White Paper

Human-centric Machine Learning

A Human-Machine Collaboration Perspective

fortiss

Human-centric Machine Learning –
A Human-Machine Collaboration Perspective

*Machine Learning Lab * | Human-centered Engineering **

Authors

Dr. Yuanting Liu

fortiss GmbH,
Guerickestr. 25
80805 Munich

Dr. Hao Shen

fortiss GmbH,
Guerickestr. 25
80805 Munich

kontakt@fortiss.org

* Equal contribution
fortiss GmbH. Correspondence to

Machine Learning Lab (MLL): mll@fortiss.org
Human-centered Engineering (HCE): hce@fortiss.org

Content

Abstract

Engineering of *reliable, safe* and *trustworthy* technical systems has entered a new era with recent advances in Machine Learning (ML) technology. Such technological progress provides engineers of this generation with tools of great potential, but also significant new challenges. Advanced automation as enabled by ML has to respect social norms and ethical standards. High levels of human control to safeguard such technical systems is therefore of paramount importance, but difficult to achieve. Thus, it is critical to place the focus of future research on a way to engineer ML systems which enables smooth collaboration between humans and machines. In this white paper, we outline a high-level framework for the human-centric engineering of ML systems. We derive a number of general research topics and more specific research questions from this framework, and describe how the joint effort of two fortiss competence fields can lead to an innovative approach for the engineering of ML-driven software systems.

Motivation

Software and systems engineering generally strive to empower humans in their ability to carry out their tasks. These systems have to be reliable, safe and trustworthy in order to fulfil their purpose. As simple as this statement is, engineering of such systems has turned out to be extremely challenging and complex, which already has led to an abundance of research for traditional systems that predictably follow well-understood algorithms. We are currently experiencing the introduction of a new wave of powerful technology, based on algorithms which can learn from given data and adapt themselves to optimize their solutions, i.e., Machine Learning (ML). Specifically, remarkable achievements have been made in well-defined domains where huge amounts of data are available with pre-defined interpretation of the data, obtained by supervision and labeling. However, the apparent success in producing seemingly intelligent decisions brings along a number of dangerous causes for misunderstandings in the communication between humans and machines. If we compare the behavior of ML systems and humans in decision making, significant differences are obvious. ML essentially provides efficient algorithmic solutions for optimizing a well-defined target function, enabling the learning of task- and data-specific patterns from a huge amount of samples or observations. In contrast, a human would rather make decisions based on ground-truth rules like causality and can transfer known solutions to new situations and domains. Although both types of decision making can be called *forms of generalization*, the human way of decision making is a harder form of generalization, sometimes termed *horizontal, strong,* or *out-of-distribution generalization*. Human decision making takes advantage of heterogeneous information sources such as interventions, domain shifts and temporal structures, which ML typically discards or even fails to model in learning processes. This shortcoming leads to a number of challenges in designing reliable, safe and trustworthy systems based on ML for human users:

Low explainability
The decision-making mechanism used by an ML system cannot be made fully transparent to humans due to their nature of learning. Consequently, behaviors of such systems become difficult to interpret for humans. Although there have been some recent efforts in providing some explanation to humans, such as in the area of image recognition, these approaches are far from being generally applicable, and even worse so when it comes to respecting human norm and value systems.

Low robustness against perturbed input data
The performance of intelligent ML-based systems depends heavily on the quality of the input data. Even small manipulations of input data, such as pixel-by-pixel perturbations, can lead to serious disturbances of the system outputs. The poor robustness to small changes in input data raise additional concerns when ML-based intelligent data processing is widely deployed in critical areas such as autonomous driving. On a higher level, input data may contain certain, sometimes unknown biases, which consequently are mirrored by the ML system.

Miscalibration of trust
Due to the unpredictable behaviour of an ML system and the high effectiveness of these systems in many cases, humans can be tempted to accept technical systems as human-like partners (anthropomorphization) and to trust the systems more than it was adequate for their actual capabilities, or humans may undertrust systems exhibiting unexpected behavior.

Low level of human control and involvement
Most ML algorithms rely on either a hypothetical model of the distribution of data or concrete interpretation (labeling) of data. Such constructions have become one major hurdle to enabling ML systems with high levels of human control, such as human-like reasoning and generalization. Specifically, it is rather difficult to find the right level of human control on which the system can effectively communicate with humans to obtain such input.

To tackle these challenges from a systems engineering perspective, we are introducing Human-Centric Machine Learning (HCML), a new paradigm in system construction taking a human-machine collaboration perspective. In this white paper,

- we provide a general framework for understanding the specific problems in human-machine interaction for ML systems;
- we outline an approach to the design of intelligent assisting systems which augment human capabilities instead of reducing the human to a controlling instance;
- we define a number of relevant research topics and research questions which fit fortiss' capabilities;
- we suggest new directions of research in ML systems which integrate the efforts of two fortiss labs.

Framework

Several high-level frameworks for Human-Centric Artificial Intelligence (HCAI) have been recently published. These frameworks aim to create a balance between human abilities and system capabilities in order to empower humans with powerful tools, but not to deprive them of control over the task being carried out.

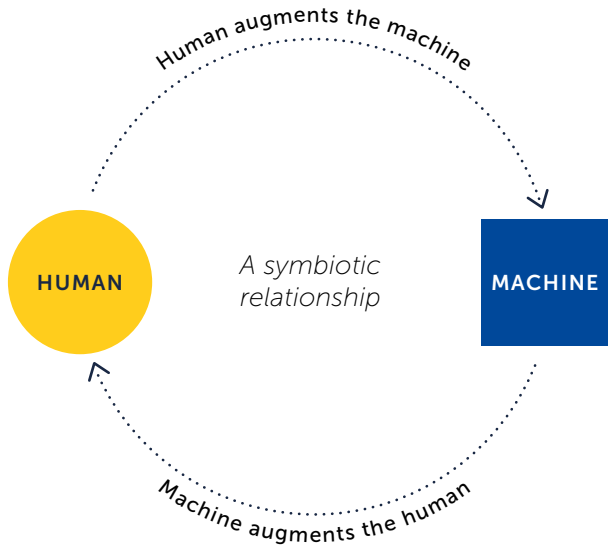

Figure 1. IBM high-level AI framework. Source: IBM.

An example of a very high-level framework is provided by IBM[1] which views intelligent systems as part of an overall ecosystem and stresses a symbiotic partnership between human and machine (Figure 1). This point of view makes it very obvious that the design of systems based on ML needs to take the human user into account from the beginning. Taking this point of view as the basis for system design, we see a number of consequences at several levels:

- **Development process**
  The development of the system has to follow a human-centered design (HCD) approach, since the ultimate design goal is the combined power of human and machine to address human needs. This type of process will be similar to the well-known User-Centered Design (UCD) method.. However, for ML systems, the development process needs to explicitly involve considerations for training data and the human impact of all data-related decisions and development activities.

- **Overall system design**
  ML models need to be integrated purposefully, following a clear idea of how they can support users. Designs that treat humans as fallbacks for imperfect ML models should be avoided. System design is dominated by control loops that take humans and machines in a single closed loop.

- **Interaction design**
  The machine has to present its inner workings on an appropriate level of abstraction such that humans can interpret it and influence what is going on.

- **Algorithm design**
  ML algorithms need to be designed in a way that input coming from humans can be integrated.

- **Advanced machine learning principles**
  In order to achieve a better match to human decision making, ML needs to take into account higher level concepts like causality.

A more detailed framework for HCAI is defined by Ben Shneiderman (Shneiderman, 2020b) which stresses the fact that human control and computer automation are not opposites, but two dimensions of a design space. We illustrate this framework here for the application of a decision support system helping humans to act in a complex situation with a huge amount of information sources of varying reliability (Figure 2). A concrete example would be a system supporting the command of a complex rescue operation. The figure shows that human control and computer automation are independent dimensions, and that we can design systems that aim to perform well in both dimensions (upper right corner). In the case of a decision support system in a rescue operation, this means the system keeps the human operator in full control but improves the basis for their decisions and eases the execution of decisions. This is in contrast to fully automated systems like in the lower right corner, where the role of the human is reduced to check system decisions for sanity (which is neither what the human nor what the system are easily capable of).

1 https://www.ibm.com/design/ai/fundamentals/

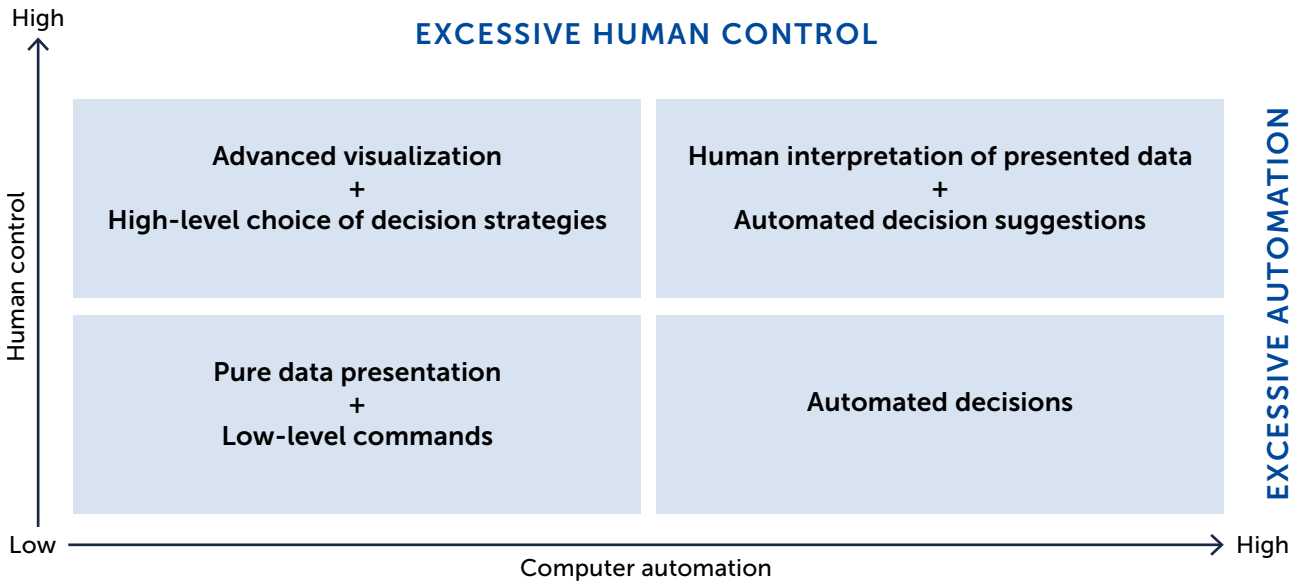

Figure 2. HCML design space, based on Shneiderman's HCAI framework (Shneiderman, 2020b)

Research Topics and Research Questions

In this section, we list a number of high level research topics for the outlined research field (Figure 3), together with examples of more concrete research questions.

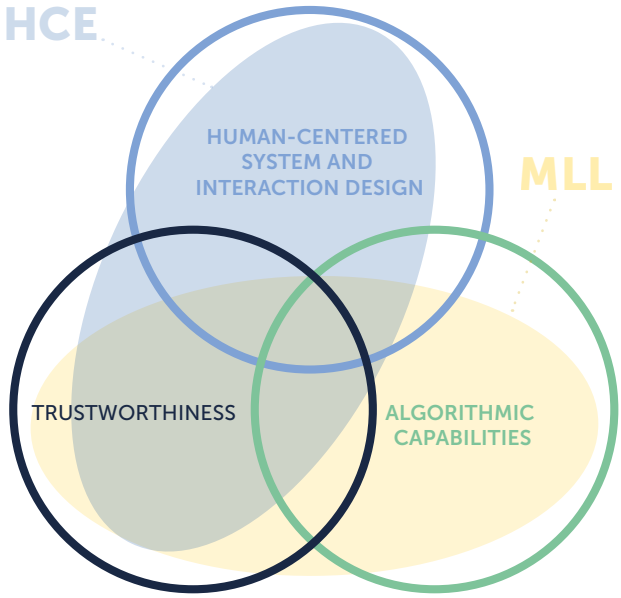

Figure 3. High level research topics for human-centered machine learning. While all three topics are distinct, there is also some overlap and interaction. HCE tends to focus on Human-Centered System and Interaction Design, while MLL tends to concentrate on Algorithmic Capabilities. Trustworthiness is the main intersection of the two competence fields, but there are interactions in the other two topics as well.

Research Topic 1

Human-Centered System and Interaction Design

The current advances in ML are rapidly pushing the boundaries of what computing systems are capable of, both in terms of their ability to make sense of and act upon their environment, as well as the possibility to adapt to their human users. However, while ML algorithms are making great strides, it is often unclear how to design ML systems such that they live up to their promises in the real world. For all their potential, ML systems also create significant challenges for the human-machine interaction as they often violate established usability principles like predictability and consistency. Much more research is necessary to understand how to employ the predictive and adaptive power of ML in a way that is aligned with human needs.

Examples of **research questions**

1.1 What are appropriate methods, frameworks, metaphors or patterns for designing ML systems around human needs from the ground up?

1.2 How to model human cognitive processes and human behavior to optimize user-adaptive systems for intuitive and efficient human-ML interactions?

1.3 How to generate semantic (structural) knowledge from observations/data in a specific domain, in order to facilitate the design of human-centered systems and actions?

Research Topic 2

Trustworthiness

Given that ML enables higher degrees of automation for more complex tasks, the trustworthiness of ML-driven systems is crucial. This concerns both the model development and the user-facing system design. Models need to be as robust and as bias-free as possible, which relies heavily on the training data and advanced ML algorithms. Furthermore, these models need to be integrated into system designs that facilitate appropriate trust of human users towards machines to avoid both misuse and disuse.

Examples of **research questions**

2.1 Under which conditions and in which form should explanations be provided to users to calibrate their trust in ML systems? What are alternatives to explanations for trust calibration?

2.2 How to develop high-level automated ML algorithms, so as to respect trustworthiness requirements in a dynamic working environment?

Research Topic 3

Algorithmic Capability

As the core component of ML pipelines, feature extractors trained in a pure data-driven manner rather than considering high-level task abstraction, fail to infer when given out-of-distribution input. Poor generalization, therefore, leads to low robustness and poor user experience due to the lack of human involvement (Madry et al., 2017). While many research efforts have been established to improve the model robustness against out-of-distribution data, the major research direction, however, neglects human needs from a system design perspective. Instead of hard-coding the entire system with off-the-shelf machine learning algorithms, exploiting interactive patterns towards better knowledge integration provides a further opportunity for HCML development. To overcome the challenges of low robustness and poor user experience caused by the lack of human involvement, we essentially focus on two perspectives, namely task-driven adaption of ML systems with respect to human needs and human-centered learning algorithm design. The former perspective suggests the adaption of ML algorithms by extending the algorithmic capability of user understanding, personalization and interaction. The latter perspective emphasizes the development of novel learning algorithms that are capable of efficiently abstracting and leveraging human knowledge.

Examples of **research questions**

3.1 What are the proper abstractions of human knowledge using graph representation learning?

3.2 How to design domain-dependent multi-modality interaction patterns for smooth integration of human knowledge?

3.3 How and to which degree can human knowledge or human-oriented metrics have an impact on the capacity of ML algorithms in terms of automation, generalizability, and explainability?

An Integrated Approach at fortiss

The aim of our development of HCML is to design ML-based intelligent systems that (a) empower humans with ML techniques and (b) simultaneously achieve high levels of human control and high levels of ML automation. Since both HCE and ML are well-established scientific disciplines which have their distinct perspectives, the uniqueness of our approach is to integrate progress in both HCE and ML to facilitate and enhance the development of the structure of HCML. Our approach as illustrated in Figure 4 can be viewed as an iterative, agile method. On the one hand, HCE benefits from both high levels of human control and high levels of machine automation to empower human capabilities. On the other hand, ML adapts to the requirement of empowering humans to construct advanced ML paradigms.

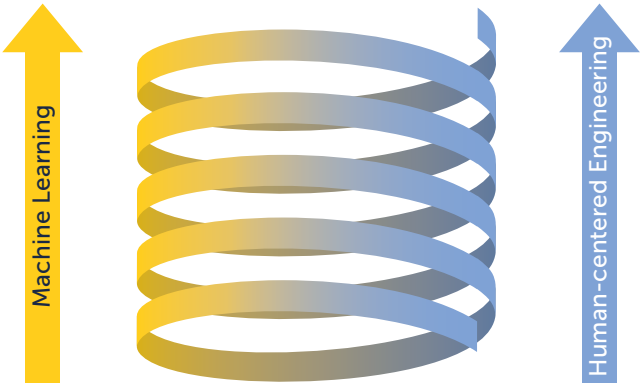

Figure 4. A spiral model of HCML interaction

HCE

According to a simplified view suggested by Shneiderman (Shneiderman, 2020a), AI researchers pursue two grand goals: the emulation goal and the application goal. The emulation goal denotes the desire to emulate and to surpass human capabilities with computers, while the application goal is concerned with deploying AI into real-world applications. Today's impressive advances in ML algorithms are primarily driven by the emulation goal. However, this also has the effect that real-world ML applications nowadays are shaped much more strongly by the emulation goal perspective than by the application goal. As a consequence, human factors are often not adequately considered in ML application designs.

The HCE competence field approaches HCML from the application goal perspective. More specifically, our approach focuses on the elements listed below.

ML for human augmentation
For one, we explore how to apply ML algorithms to augment human skills. Today's ML system designs are usually centered around the ML model that is designed to solve a human task as automatically as possible. In these systems, HCI issues are usually only considered to get the human "back into the loop" as a fallback for shortcomings of the ML model. A common example are explanation interfaces for trust calibration in decision support systems. Here, the explanations are designed to help users notice when they need to override erroneous model outputs. Another example is interactive machine learning (iML), where the objective is often to improve the ML model through user inputs.

We want to take a step back and design ML systems around humans rather than models. We therefore prefer to think of ML-in-the-loop rather than the popular human-in-the-loop. For instance, for a decision support system, our vision would be a "mental prosthesis" rather than an automated decision making machine with explanation interface (Zhang et al., 2021). The goal of the former is to contextually augment the human-led decision-making process, while the latter means that users either follow a machine decision or make a decision on their own. To this end, we try to understand the actual concerns, needs and tasks of humans before designing solutions or judging the quality of solutions, such as through interviews or contextual inquiries (RQ **1.1**).

A key concern in our approach is the impact of inevitable model errors on the human-machine interaction (Zhang & Hußmann, 2021) and how to design ML systems such that model errors are detectable, correctable and non-critical. To achieve such designs that are tolerant of model errors, we adopt a holistic view of the design space of human-ML interaction. For instance, we look beyond explainability to design trustworthy ML systems, taking into consideration more interactive system designs as well (RQ **2.1**) (Zhang et al., 2021).

ML systems understanding their human users
Further, we conduct research into ML systems that have a better understanding of their human users. The basic idea is to build models of human behavior (general models and individually specific models) with the goal of integrating them into systems that are then able to adapt to their users and interact smoothly with humans (RQ **1.2**).

To address this challenge, we build a tool to learn formal models of human behavior. In such a system, we let the user experience different stimuli and record the resulting human reaction. This enables us to build a system that can understand the human

black box, based on human analytics and reasoning of the users' experienced stimuli and the users' resulting reactions. As a first concrete example, we currently focus on human behavior influenced by stress. We have developed virtual reality (VR) experiences which that the user on a journey involving different stimuli. By observing user behavior and physiological data, we try to understand which stimuli causes which reaction of the user and build a corresponding model. This is currently being studied for stress detection and control, but the long-term goal is to tailor a personalized experience to the user that enables smooth collaboration between intelligent system and human by making the system understand its user.

Human-computer collaboration in joint learning situations

In addition, we investigate two common components and one major issue of human-computer collaboration in a joint learning setting, where a human is enabled to influence the decision-making results either directly or indirectly. The first component is the proper knowledge abstraction (RQ 3.1) which is considered as a medium that connects humans and algorithms in the learning process and continuously offers hints. Typically, those hints, such as causal relation, must be invariant to noise and data perturbation. Han et al., 2020 shows that a recommender benefits from sequential relationships, with an attention mechanism learning the effectiveness of fed human knowledge. As we approach HCML mainly from an application perspective, it is necessary to study the task-dependent multi-modal interaction design (RQ 3.2). As an example, a DIP-based (Ulyanov et al., 2018) inpainting tool for a human-machine collaborative ML system was developed for the area of image restoration (Weber et al., 2020). The proposed approach translated human painting intuitions into pixel-wise modifications that can be seamlessly refined by our system. Finally, yet importantly, we plan to address the out-of-distribution generalization problem by leveraging acquired human knowledge such as graph-structured data, which is insightful but intractable to learn in a pure data-driven fashion. Our first step here is to design textual queries for scene grounding. Our aim is essentially to embed causation into queries for a more robust and explainable scene grounding (RQ **2.2**).

MLL

MLL focuses mainly on two research challenges

Enhancement of human control with automated ML techniques

In order to enhance human control over an intelligent system, it is crucial to have effective and efficient ways to 1) understand human intentions and 2) interpret the environment or situation around the human. One challenge to understanding human intention lies in the fact that human thinking is often structural and implicit. We thus propose employing ML paradigms that are capable of dealing with structural data to extract human intentions (RQ 1.3).

- Neural structural generative models
  Knowledge that is perceivable or understandable to humans is mostly structural, such as graphs, relations and networks. Thus, in order to enable efficient high-level human control over ML systems, it is important to require the ML techniques to be able to represent or generate new knowledge from data in a human-understandable format.

- Causal representation learning
  Arguably, a core problem for AI is causal representation learning, such as the discovery of high-level causal factors from low-level observations. As a central property of human intelligence, causal representation learning is expected to enable better understanding of machine decisions, so as to promote high-level human control in ML systems.

To the second task, original observations of the environment are generally far too overwhelming for humans to interpret quickly. Therefore, authentic, human-level annotations or instructions are critical to facilitating high levels of human control (RQ 1.3). Candidate ML concepts to be developed further in this topic are

- Question answering
  In order to facilitate high levels of human control in ML systems, it is crucial to equip the ML techniques with the capability of semantic understanding. With such a functionality, an HCML system takes semantic inquiries from humans, interprets human intentions automatically, and finally empowers humans with high-level outcomes from the ML methods.

Automation of ML techniques with implicit human input

As discussed above, human intention is often implicit, or even subtle. Hence, the most essential task is arguably to extract a reliable representation from observations aligned with human preferences or values, to enable development and automation of effective and efficient ML algorithms. One concrete challenging scenario is that humans often deploy a relatively small amount of samples and negative examples for learning. We thus propose to develop human-centered representation learning algorithms in order to automate ML solutions in intelligent systems. In this context, we will focus on the following paradigms:

- Self-supervised learning
  This is a technique for learning representations from typically high-dimensional signals by predicting a derived signal, such as the reordering of a shuffled audio sequence. Instead of manual labels, this technique relies on general information that applies to the data, such as temporal coherence in the previous example. The learned structure makes it easier to solve downstream tasks (more efficiently than based on the original raw data) and allows for better visualizations that facilitate understanding of the data.

- Contrastive learning
  Also known as "learning by comparing", this is closely related to self-supervised learning. However, instead of predicting some auxiliary signal, contrastive learning makes use of general relations between data instances. For example, shifting an audio signal by a short time will not significantly change the content or the speaker identity. This allows a user to encode high-level information.

- Continual learning
  User-facing machine learning algorithms need to *self-evolve* together with humans. For example, wearable devices should fit better with their unique owners because of more collected user-specific data, or an autonomous driving system can become familiar with the daily commute traffic. Such adaptations come not only from self-supervised learning, but there are also strict requirements from continual learning which avoid the so-called *catastropic forgetting* of previously learned abilities.

The focus on these frameworks allows us to seamlessly integrate another concept referred to as *Knowledge-Augmented Machine Learning*, which aims to transfer purely data-driven approaches towards knowledgeable systems that can easily adapt to new scenarios and environments (RQ 3.3).

References

Han, Z., Anwaar, M. U., Arumugaswamy, S.,Weber, T., Qiu, T., Shen, H., Liu, Y., and Kleinsteuber, M. Metapath-and entity-aware graph neural network for recommendation. *arXiv preprint arXiv:2010.11793,* 2020.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083,* 2017.

Shneiderman, B. Design lessons from AI's two grand goals: Human emulation and useful applications. *IEEE Transactions on Technology and Society,* 1(2):73–82, June 2020a. ISSN 2637-6415. doi: 10.1109/TTS.2020.2992669.

Shneiderman, B. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human– Computer Interaction, 36*(6):495–504, April 2020b. ISSN 1044-7318, 1532-7590. doi: 10.1080/10447318.2020.1741118.

Ulyanov, D., Vedaldi, A., and Lempitsky, V. Deep image prior. In *Proceedings of the IEEE conference on computer vision and pattern recognition,* pp. 9446–9454, 2018.

Weber, T., Han, Z., Matthes, S., Hussmann, H., and Liu, Y. Draw with me: Human-in-the-loop for image restoration. In *Proceedings of the 25th International Conference on Intelligent User Interfaces,* IUI '20, pp. 243–253, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450371186. doi: 10.1145/3377325.3377509. URL https://doi. org/10.1145/3377325.3377509.

Zhang, Z. T. and Hußmann, H. How to manage output uncertainty: Targeting the actual end user problem in interactions with AI. In *Joint Proceedings of the ACM IUI 2021 Workshops co-located with the 26th ACM Conference on Intelligent User Interfaces,* IUI '21, April 2021.

Zhang, Z. T., Liu, Y., and Hußmann, H. Forward reasoning decision support: Toward a more complete view of the human-ai interaction design space. In *Proceedings of the 14th Biannual Conference of the Italian SIGCHI Chapter, in press,* CHItaly '21, July 2021.

Imprint

Publisher
fortiss
www.fortiss.org
© 2021

Authors
Dr. Yuanting Liu
Dr. Hao Shen

Layout
Sonja Taut

Editing
Daniel Hawpe

Print
viaprinto | CEWE Stiftung & Co. KGaA
Martin-Luther-King-Weg 30a
48155 Münster

ISSN Print
2699-1217

ISSN Online
2700-2977

1. Edition, September 2021

Picture Credits
Titel: adobe stock © sdecoret
Seite 4: adobe stock © lassedesignen
Seite 8: adobe stock © Deepak
Seite 10: adobe stock © Sikov
Seite 12: adobe stock © faraktinov
Seite 14: fortissGmbH ©Kathrin Kahle

fortiss is the Free State of Bavaria research institute for software-intensive systems based in Munich. The institute's scientists work on research, development and transfer projects together with universities and technology companies in Bavaria and other parts of Germany, as well as across Europe. The research activities focus on state-of-the-art methods, techniques and tools used in software development and systems & service engineering and their application with cognitive cyber-physical systems such as the Internet of Things (IoT).

fortiss is legally structured as a non-profit limited liability company (GmbH). The shareholders are the Free State of Bavaria (majority shareholder) and the Fraunhofer-Gesellschaft zur Förderung der angewandten Forschung e.V.

Although this white paper was prepared with the utmost care and diligence, inaccuracies cannot be excluded. No guarantee is provided, and no legal responsibility or liability is assumed for any damages resulting from erroneous information.

fortiss GmbH
Guerickestraße 25
80805 München
Deutschland
www.fortiss.org
Tel.: +49 89 3603522 0
E-Mail: info@fortiss.org

fortiss