, Hassabis D (2017) Mastering the game of go without human knowledge. Nature 550(10):354–359
30. Skinner BF (1965) Science and Human Behavior. 92904. Simon and Schuster, New York
31. Stehman SV (1997) Selecting and interpreting measures of thematic classification accuracy. Remote Sens Environ 62(1):77–89
32. Stevenson H, Stigler JW (1994) Learning gap: why our schools are failing and what we can learn from Japanese and Chinese education. Simon and Schuster, New York
33. Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2015) Rethinking the inception architecture for computer vision. Comput Vis Pattern Recognit 11(12):1–10
34. Todd P, Benbasat I (1987) Process tracing methods in decision support systems research: exploring the black box. MIS Q 11(4):493–512
35. VanLehn K (2011) The relative effectiveness of human tutoring, intelligent tutoring systems, and other tutoring systems. Educ Psychol 46(4):197–221. https://doi.org/10.1080/00461520.2011.611369
36. Vitalari NP (1985) Knowledge as a basis for expertise in systems analysis: an empirical study. MIS Q 9(3):221–241
37. Wang X, Peng Y, Lu L, Lu Z, Bagheri M, Summers RM (2017) Chestx-ray8: hospital-scale chest X-ray database and benchmarks on weakly-supervised classification and localization of common thorax diseases. IEEE CVPR 2017 27(9):2097–2106
38. Yann L, Yoshua B, Geoffrey H (2015) Deep learning. Nature 521(1):436–444
39. Yao L, Poblenz E, Dagunts D, Covington B, Bernard D, Lyman K (2017) Learning to diagnose from scratch by exploiting dependencies among labels. Comput Vis Pattern Recognit 10(1):1–12