https://doi.org/10.20378/irb-49775

Explainability is not Enough: Requirements for Human-AI-Partnership in Complex Socio-Technical Systems

Toni Waefler[1] and Ute Schmid[2]
[1]University of Applied Sciences and Arts Northwestern Switzerland, Olten, Switzerland
[2]Cognitive Systems, University of Bamberg, Germany
toni.waefler@fhnw.ch
ute.schmid@uni-bamberg.de
https://orcid.org/0000-0002-1301-0326

DOI: 10.34190/EAIR.20.007

Abstract: Explainability has been recognized as an important requirement of artificial intelligence (AI) systems. Transparent decision policies and explanations regarding why an AI system comes about a certain decision is a pre-requisite if AI is supposed to support human decision-making or if human-AI collaborative decision-making is envisioned. Human-AI interaction and joint decision-making is required in many real-world domains, where risky decisions have to be made (e.g. medical diagnosis) or complex situations have to be assessed (e.g. states of machines or production processes). However, in this paper we theorize that explainability is necessary but not sufficient. Coming from the point of view of work psychology we argue that for the human part of the human-AI system much more is required than intelligibility. In joint human-AI decision-making a certain role is assigned to the human, which normally encompasses tasks such as (i) verifying AI based decision suggestions, (ii) improving AI systems, (iii) learning from AI systems, and (iv) taking responsibility for the final decision as well as for compliance with legislation and ethical standards. Empowering the human to take this demanding role requires not only human expertise but e.g. also human motivation, which is triggered by a suitable task design. Furthermore, at work humans normally do not take decisions as lonely wolves but in formal and informal cooperation with other humans. Hence, to design effective explainability and to empower for true human-AI collaborative decision-making, embedding human-AI dyads into a socio-technical context is necessary. Coming from theory, this paper presents system design criteria on different levels substantiated by work psychology. The criteria are described and confronted with a use case scenario of AI-supported medical decision making in the context of digital pathology. On this basis, the need for further research is outlined.

Keywords: Companion Technology, Explainable AI, Interactive Learning, Human Factors, Socio-Technical Systems, Motivation

1. Introduction

Artificial intelligence (AI) may be deployed to automate or to informate decision-making (Zuboff, 1988). The former aims at reducing skill and labor requirements. In contrast, the latter aims at increasing the quality of decisions by allowing for a more informed human decision-making. Automation is not the subject of this paper. Effects of automation on humans are described in detail elsewhere (e.g. Bainbridge, 1987; Parasuraman, Mouloua & Molloy, 1996; Grote, Weik & Waefler, 1996; Grote, 1997; Waefler et al, 2003; Sheridan & Parasuraman, 2005; Manzey, 2012). Important negative consequences of automation on humans include: Overstraining humans when required to monitor automated processes, human over-confidence in technology, loss of situation awareness, loss of skills and experience, as well as demotivation as a result of automation.

However, in this paper we advocate a deployment of AI that informates decision-making. Thus, decision-quality is increased by an interaction of AI and the human, elsewhere referred to as augmented intelligence or augmented cognition (Crowe, LaPierre & Kebritchi, 2017; Kirste, 2019). Humans should be empowered to make more precise decisions (Scherk, Pöchhacker-Tröscher & Wagner, 2017). This includes AI support regarding the humans' susceptibility to errors (Both & Weber, 2014). Designing and implementing AI for joint human-AI decision-making is the overall objective of this approach.

In joint human-AI decision-making a specific role is assigned to the human. For example, Samek, Wiegand and Müller (2017) suggest four major tasks for the human: (i) verification of AI based decision suggestions; (ii) improvement of AI systems, which involves the identification of biases in AI based decision suggestions that might emerge from biases in the data set used for training and/or from deficient decision models; (iii) learning from AI systems, that is, the human improves his or her knowledge by interacting with the system; and (iv) taking responsibility for the final decision as well as for compliance with legislation and ethical standards. All four tasks require human understanding of both, the subject matter as well as the decision-making process. In line with this, the machine learning pioneer Donald Michie already characterized machine learning systems supporting human learning as 'ultra strong' (Muggleton et al, 2018).

For humans, to understand and learn requires explainable AI. This refers to the language in which humans and AI communicate as well as to the communication's object. Both should contribute to the AI's explainability by making AI based decisions transparent to the human (Hager et al, 2017; Samek, Wiegand & Müller, 2017; Crandall et al., 2018). Some authors go even further by postulating that AI should provide information proactively even if the human is not (yet) looking for it (Ittermann et al., 2016). This may also include that the machine detects the human's intentions and needs (Heim, 2011; Ludwig, 2015).

Explainability is especially required for sub-symbolic approaches of machine learning (Adadi & Berrada, 2018), where decision-making models emerge and refine in neural networks by training or in reinforcement learning by operant conditioning (Mitchell, 2019). These algorithms adjust a large number of parameters in a self-learning manner, which is no longer transparent for humans. Thus, machine learning acquires sort of a tacit knowledge that is difficult to communicate to humans. As a result, AI-based decisions are non-transparent for humans and accordingly cannot be traced. Consequently, machine learning systems become a black box for humans.

However, in this paper we strongly emphasize that explainability is a necessary but not sufficient precondition for joint human-AI decision-making. We consider joint human-AI decision-making a process, occurring within a socio-technical system, where the AI forms the technical subsystem and humans form the social subsystem (e.g. Hollnagel & Woods, 2011). Furthermore, on the human side of the socio-technical system there is normally not a single human, i.e. an individual but rather many humans, i.e. a team, a department or even a hole organization. Regarding this socio-technical system, explainability is a design requirement that needs to be considered when engineering the human-AI interface. This is a necessary precondition for joint human-AI decision-making. But it is not enough. For humans to really take an active role in joint decision-making, more levels of socio-technical design requirements need to be considered (Waefler et al, 2003). This is because humans are human beings, which encompasses much more than being a cognitive information processor. Hence, providing information is not enough. Rather, socio-technical system design needs to make sure that humans actively aim for high quality decisions, engage in continuous improvement of decisions and take responsibility. If such aspects are not considered systematically when designing and implementing the human-AI system, it is quite likely that the humans will not play the role expected by the system designers and as a consequence the system will not perform according to the designers’ intentions (e.g. Ulich, 2011).

This is a theoretical paper. Its main purpose is to reflect on system design requirements for joint human-AI decision-making from the perspective of work psychology. The aim is to identify research topics regarding true human-AI partnership that go beyond explainability, considering insights from the tradition of socio-technical system design. To do so, two domains of system design are discussed: (i) design of the human-AI system on the one side, and (ii) design of the socio-technical integration of the human-AI system into a work organization on the other side. These two domains are explored more deeply in the following sections. Afterwards, digital pathology is presented as a use case. We will discuss the requirements of this application based on the introduced principles for system design and propose measures that need to be taken into account when introducing that kind of systems. Although we present a case from the medical domain, we strongly assume that similar aspects need to be considered when applying AI to other domains where human experts need to take critical decisions, e.g. production plants, energy systems, or rail transport. However, we conclude with the insight that in order to apply human-AI systems successfully, focusing on the design of human-AI interaction is not sufficient. Rather, a broader view on system integration needs to be adopted. Appropriate design solutions are still to be developed.

2. Design of the human-AI system

Combining (new) technologies with humans always transforms the humans' tasks and with it the conditions under which the humans perform and learn. This can cause several problems. Bainbridge (1983) described the "ironies of automation" which emerge when automation is used to replace humans where the automation outperforms them. However, often the task remaining with the humans is to supervise the automated processes. The irony is that the humans are expected to supervise a performance, by which they are outperformed. Thus, this task is beyond their capabilities and hence unaccomplishable.

Furthermore, the human is expected to intervene when the automation fails. Often this is another unaccomplishable task due to deskilling effects of automation (Manzey, 2012). Since humans learn skills by doing, they will lose those skills related to activities that are replaced by automation. Imagine a car driver who is expected to supervise an autonomously driving car. Where does he or she train the driving skills, such as the estimation of a breaking distance or the estimation whether or not the car's speed is soundly adapted to the weather conditions, when he or she never drives manually? Skills like these will disappear and in turn the human driver will not be able to supervise the autonomous driving. When designing human-machine systems it is therefore important to consider the competencies the humans need to perform the role they are expected to take. Many, if not most of these competencies – especially regarding human expertise – base on tacit knowledge, which is acquired by experience, i.e. by doing. This refers to both aspects of expertise, the know-how as well as the know-why. In addition to the risk of loss of relevant expertise, the implementation of new technologies normally also brings about the need of new competencies (Parasuraman, Mouloua & Molloy, 1996). As a consequence, joint human-AI decision making requires not less, but even more human expertise.

When designing human-machine interaction, different levels of interfaces need to be taken into consideration. Though usability and user experience are important features of human-machine interaction, they are not at the core of the problems sketched above. These problems refer rather to human-machine function allocation, which determines the human's task when cooperating with the machine, and hence the human's opportunities for learning and developing expertise (Grote, Weik & Waefler, 1996). In the car driver example mentioned above, a good usability cannot compensate for lacking skills. Rather function allocation needs to make sure, that the human has the opportunity to acquire task-relevant expertise. This is not reached by the human interacting with the interface, but rather by interacting through the interface with the process (Hollnagel & Woods, 2005).

Current concepts regarding human-machine function allocation consider humans and machines as complementary (Waefler et al, 2003). The basic assumption is, that humans and machines are not comparable in a quantitative way. Rather they differ qualitatively from each other. What is an easy task for machines (e.g. playing chess, computing a huge amount of data) is very demanding for humans, and vice versa (e.g. loading the dishwasher, acting in a unstructured environment). The aim of complementary system design is to combine humans and machines in a way allowing for mutual fostering of strengths as well as for mutual compensation of weaknesses. Current concepts assume that the coping with ill-defined problems is a human strength, whereas the machine is good in handling well-defined problems. With this background, criteria for assessment and design of human-machine function allocation where developed that allow for human control over automated processes. These criteria encompass aspects such as process transparency, human-machine coupling, authority over information, authority over process control, and flexibility of function allocation (e.g. Waefler et al, 2003; see Fig. 1).

With the emergence of AI systems in real world domains, corresponding concepts and criteria need to be developed further. This is because AI systems show capabilities formerly considered exclusive human. Probably most prominent is the capability of machine learning systems to recognize patterns in unstructured data and hence to handle ill-defined problems (Muggleton et al, 2018). Whether humans and AI are still complementary is a question that consequently needs to be rethought. Of course, there is already evidence of human-AI complementarity. Mitchel (2019) describes shortcomings of machine learning approaches like overfitting to and biases in training data. In general, in most AI systems perception is realized very different from the way humans perceive. Most prominent is that AI systems do not have background knowledge and therefore cannot understand objects in their context.

Humans on the other hand are also prone to cognitive biases (e.g. Kahneman, 2011). In contrast to AI however, cognitive failures of humans often emerge from context information that biases human perception and decision-making (e.g. confirmation bias). Consequently, one aspect of human-AI complementarity is the AI's capability of digging deep into the data whereas the human capability is to embed perception into broad background knowledge allowing for understanding and interpreting (Brynjolfsson & McAfee, 2014). Or as Floridi (2014) states: Computers are good at computing and humans are good at thinking, and these are different capabilities. With this background, human-AI system design does not only need to make sure that the human does not lose his capability to think. Even more, the humans' ability to think – that is e.g., the humans' ability to understand – needs to be fostered by the way the humans cooperate with the AI systems – specifically with machine learning systems, which can detect regularities in highly complex information (Muggleton et al, 2018).

However, within the human-AI system the human is not a cognitive resource only. Disposing of respective human expertise is a prerequisite only for the human to take the intended role. There is yet another aspect of human-AI complementarity required, which refers to genuine human attributes such as motivation and responsibility taking. Whether or not such human qualities emerge is not independent from socio-technical system design and hence from the way a human-AI system is integrated into a work organization. The next section will reflect on this topic.

3. Design of the socio-technical integration of the human-AI system

As it is well known in work psychology, motivation and responsibility taking requires a suitable job design (e.g. Ulich, 2011). Whether or not a person is motivated, is only partly due to his or her personality. Many context factors do affect motivation too, extrinsically as well as intrinsically. Especially context factors regarding intrinsic motivation are set by the concrete design of human-machine systems. Hackman and Oldham (1980) with their 'Job Characteristic Model' identified three critical psychological states required for intrinsic motivation: (i) knowledge of the actual results of the work activities, (ii) experienced responsibility for outcomes of the work, and (iii) experienced meaningfulness of the work. For motivation and responsibility taking it is crucial that humans really do experience these states. It is not sufficient to impose responsibility to the humans by job description. Rather it is important that the humans feel a sense of responsibility. This feeling is triggered, among other aspects, by autonomy. This is because humans do not feel responsible for decisions taken elsewhere, be it by another human or by an AI.

To allow for the three critical psychological states, Hackman and Oldham (1980) identified five core dimensions of job design: (i) skill variety, (ii) task identity, (iii) task significance, (iv) autonomy, and v) feedback from the task. Job design criteria like these are shaped when a human-AI decision-making system is created and when it is implemented into organizational processes. If this shaping is not deliberately designed but emerges randomly as a by-product of technology engineering, it is likely, that motivational preconditions are not set adequately. Imagine a human is expected to decide jointly with an AI system. If the human is not experiencing autonomy in the decision-making process, it will be likely that he or she just pitches under the AI-generated decision suggestion. Even though the system designers intended to assign the human an active role in joint human-AI decision-making, he or she will not live up to it. And he or she will likely not feel responsible even if responsibility is formally assigned to him or her. However, what human autonomy really means in joint human-AI decision-making is yet to be better understood. The same applies to the other core dimensions of job design.

With this background different methods for job design provide design criteria regarding the integration of humans, technology and organization on the level of individual tasks as well as on the level of organizational design. As a representative of such methods the KOMPASS criteria (see Fig. 1; Waefler et al, 2003) are mentioned in the following. On the individual task level these are: Task completeness, decision-making requirements, communication requirements, opportunities for learning, variety, transparency of work flow, influence over work conditions, and temporal flexibility. On the level of organizational design, these criteria are: Complete task, independence of the work system, fit between regulation requirements and opportunities, polyvalence of work system members, autonomy of work group, and boundary regulation by the supervisor.

When engineering the AI part of joint human-AI decision-making systems, criteria like these are not directly applicable. However, this does not mean that they are unimportant. Neglecting them results in an underperforming system since the preconditions for the human to play his or her intended role in joint human-AI decision-making are not met.

As a consequence, concepts for integrating AI into task design on both levels are yet to be elaborated, i.e. on the level of individual work tasks, as well as on the level of integrating AI systems into organizational processes. We will illustrate this for a scenario of medical decision making. The presented analysis can be applied likewise for other risky decision-making, e.g. in industrial production, where deciding whether a machine is in a safe state can be critical when an unnecessary stop results in financial loss, but omitting to stop the machine can result in serious injuries of workers and damage of goods.

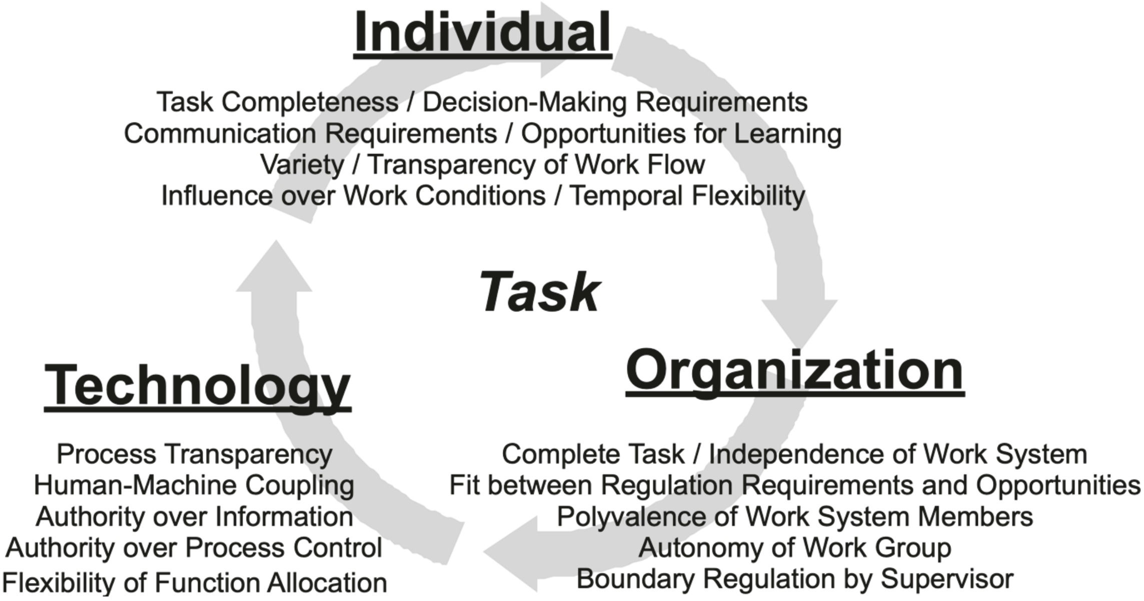

Individual
Task Completeness / Decision-Making Requirements
Communication Requirements / Opportunities for Learning
Variety / Transparency of Work Flow
Influence over Work Conditions / Temporal Flexibility
Task
Technology
Process Transparency
Human-Machine Coupling
Authority over Information
Authority over Process Control
Flexibility of Function Allocation
Organization
Complete Task / Independence of Work System
Fit between Regulation Requirements and Opportunities
Polyvalence of Work System Members
Autonomy of Work Group
Boundary Regulation by Supervisor

Figure 1: The Individual-Technology-Organization cycle (Waefler et al, 2003)

4. Use case: Human-AI-partnership for expert medical decision-making

Reliable diagnosis of cancer at an as early stage as possible is one of the most challenging domains of medical diagnosis. Missing a malign tumor has severe consequences and erroneous classification of the type of a tumor can lead to inefficient treatment – possibly with undesired side-effects for the patient. Furthermore, false alarms bring unnecessary anxiety to patients and can result in superfluous surgical interventions. Many types of cancer are diagnosed based on imaging techniques such as radiography, computed tomography of specific body parts or microscopy of tissue samples (Najarian & Splinter, 2005). Experts need long training times to be able to interpret such medical images. To analyze image-based data, human experts often take into account spatial information. In colon cancer diagnosis, medical experts analyze the tissue composition and the depth of invasion of tumors. For instance, if tumor tissue already touches fat, the tumor class is more critical compared to a situation where the tumor is included in fascial tissue (Wittekind, 2016). In general, there are five classes of tumors, ordered by severeness, labelled as pT0 (healthy), pT1, . . . pT4.

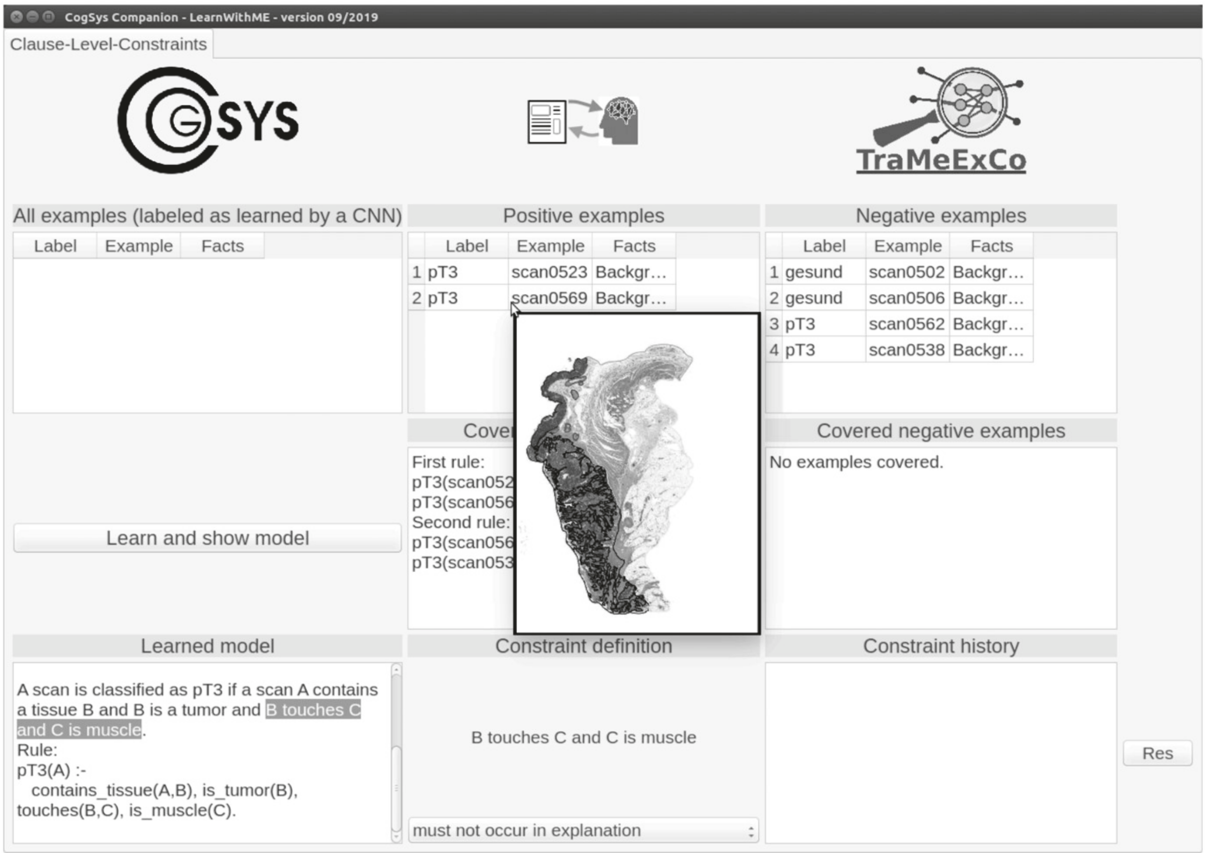
Digital pathology provides image analysis techniques to support medical experts when analyzing tissue samples. In an ongoing research project, we design a machine learning based assistance system to support medical decision-making (Schmid & Finzel, 2020). The TraMeExCo (transparent medical companion) system combines convolutional neural networks (CNNs) and inductive logic programming (ILP). CNNs show impressive results for image classification (Krizhevsky, Sutskever & Hinton, 2012; Li et al, 2014). CNNs allow end-to-end learning from raw data – here bitmaps – to class, thereby making unnecessary a preprocessing step to extract features. However, CNNs have the usual problems of data intensive deep learning approaches: To train a CNN, huge amounts of pre-labelled data are prerequisite. And – as mostly is the case in the medical domain – there is no real ground truth to label the data. At best, there are diagnostic measures with high reliability (so called gold standard). Furthermore, a trained CNN is a black box giving the human only the class decision but no indication on how it came about this decision (Adadi & Berrada, 2018).

In contrast, ILP approaches (Muggleton & De Raedt, 1994) can be trained with small sets of data (Gulwani et al, 2015). They belong to the class of interpretable machine learning approaches together with decision trees and related approaches (Furnkranz & Kliegr, 2015). Learned models are white box – i.e., represented in a symbolic, human readable, explicit form (Doshi-Velez & Kim, 2017). It has been shown that rules learned with ILP can support human decision-making in complex domains (Muggleton et al, 2018). Transforming such rules into verbal explanations can be done with similar methods as have been introduced in the context of expert systems (Clancey, 1983; Siebers & Schmid, 2019).

To provide explainability and hence to support joint human-AI decision-making, normally an explanation interface is provided. Typically, explanations for CNNs are given visually – marking that regions in the image, which had the most influence on the classification (Samek et al, 2016; Ribeiro, Singh & Guestrin, 2016). However, to explain a tumor class based on a tissue scan, human experts rely on spatial relations as pointed out above. A visual explanation can only highlight information which is present in an image. It cannot convey special relations between different components (e.g. a tumor tissue that touches fat). However, ILP learned models can generate such relational explanations (Rabold et al, 2019).

For the TraMeExCo system we enriched the ILP learned model with a background theory for spatial relations (Schmid & Finzel, 2020). Furthermore, to take into account that the initial class labels with which TraMeExCo has been trained contain erroneous labels and noise, our learning approach is human-AI interactive (Ware et al, 2001), i.e. TraMeExCo allows the human expert to modify current explanations in addition to class corrections (for details see Finzel, 2019).

In Figure 2 we present our human-AI mutual explanation interface. In the upper part, a selection of tissue scans is presented which have been classified – e.g. by a CNN classifier. Four scans have been classified as tumor class pT3 and the ILP learner induced a model characterizing these scans in contrast to two scans classified as healthy ('gesund'). A human expert inspects the learned rules given in the bottom of the interface and discovers that one of the rules contains an erroneous 'touches relation'. He or she marks the erroneous part and can define the constraint that this part should be excluded from future models (see bottom middle of the interface). The model is updated and as a result, scans previously classified as pT3 are now moved to the negative examples (see top right of the interface). The expert can inspect these scans and either change their label or modify the rules again.



All examples (labeled as learned by a CNN)

Label	Example	Facts

Positive examples

	Label	Example	Facts
1	pT3	scan0523	Backgr…
2	pT3	scan0569	Backgr…

Negative examples

	Label	Example	Facts
1	gesund	scan0502	Backgr…
2	gesund	scan0506	Backgr…
3	pT3	scan0562	Backgr…
4	pT3	scan0538	Backgr…

Cove

First rule:
pT3(scan052
pT3(scan056
Second rule:
pT3(scan056
pT3(scan053

Covered negative examples

No examples covered.

Learned model

A scan is classified as pT3 if a scan A contains a tissue B and B is a tumor and B touches C and C is muscle.
Rule:
pT3(A) :-
contains_tissue(A,B), is_tumor(B), touches(B,C), is_muscle(C).

Constraint definition

B touches C and C is muscle

must not occur in explanation

Constraint history

Figure 2: User interface for joint human AI decision-making in digital pathology

On the top, digital image data first is presented as classified by a CNN. Data is referred to as 'scan' followed by a number. Class labels are 'gesund' (healthy) or a tumor class name such as pT3. The medical expert can sort a sample of data into two groups: positive examples for a given class (here pT3) and negative examples. In the top right area, it is shown that the expert contradicts the classification decision of the CNN and classifies scan562 and scan538 as negative instances, which should not be classified as pT3. The digital image is shown by a mouse-over. Furthermore, its representation in symbolic form can be inspected. This information is represented in the programming language and consists of a set of facts describing an image, such as 'contains_tissue' (scan0523,t1), 'is_tumor(t1)' (scan0523). The middle part of the interface offers a button to start learning a new set of rules on the basis of the corrections made by the human expert. In the shown episode, the learned model covers all shown instances correctly, that is, none of the examples classified as negative by the expert are classified as pT3. At the bottom of the interface, an example of the learned rules is described in natural language and in Prolog. This is the explanation why the system classified the examples as belonging to tumor class pT3. The human expert can inspect the explanations and select parts of the explanations for which he or she can determine that this part must or nor must not occur in the classification rule. The information in Prolog is not intended for the domain expert but for the developer.

5. Discussion of the use case

As described above, optimal integration of individuals, technology and organization requires consideration of two domains of system design: Design of the human-AI system, and design of the socio-technical integration of the human-AI system into a work organization.

Regarding the human-AI system, the mutual explanation interface of the TraMeExCo system meets corresponding design criteria (cf. figure 1) quite well. It provides process transparency as it makes AI-based decision-making criteria transparent. It allows for human authority over information, as the human has the possibility to influence what information he or she gets from the AI. And it allows for authority over process control as the human can shape the AI's decision model and hence the AI's decision-making process. However, more research will provide further possibilities to meet these criteria even better.

Other criteria regarding the human-AI system are not yet reflected in the design of the TraMeExCo interface. This refers especially to human-machine coupling and to flexibility of function allocation. The former concerns whether or not the machine determines the way of task execution. The latter aims to enable switching between different degrees of distributing process control to the human and the machine. The aim of both criteria is to allow the human to personalize the way of task execution. From a work psychological point of view this is important regarding aspects like supporting different levels of human expertise or allowing for the experience of self-efficacy – just two examples for setting the preconditions regarding active human role taking. However, to develop design solutions regarding such aspects, a better understanding of the human decision-making process and especially of different decision-making styles is required. This may lead to an AI system design allowing the individual to adapt the system to his or her personal work style.

The second domain of system design refers to the socio-technical integration of the human-AI system into a work organization. As outlined above, in this domain two levels need to be distinguished: Individual task design and organizational design (cf. figure 1). At first glance, these two levels are not directly affected by the design of the TraMeExCo system's mutual explanation interface. Nevertheless, both levels are important to prepare the human for active role taking in joint human-AI decision making. On the individual level, the human may develop competencies required to classify digital images when performing sub-tasks with no AI interaction. As a consequence, task completeness for example may be critical for the availability of required human expertise.

The level of organizational design refers for instance to the way the joint human-AI decision making is integrated into organizational decision-making processes. This is important because different people with different expertise may increase their collective decision-making quality by creating human-human synergy. Polyvalence of work system members for example might therefore be important to promote mutual understanding and hence the ability to cooperate. Although aspects like these are not directly dependent of human-AI system design, they are nevertheless crucial for human-AI system success. Since we do not yet have a sufficient understanding of how the AI-system design shapes such aspects, further research is required to find new design solutions.

6. Conclusion

This is a theoretical paper. The core assumptions of our reflections are that AI systems are currently not mature enough to provide high quality decisions autonomously. Therefore, humans need to take an active role in the decision-making process. Furthermore, we assume that humans will keep an active role although AI systems may become more mature in future. We take this assumption, because we adopt a complementary approach, that considers humans and technology as qualitatively different. Hence, they are not competing for which of the two is better regarding capabilities that are comparable in a quantitative way. Rather, they complement each other with qualitatively different capabilities. As Floridi (2014) states, computers are good in computing, humans are good in thinking, and this is not the same. Consequently, a clever human-technology combination will always perform better in comparison to what the human or the technology could deliver each on its own. Regarding AI systems we strongly assume that this is especially true, where decision-making requires understanding and responsibility taking, both major human characteristics.

However, in order for a joint human-AI system to effectively perform it is not enough to assign the human an active role in the decision-making process. Rather, the human needs those preconditions required for high level of engagement at work. The reason for this is that being active, showing commitment or taking responsibility is not something that can be forced just by order or by assignment. That kind of human contributions must be intrinsically motivated by an adequate job design. To do so, work psychology provides job design criteria on three levels of socio-technical system design: Human-machine function allocation, job design on the individual level, and integration into organizational processes (cf. figure 1).

Applying this approach to our case of AI supported medical decision-making shows, that on the level of the human-machine function allocation, explainable AI addresses some of the relevant criteria. If explainability is mutual, that is, if the AI does not only make decisions transparent to the human but the human can also influence the AI, these design criteria are met even better. However, much more research is required to reach true complementarity in the interaction of humans an AI in joint decision-making. On the other two levels of socio-technical system design – i.e. job design on the individual level, and integration into organizational processes – there is still much less knowledge available regarding how the introduction of AI impacts the corresponding criteria. Though these two levels are not directly related to AI system design, they are important for the success of joint human-AI decision-making systems. This is because work design on these levels is crucial for human behavior at work and if the work design is bad, humans will not take the active role expected from them when cooperating with AI. Much more research is required on these two levels to better understand how they are affected by the introduction of AI.

Acknowledgements

We thank Bettina Finzel for providing the screenshot of the Digital Pathology Interactive Interface. The work presented in this paper is partially supported by the BMBF ML- 3 grant FKZ 01IS18056 B – project Transparent Medical Expert Companion (TraMeExCo).

References

Adadi, A. and Berrada, M. (2018) “Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)”, *IEEE Access*, Vol. 6, pp 52 138–52 160.

Bainbridge, L. (1983) „Ironies of automation“, *Automatica*, Vol 19, No. 6, pp 775-779.

Both, G. & Weber, J. (2014) *Hands-Free Driving? Automatisiertes Fahren und Mensch-Maschine Interaktion.* In E. Hilgendorf (Hrsg.) Robotik im Kontext von Moral und Recht, Nomos, Baden-Baden S. 171-188.

Brynjolfsson, E. and McAfee, A. (2014) *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*, WWNorton & Company, New York.

Clancey, W.J. (1983) “The epistemology of a rule-based expert systema framework for explanation”, *Artificial Intelligence*, Vol. 20, No. 3, pp 215–251.

Crandall, J. W., Oudah, M., Chenlinangjia, T., Ishowo-Oloko, F., Abdallah, S., Bonnefon, J.-F., Cebrian, M., Shariff, A., Goodrich, M. A. & Rahwan, I. (2018) "Cooperation with machines", *Nature Communications*, Vol. 9, online, www.nature.com/articles/s41467-017-02597-8.

Crowe, D., LaPierre, M. & Kebritchi, M. (2017) "Knowledge Based Artificial Augmentation Intelligence Technology: Next Step in Academic Instructional Tools of Distance Learning", *TechTrends*, Vol. 61, pp 494-506

Doshi-Velez, F. and Kim, B. (2017) “Towards a rigorous science of interpretable machine learning”, *arXiv preprint arXiv:1702.08608*.

Finzel, B. (2019) “Explanation-guided constraint generation for an inverse entailment algorithm,” Master’s Thesis, University of Bamberg.

Furnkranz, J. and Kliegr, T. (2015) “A brief overview of rule learning,” paper read at International Symposium on Rules and Rule Markup Languages for the Semantic Web, Berlin, August.

Grote, G. (1997) *Autonomie und Kontrolle*. vdf Hochschulverlag, Zürich.

Grote, G., Weik, S. & Waefler, T. (1996) *KOMPASS: Complementary allocation of production tasks in sociotechnical systems.* In S. A. Robertson (Ed.) Contemporary Ergonomics 1996, Taylor & Francis, London, pp. 306-311.

Gulwani, S., Hernandez-Orallo, J., Kitzelmann, E., Muggleton, S., Schmid, U. and Zorn, B. (2015) “Inductive programming meets the real world”, *Communications of the ACM*, Vol. 58, No. 11, pp 90–99.

Hackman, J.R. & Oldham, G.R. (1980) *Work redesign*, Addison-Wesley, Reading MA.

Hager, G. D., Bryant, R., Horvitz, E., Matarić, M. & Honavar, V. (2017) *Advances in Artificial Intelligence Require Progress Across all of Computer Scienc*e, Computing Community Consortium Catalyst, Washington, D.C.

Heim, P. (2011) *Interaktive Angleichung als Modell für die Mensch-Computer-Interaktion im Semantic Web*. Unveröffentlichte Dissertation, Universität Stuttgart

Hollnagel, E. and Woods, D. D. (2005) *Joint cognitive systems: Foundations of cognitive systems engineering,* CRC Press, Boca Raton, FL.

Ittermann, P., Niehaus, J., Hirsch-Kreinsen, H., Dregger, J. & ten Hompel, M. (2016) *Social Manufacturing and Logistics. Gestaltung von Arbeit in der digitalen Produktion und Logistik*. Technische Universität, Dortmund

Kahneman, D. (2011) *Schnelles Denken, langsames Denken,* Random House, München.

Kirste, M. (2019) *Augmented Intelligence - Wie Menschen mit KI zusammen arbeiten*. In: V. Wittpahl (Hrsg.) Künstliche Intelligenz, Technologien, Anwendung, Gesellschaft, Springer, Berlin.

Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012) “ImageNet classification with deep convolutional neural networks”, *Advances in Neural Information Processing Systems*, Vol. 25, pp 1097–1105.

Floridi, L. (2014) *The fourth revolution: How the infosphere is reshaping human reality,* Oxford University Press, Oxford.

Li, Q., Cai, W., Wang, X., Zhou, Y., Feng, D.D. and Chen, M. (2014) “Medical image classification with convolutional neural network", paper read at 13th International Conference on Control Automation Robotics & Vision, Singapore, December.

Ludwig, B. (2015) *Planbasierte Mensch-Maschine-Interaktion in multimodalen Assistenzsystemen,* Springer, Berlin.

Manzey, D. (2012) Systemgestaltung und Automatisierung. In P. Badke-Schaub, G. Hofinger & K. Lauche (Hrsg.), Human Factors. Psychologie sicheren Handelns in Risikobranchen, pp 333-352, Berlin: Springer.

Mitchell, M. (2019) “Artificial intelligence hits the barrier of meaning”, *Information*, Vol. 10, No. 2, p. 51-53.

Muggleton, S. and De Raedt, L. (1994) “Inductive logic programming: Theory and methods”, *The Journal of Logic Programming*, Vol. 19, pp 629–679.

Muggleton, S., Schmid, U., Zeller, C., Tamaddoni-Nezhad, A. and Besold, T. (2018) “Ultra-strong machine learning: comprehensibility of programs learned with ilp”, *Machine Learning*, Vol. 107, No. 7, pp. 1119– 1140.

Najarian, K. and Splinter, R. (2005) *Biomedical signal and image processing*, CRC press, Boca Raton FL.

Parasuraman, R., Mouloua, M. and Molloy, R. (1996) "Effects of Adaptive Task Allocation on Monitoring of Automated Systems", *Human Factors,* Vol. 38, pp 665-679.

Rabold, J., Deininger, H., Siebers, M. and Schmid U. (2019) “Enriching visual with verbal explanations for relational concepts – Combining LIME with Aleph,” paper read at *Advances in Interpretable Machine Learning and Artificial Intelligence Workshop (AIMLAI), Würzburg, June*.

Ribeiro, M.T., Singh, S. and Guestrin, C. (2016) “Why should i trust you?: Explaining the predictions of any classifier,” paper read at 22nd ACM SIGKDD international conference on knowledge discovery and data mining, San Francisco, August.

Samek, W. Wiegand, T. Müller K.R. (2017) "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models", *ITU Journal: ICT Discoveries, Special Issue, The Impact of AI on Communication Networks and Services*, No. 1, pp 1–10.

Samek, W., Binder, A., Montavon, G., Lapuschkin, S. and Müller, K.-R. (2016) “Evaluating the visualization of what a deep neural network has learned,” *IEEE Transactions on Neural Networks and Learning Systems*, Vol. 28, No. 11, pp 2660–2673.

Scherk, J., Pöchhacker-Tröscher, G. & Wagner, K. (2017) *Künstliche Intelligenz – Artificial Intelligence,* Pöschhacker Innovation Consulting, Linz.

Schmid, U. & Finzel, B. (2020) "Mutual Explanations for Cooperative Decision Making", *Medicine. KI - Künstliche Intelligenz*. https://doi.org/10.1007/s13218-020-00633-2

Sheridan, T. B. & Parasuraman, R. (2005) "Human-Automation Interaction", *Reviews of Human Factors and Ergonomics*, 1, 89-129.

Siebers, M. and Schmid, U. (2019) “Please delete that! Why should I? – Explaining learned irrelevance classifications of digital objects,” *KI- Künstliche Intelligenz*, Vol. 33, No. 1, pp 35–44.

Ulich, E. (2011) *Arbeitspsychologie*, vdf Hochschulverlag, Zürich.

Waefler, T., Grote, G., Windischer, A. and Ryser, C. (2003) KOMPASS: A Method for Complementary System Design. In: E. Hollnagel (Ed.) Handbook of Cognitive Task Design. Lawrence Erlbaum, Mahwah, NJ, pp 477-502.

Ware, M., Frank, E., Holmes, G., Hall, M. and Witten, I.H. (2001) “Interactive machine learning: letting users build classifiers”, *International Journal of Human-Computer Studies*, Vol. 55, No. 3, pp 281–292.

Wittekind, C. (2016) TNM: Klassifikation maligner Tumoren, John Wiley & Sons, Weinheim.

Zuboff, S. (1988) *In the age of the smart machine*, Basic Books, New York.