INAUGURAL-DISSERTATION

zur

Erlangung der Doktorwürde

der

Gesamtfakultät für Mathematik, Ingenieur- und Naturwissenschaften

der

Ruprecht–Karls–Universität

Heidelberg

vorgelegt von

M.Sc. Martin Schüßler

aus: Burg b. Magdeburg

Tag der mündlichen Prüfung:

User-Centred Evaluations in Computer Vision: Empirical Insights on Explanation Methods and Gaze-Aware Videoconferencing

Gutachter: Prof. Dr. Carsten Rother

Abstract

Computer vision has played a crucial role in the recent increase in interest in artificial intelligence. Neural networks, in particular, have led to breakthroughs in many application areas, ranging from the recognition of lung cancer in CRT images to novel ways of creating virtual immersive environments and photorealistic avatars.

Most computer vision research focuses on demonstrating new technological solutions and showcasing their capabilities. This dissertation attempts a change of perspective from a technology-centred to a user-centred focus to ensure the successful deployment of innovations in real-world scenarios. It explores the intersection of human-computer interaction (HCI) and computer vision, with particular emphasis on the domains of interpretable vision and gaze-aware video conferencing. Comparative user studies with robust baseline conditions are central to this work and form the cornerstone of the methodological strategy.

The first part of this dissertation delves into interpretable vision, evaluating the efficacy of different explanation methods in enhancing users' understanding of image classifiers. Through rigorous experimental design and the development of a novel synthetic dataset, two studies provide nuanced insights into the effectiveness of these explanation methods. Results show that saliency maps can draw users' attention to specific features, while counterfactuals help discover model biases. Notably, results also show that simple example-based explanations can be overall just as effective as more sophisticated methods while being easier to implement. We argue that these explanations should serve as a benchmark for evaluating any future explanation methods. These results highlight the importance of measuring how well users can reason about a model rather than solely relying on technical evaluations or proxy tasks when assessing the explanation techniques.

The second part of this dissertation shifts the focus to image synthesis. It addresses the quality of the video-conferencing user experience by exploring a conceptual system capable of conveying gaze and attention. Gazing Heads is a round-table virtual meeting concept that enables direct eye contact and signals gaze via controlled head rotation. We built a four-party camera-based simulation to evaluate Gazing Heads against a conventional "Tiled View" video-conferencing system. In contrast to prior concepts, Gazing Heads increases social presence, mutual eye contact, and user engagement. We attribute these novel results

to the amplifying effect of head rotations for conveying gaze. In its current design, Gazing Heads unequivocally enhances the experience of users in highly interactive small group meetings. Our work also highlights the remaining challenges in implement Gazing Heads on commodity hardware and in achieving seamless integration into daily video-conferencing.

Overall, this dissertation contributes to the fields of HCI and computer vision by providing empirical insights into the benefits and limitations of current computer vision applications from a user-centred perspective. Published across top-tier machine learning and HCI venues, this research emphasises the need for more meticulously designed user studies in computer vision. It provides foundational artefacts, such as benchmark datasets, study designs, and system concepts, which can serve as a starting point for future research.

Zusammenfassung

Fortschritte in der Wissenschaft der Computervision (CV), die sich mit computerbasiertem Sehen beschäftigt, haben in den letzten Jahren zu einem gestiegenen Interesse an künstlicher Intelligenz (KI/AI) beigetragen. Mit Hilfe von neuronalen Netzen kam es in einer Vielzahl von Bereichen zu unerwarteten Durchbrüchen. Es ist zum Beispiel nun technisch möglich, Lungenkrebs in Computertomografiebildern automatisierter und zuverlässiger zu erkennen. Darüber hinaus lassen sich immersive virtuelle Umgebungen schaffen, die nicht nur sehr realistisch wirken, sondern menschliche Akteure darin fast realitätsgetreu nachbilden.

Angesichts dieser Errungenschaften ist es nicht verwunderlich, dass primär eine technologiezentrierte Betrachtungsweise in den beteiligten Wissenschaften vorherrscht. Es wird besonderer Wert auf die Entwicklung innovativer Lösungen gelegt, die möglichst eindrucksvoll präsentiert werden sollen. Diese Arbeit unternimmt den Versuch eines Perspektivwechsels und nähert sich dem Feld mit einer kritischeren, nutzerzentrierten Sichtweise. Dabei sind Nutzerstudien, die neue Ansätze evaluieren und mit einfacheren, bereits etablierteren Lösungen vergleichen, integraler Bestandteil der Methodik.

Diese Arbeit hat zwei Teile, wobei sich jeder mit einem eigenen Feld der Computervision beschäftigt. Der erste Teil befasst sich mit erklärbarer künstlicher Intelligenz. Es werden algorithmische Erklärungsmethoden evaluiert, die Bildklassifizierungsmodele verständlicher machen sollen. Gegenstand der vorgestellten Studien sind Saliency-Maps, Counterfactuals und Concept-Maps. Zum Vergleich kommt eine einfachere Methode zum Einsatz, die eine Auswahl von Bildern und deren Klassifizierungsergebnisse zeigt. Die Interpretation der Klassifikationsgründe wird also mehr dem Nutzer überlassen. Diese Studien wurden iterativ und mit besonderer Sorgfalt entworfen. Dabei wurde auch ein synthetischer Datensatz geschaffen, der auf den Einsatz in Nutzerstudien abgestimmt ist. Diese strukturierte Vorgehensweise erlaubt einen nuancieren Einblick in die Wirksamkeit verschiedener Erklärungsmethoden. Die Ergebnisse der zwei Studien sind überraschend. Die einfach zu implementierende Methode, die Beispielbilder verwendet, erzielt vergleichbare und zum Teil sogar bessere Ergebnisse als die komplexeren Methoden. Daraus wird unter anderem die Erkenntnis gezogen, dass Erklärungen mit Beispielbildern als Benchmark für alle zukünftigen Evaluierungen von Erklärungstechniken dienen sollten. Neben den detaillierten Erkennt-

nissen zu jeder evaluierten Erklärungstechnik leistet diese Arbeit auch einen wissenschaftlichen Beitrag, indem sie die konkrete Umsetzung der Benchmarkmethode, ein passendes Studiendesign und einen auf Nutzstudien abgestimmten Datensatz bereitstellt.

Im zweiten Teil der Arbeit verlagert sich der Fokus auf die Bildsynthese. Es wird untersucht, ob sich Videokonferenzen verbessern lassen, indem man die Blickrichtung der Teilnehmer mittels synthetisierter Kopfdrehungen visualisiert. Es wird ein Konzept für ein solches System namens Gazing Heads vorgestellt. Dieses wird mit Hilfe von mehreren Kameras prototypisch umgesetzt und einer umfassenden Nutzerstudie unterzogen. Als Vergleichssystem dient die marktübliche Videokachelansicht. Dabei wird untersucht, wie sich Gazing Heads auf die Kommunikation, soziale Präsenz und Natürlichkeit der Interaktion auswirkt. Die Ergebnisse zeigen, dass sich die Blickrichtungen der Teilnehmer eindeutig mit synthetischen Kopfrotationen vermitteln lassen. Dies hatte im Experiment zur Folge, dass die Probanden Gazing Heads klar bevorzugten. Sie fühlten sich stärker in die Unterhaltung eingebunden und nahmen eine gesteigerte soziale Präsenz wahr. Daraus wird die Schlussfolgerung gezogen, dass sich das Videokonferenzerlebnis mit dem Einsatz von synthetisierten Kopfdrehungen signifikant verbessern lässt. Um diese Vorteile allerdings in der Praxis nutzen zu können, muss eine nahtlose Integration in bestehende Anwendungsfälle gewährleistet werden. Darüber hinaus gilt es, technische Herausforderungen zu überwinden, damit Kopfrotationen und Gesichter als möglichst realitätsgetreu empfunden werden.

Insgesamt leistet diese interdisziplinäre Arbeit Beiträge zu zwei Wissenschaftsbereichen, den der Mensch-Computer-Interaktion und den der Computervision. Sie liefert empirische Erkenntnisse zu den Vorteilen und Grenzen aktueller Anwendungen von künstlicher Intelligenz. Darüber hinaus zeigt sie den Mehrwert empirischer Netzwerkstudien im Computervision-Bereich auf. Da dieser Perspektivwechsel wichtige wissenschaftliche Erkenntnisse liefert, erkennt man nicht zuletzt auch daran, dass die hier vorgestellten Studien auf führenden Konferenzen und in Journalen veröffentlicht wurden.

Neben den neuen Erkenntnissen und offenen Forschungsfragen liefert diese Arbeit auch wiederverwendbare Artefakte wie Systemkonzepte, Datensätze und Studiendesigns für zukünftige Forschungsvorhaben.

Contents

List of Figures

List of Acronyms

AI Artificial Intelligence

ANOVA Analysis of Variance

AP Average Precision

AR Augmented Reality

BIFOLD Berlin Institute for the Foundations of Learning and Data

CNN Convolutional Neural Network

CRT Cathode-ray tube

DARPA Defense Advanced Research Projects Agency

DNN Deep Neural Network

F F-Statistic

FFP2 Filtering Face Piece type 2

FN False Negative

GAN Generative Adversarial Network

GDPR General Data Protection Regulation

GH Gazing Heads

HCI Human-Computer Interaction

IL Interlocutor

IML Interpretable Machine Learning

INN Invertible Neural Network

LRP Layer-wise Relevance Propagation

M Mean

ML Machine Learning

MSE Mean Squared Error

N Number of Participants

NN Neural Network

p P-Value

RQ Research Question

SARS-CoV-2 Severe Acute Respiratory Syndrome Coronavirus type 2

SD Standard Deviation

TV Tiled View

UDP User Datagram Protocol

UX User Experience

VGG Visual Geometry Group

VOC Visual Object Classes

VR Virtual Reality

XAI Explainable Artificial Intelligence

Acknowledgements

This dissertation was completed with the support and guidance of many individuals. While scientific contributions are detailed in the respective chapters, I would like to express my heartfelt thanks to those who helped me along my PhD journey.

First and foremost, I owe my deepest gratitude to my primary supervisor, Prof. Dr. Carsten Rother. His willingness to embrace my human-centered perspective within his computer vision lab made this work possible. During the pandemic, we faced the challenge of conducting an ambitious, large-scale, in-person user study, and Carsten's unwavering support ensured its success.

I would also like to thank Prof. Dr. Enrico Costanza for the invaluable suggestion to collaborate with his exceptional PhD student, Dr. Ahmed Alqaraawi. This partnership and Enrico's guidance led to my most cited work to date.

Dr. Leon Sixt's patience and willingness to engage with my human-centered inquiries made him a standout in the machine-learning field. Our collaboration culminated in my most challenging paper. Even after two rejections, his supervisor, Prof. Dr. Tim Landgraf, kept us motivated and brought us to completion. Prof. Dr. Andrew Blake deserves special thanks for sharing his vast experience in research and computer vision and for showing me how to write clearly and precisely. I am also grateful to Prof. Dr. Raimund Dachselt and Dr. Martin Spindler, who taught me how to think like a scientist over a decade ago. Working with Raimund again on my latest study was a pleasure. I am also appreciative of Luca Horman, who patiently implemented countless changes to the Gazing Heads prototype, and to the founders of Copresence, Dr. Titus Leistner and Radek Mackowiak, for their valuable input on the same project. The taxonomy I developed with Dr. Michael Chromik was instrumental in shaping my XAI studies despite requiring many late-night work sessions.

My journey began at the Weizenbaum Institute, where Dr. Milagros Miceli broadened my perspective with her critical sociological insights into data annotation and machine learning. I am thankful to the other colleagues at the Institute for their stimulating conversations. I also wish to thank Prof. Dr. Ina Schieferdecker, Prof. Dr. Bettina Behrendt, Dr. Diana Serbanescu, and Dr. Stefan Ullrich for their continued support of my interdisciplinary research ideas. Special thanks to Evelyn Adams, without whom navigating the bureaucratic

complexities would have been impossible.

I have been fortunate to supervise and collaborate with talented students, including Philipp Weiß, Tianling Yang, Marie Theres Thomas, and Žan Jonke, and I am grateful for their hard work. I also appreciate the insights shared by Dr. Fenne große Deters, Dr. Hannes-Vincent Krause, Berit Wiegmann, and Otto Lutz on experimental design and quantitative analysis. I also wish to thank all study participants for their time. [1] Lastly, I express my deepest gratitude to my parents for their unwavering support throughout my academic journey. Their encouragement made this pursuit possible. This work was supported through funding from the German Federal Ministry of Education and Research (BMBF) under grant number 16DII113. The respective publications acknowledge additional sources of funding.

[1] Participants were ethically compensated. In total, over 12000 Euros were spent on compensation.

Conceptual and Methodological Prelude

User Studies as a Contribution to the Field of Computer Vision

In my collaborative projects, my primary focus was on the conceptualisation and orchestration of user studies. I must point out that, despite their relative scarcity in computer vision, they are no less significant or complex than technical contributions. Throughout the peer-review process, in which I engaged as both an author and reviewer, I encountered a shared challenge among my colleagues: there is a lack of recognition in the field of computer vision for the time and effort required in conducting such evaluations (a sentiment echoed by Doshi-Velez and Kim [52]).

Alongside technical contributions, the conceptualisation of user studies is a meticulous and iterative endeavour. It begins with a foundational theoretical idea and culminates in a robust design that offers credible empirical insights. The many iterations one must go through may not be obvious here, as I have intentionally omitted details of intermediate phases, such as my prototypes and pilot studies, from the main text to maintain clarity and conciseness. However, minute inaccuracies in study designs have the potential to skew final measurements, leading to erroneous interpretations. Therefore, I engage in a thorough reflection on the validity of my studies in this work, alongside a critical examination of related literature. The designs presented here are intended to lay the groundwork for future rigorous user studies in computer vision, hopefully inspiring and guiding subsequent human-centred research in this domain.

Key Psychological Research Design Concepts

This section offers a concise overview of key principles in psychological experimental design, drawn from a lecture presented by the author at the BIFOLD summer school. This summary is particularly relevant for readers less familiar with experiments involving human subjects. In line with the theme of this work, the review is tailored to the context where

an experimenter evaluates different computer vision systems or their variations through a user study. The definitions and frameworks discussed here are primarily influenced by the works of MacKenzie [139] and Field et al. [67].

Validity of discovery is the primary concern in any experimental design. The principal challenge lies in managing the delicate equilibrium between internal and external validity. **Internal validity** indicates the extent to which the outcomes observed can be attributed unequivocally to the test conditions. **External validity** refers to the degree to which the results of the experiment can be extrapolated to other contexts, essentially questioning whether the effects on user experience observed in the study can be confidently expected to occur under similar conditions elsewhere.

Experimental Variables

In user study design, broad research inquiries, such as "What impact does System X have on user experience?", are distilled into precise, testable hypotheses, such as "System A positively affects Aspect X." It is essential to carefully identify and consider all pertinent *experimental variables* in this process, ensuring a comprehensive and targeted exploration of the hypotheses.

Factors (also referred to as *independent variables*) are manipulated by experimenters to test hypotheses and elucidate effects. These typically include categorical variables with distinct levels. For instance, the factor "System" might have levels, such as "System A", "System B", etc. A minimum of two levels per factor is essential for conducting comparative statistical analyses.

Outcome or response variables (also termed *dependent variables*) quantify the experiment's results and are suspected to be influenced by the chosen factors. For instance, the "System" factor could influence a user experience outcome, such as the error rate.

Measurement scales are pivotal in defining the outcome variable's resolution and applicable statistical tests. Measurement scales can be categorised as follows: *nominal*, as in the case of categorical outcomes, such as users choosing a favourite system from A, B, or C; *ordinal*, which involves ordered categorical outcomes, such as users ranking systems A–C by preference; *interval*, which refers to a continuous scale with equal intervals indicating equal distances, such as task completion time; and *ratio*, an interval scale with a meaningful zero, such as percentage of correct answers. The classification of scales for self-reported

data obtained from questionnaire ratings remains an ongoing subject of debate. The prevailing opinion is that self-reported questionnaire ratings should be considered ordinal unless a verified *Likert-Scale questionnaire* has been used [67]. Yet, the problematic practices of treating unverified Likert items or custom questions as interval scales persist in many studies, often without ensuring that the underlying assumptions of statistical tests have been met.

Circumstantial factors (for example, participant demographics) may affect experiment outcomes but are not directly tied to the hypothesis. *Randomisation* approaches, such as inviting random participants from the target population, can be used to mitigate circumstantial factors and are advisable when generalisability is the priority, despite the potential to impact internal validity by adding noise and introducing the risk of overlooking relevant effects, such as those related to gender or education. *Fixed settings* that avoid extremes, such as equal male-female participation or a fixed common display size for the experimental interface, improve internal validity by controlling variability but might limit generalisability to real-world scenarios.

Confounding factors alter the outcome and co-vary with independent variables, posing a substantial threat to validity. Consider a study where participants invariably interact with System A first and System B second while performing a demanding task. If System B shows lower performance, it could be partially due to increasing fatigue rather than issues with the system. This scenario demonstrates how the order of administering conditions acts as a confounding factor and misleads the interpretation of results.

Conditions

In most studies, participants are allocated to Conditions, which are combinations of factor levels. When conditions encompass all possible combinations of factor levels, this constitutes a *full-factorial design*. Paired with suitable statistical methods, such designs enable the clear delineation of each factor's impact on the outcome. Studies in HCI typically limit their designs to 1–3 factors because of the exponential increase in statistical complexity with the inclusion of additional factors and levels [139].

Baseline

When evaluating systems, it is advisable to think of them as "treatments" administered to participants. The inclusion of a *baseline condition* is essential in the assessment of the effectiveness of these treatments. Here the treatment system or configuration is absent but

replaced by a viable alternative, such as a commonly used alternative or the default configuration. For instance, in evaluating a novel spatial navigation technique for handheld devices, a suitable baseline might be multi-touch navigation, commonly used in everyday handhelds [2]. This comparative approach enables quantifying improvements over the state of the art and is used in all the author's studies in this work.

[2]Refer to a prior study by the author for such a study design [217].

Part I

Evaluating Explanation Methods of Image Classifiers through User Studies

Chapter 1

Introduction

Machine learning continues to be integrated into numerous computer applications, thereby extending its influence on society across a diverse range of fields. Given its broad utility, its deployment can influence decision-making in many domains, including those that are more sensitive within society. Examples range from predictive policing [149], to healthcare [33], to social services [116], and many others [223, 34]. These predictors may yield remarkable accuracy, particularly when trained on abundant data, but many function as *black-box models*. Their "black-box" nature can stem from either proprietary closed models or intricate architectures, such as deep neural networks [190]. This means their internals "are either unknown to the observer or they are known but uninterpretable *by humans*" [78]. Therefore, there is a growing acceptance that they need to be made accountable and capable of explaining their behaviour in human-understandable terms [205, 231]. There has also been a scholarly debate as to whether the General Data Protection Regulation (GDPR) necessitates the provision of explanations for automated systems [244, 43]. Whether engineer or user, insurance company, or regulatory body, all require reliable information about what the model has learned or why the model provides a certain output. To address this challenge, much research is being conducted within the domains of *explainable artificial intelligence* (XAI) and *interpretable machine learning* (IML) on developing interpretable models, methods, and interfaces [78, 152].

However, research within the field largely focuses on computational issues, with a notable lack of human subject experiments to verify their effectiveness [3, 52, 147, 53, 240, 170, 162]. There is no consensus on evaluation methods for suggested approaches and scholars agree that the comparison and validation of diverse explanation techniques is an important open challenge [3, 52, 147, 53]. The work we present here takes an HCI perspective on this challenge and conducts rigid empirical evaluations with user studies. We focused on the evaluation of explanation techniques for image classification models. We first review applicable explanation methods and survey existing literature for best practices

to evaluate them. We then conduct two peer-reviewed studies that investigate the ability of four explanation types to increase model understanding. They reveal the unexpected result that example-based explanations are a robust competitor to other explanation methods and that saliency maps, counterfactual explanations, and automatically discovered concepts are less effective for images than previously claimed. Based on this, we emphasise the need for more rigorous studies that include example-based explanations as a baseline. We contribute a synthetic dataset with known feature importance and a baseline explanation method, both specifically designed for such evaluations. In addition, we identify usability issues in all methods and find that users' cognitive heuristics limit the usefulness of currently available local explanations. The work presented here contributes to bridging the gap between AI/ML and HCI communities and establishes more rigorous human evaluation procedures in interpretable vision. The open-source materials, including the Two4Two dataset, study designs, and code, are intended to support the replication of results and adaptation of the study design to other explanation techniques.

Chapter 2

Foundations and Literature Review

2.1 Research Questions and Publication

This chapter introduces the interdisciplinary field of *explainable artificial intelligence* (XAI), highlighting the empirical evaluation debate and clarifying key terminologies and concepts essential for the studies presented in later Chapters 3–4. We first present a high-level literature review, before narrowing our focus to explanation techniques for *image classification* to increase *model understanding*.

Research Questions

- **RQ 1** — Which explanation methods are suitable for explaining image classifiers?
- **RQ 2** — Which evaluation approaches have been proposed and discussed across disciplines in the field of XAI?
- **RQ 3** — Which study design decisions have researchers made in previous evaluations with human subjects?
- **RQ 4** — What study results (including our own) have been obtained regarding the ability of explanation techniques to facilitate users' understanding of image classification models?
- **RQ 5** — What are the best practices for assessing model understanding of image classification models in user studies?

We provide a technical review of explanation methods for image classification, addressing RQ 2 (Section 2.3.) We then derive a taxonomy that describes the main aspects of user studies in the field of XAI, thereby answering RQ 2–3 (Section 2.5.) We conduct a focused literature review in chronological order, which covers RQ 4 (Section 2.6) to show how our

work has advanced the state of the art and relates to other studies. Finally, we answer RQ 5 based on our review, taxonomy, and experience from conducting studies by deriving recommendations for future studies (Section 2.7.)

Publication The taxonomy shown in Figure 2.8 has been published as an ExSS-ATEC workshop paper, which has been cited 78 times since its publication. It was presented virtually at the 2020 Annual ACM Conference on Intelligent User Interfaces (IUI).

> Michael Chromik and Martin Schuessler. 2020. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. in *Proceedings of the IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies* (IUI). Vol. 2582. `http://ceur-ws.org/Vol-2582/paper9.pdf`

Author Contributions The taxonomy in Section 2.5 was jointly developed with Michael Chromik. Michael Chromik formulated the research idea, concept, and methodology. Martin Schuessler contributed a literature review that he had conducted individually before starting this work. Paper writing was a shared effort. Martin Schuessler revised and extended the description of the dimensions after the initial publication. Sections 2.5.3–2.5.3 are still based on the original publication but have been substantially revised. All remaining sections in this chapter are unpublished and exclusive to this seminal work.

2.2 Explaining Machine Learning Models

Understanding the mechanisms that lead to the prediction of complex machine-learning models is extremely challenging. This is problematic, since users, even those who lack expertise in ML, ought to have the capacity to determine when they should rely on predictions. Several disciplines are currently involved in the effort to solve this challenge.

Machine Learning Interpretable machine learning (IML) commonly refers to studies on transparent models and algorithms, while explainable AI (XAI) primarily addresses the generation of explanations for complex, opaque models [190, 20]. Interpretability in machine learning, as Lipton [134] notes, is not a monolithic concept, rather, it is an effort to ensure important aspects like fairness, reliability, and trust in different machine learning contexts [52]. This is evident in the goal statement of DARPA's influential XAI program:

> "Enable human users to understand, appropriately trust, and effectively manage the emerging generation of artificially intelligent partners" — Gunning and Aha [83].

The field is characterised by a rapidly expanding literature and a diverse range of technical solutions claiming to enhance *interpretability* while lacking empirical validation for this claim [3, 52, 147, 53, 240, 170, 162]. This work addresses this gap.

Human-Computer Interaction Users' understanding of systems is a primary research area in the HCI community, as evidenced by an extensive body of literature. The concept of mental models, a user's internal representation of a system [168, 155], is fundamental to the field and this work. Reflections on the impact of deep neural networks (DNN) on interpretability can be traced back to as early as 1992 [49]. Even before their prevalence, research on explanations for improved system understanding was conducted in areas such as information retrieval [118], recommender systems [91, 44, 122], and context-aware systems [132, 46]. Studies from this era show that accurate mental models enhance system interaction efficiency [26, 16, 106] and user satisfaction [44, 122], while inaccurate models lead to confusion, misconceptions, dissatisfaction, and erroneous interactions [121, 236]. Overestimating a system's capabilities also adversely affects user interaction [6, 114], potentially resulting in over-reliance [131, 32], decreased vigilance towards system failures [258], and unrealistic expectations [258].

Social Science and Psychology Explanations are a subject of longstanding interest in psychology and social science. Miller defines explanation as either a process or a product [146]. Explanation, as a process, involves identifying the causes of events and also functions as a social interaction between an *explainer* and an *explainee*, aiming to transfer knowledge about the cognitive process. Explanation, as a product, represents the outcome of this process, crafted to address a specific question. In this work, we adopt the product perspective. The evaluation of explanations involves the explainee's assessment of an explanation's adequacy [146]. Research in these fields, involving presenting varied explanations to participants, reveals that cognitive biases and heuristics often influence preference for one explanation over another [104]. Key criteria for explainees include consistency with prior beliefs, simplicity, generalisability, and alignment with current information needs [146]. This suggests that the context of use is crucial in determining an explanation's effectiveness. Research on trust in automation further reveals that users' trust in machine learning models is potentially influenced by numerous factors [131], which are very challenging to manage in a study [219].

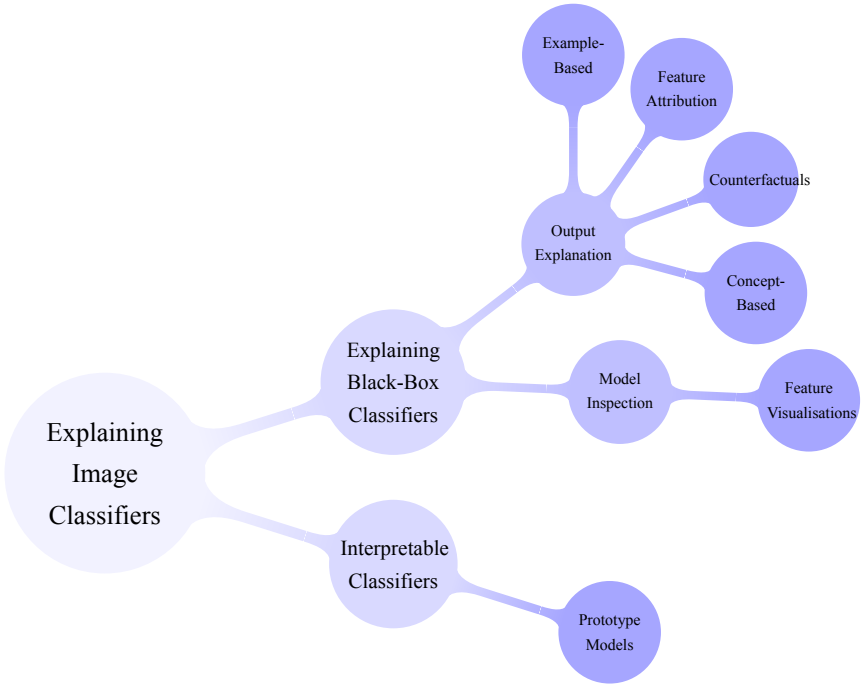


Figure 2.1. Categorisation of explanation types: We categorise relevant approaches using the taxonomy of Guidotti et al. [82]. A key distinction is whether an approach pursues inherently transparent systems or seeks to explain a black-box model without fully uncovering the underlying mechanisms [82, 134]. The latter category is subdivided as follows: output explanation, which involves creating local explanations for individual predictions, and model inspection, which includes creating visualisation techniques that help to elucidate how a model functions internally.

2.3 Explaining Image Classifiers – A Technical Review

Image classification models, such as convolutional neural networks (CNN), assign labels to images or image regions. They have demonstrated impressive results in computer vision, with a broad range of applications [179]. Such models are typically trained with large amounts of labelled images (supervised learning). They are inherently complex and often learn sub-symbolic patterns devoid of semantic meaning [21, 134]. These properties make it very challenging for users to understand how they operate. Many different methods are available to explain them. However, when we began our investigations in 2018, only two studies had investigated their ability to increase users' understanding [184, 108]. We set out to address this research gap. Consequently, we limit our technical review to interpretability approaches for image classification but refer the interested reader to reviews covering additional methods [134, 78, 82, 3, 152]. Figure 2.1 provides an overview of the covered approaches.

2.3.1 Example-Based Explanations

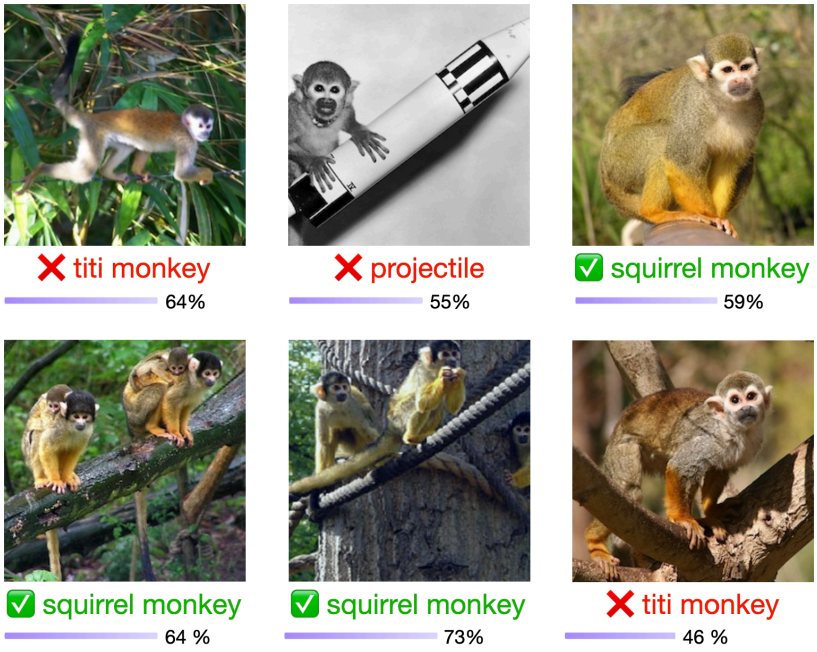


Figure 2.2. Example-based explanation: Showing correct and incorrect predictions for images of squirrel monkeys may help users understand how the system predicts this class and under which circumstance it is more likely to make a mistake. This illustrative example was created with real predictions from a visual transformer model [255].

Various methods use samples from the dataset and their corresponding predictions by the model to generate explanations. The beauty of this approach lies in the decoupling of the explanation generation from the model's internal processing, making it applicable to any model. These model-agnostic approaches are usually easier to implement and compute than more complex model-dependent methods. However, the real challenge lies in selecting and presenting the most informative examples in an interface that facilitates understanding. As Figure 2.2 illustrates, selected samples could highlight similar outcomes, such as images classified under the same category [33, 31]. They could also highlight a model's shortcomings by showing where it diverges from ground-truth data [107].

2.3.2 Saliency Maps

Feature attribution methods calculate a relevance score for each feature of an input that is used for prediction [134]. As shown in Figure 2.3, these scores can be visualised as heat maps, known as *saliency maps*, which highlight the most relevant features of an image. Such *"why"* explanations are claimed to make it easy for even novice users, to spot unusual behaviour [128], and build appropriate trust in the system [184]. However, it is important

(a) Gradient [208] **(b)** LRP [12] **(c)** Grad-Cam [202] **(d)** LIME [184]

Figure 2.3. Saliency map explanations: Different feature relevance explanations for the same "squirrel monkey" prediction of a ResNet50 model [87]. Red indicates supporting evidence, while blue indicates contradicting evidence. Different methods for calculating feature relevance yield different saliency maps. The first three explanations are taken from Schulz et al. [200]; the last image was generated using LIME [184].

to assess whether the relevance scores themselves, and not just the sample images for which they are computed, are contributing to the user's understanding. There are two broad approaches to calculating relevance scores: gradient-based and surrogate methods [5, 151].

Gradient-Based Methods (e.g. [12, 208, 202, 215, 218, 154]), compute the gradient of a prediction (e.g., the likelihood of a "cat" label) against input features (such as pixels). However, as Figure 2.3a shows, vanilla gradient-based saliency maps [208] are not very informative, which is why several variations of these methods exist. Explanation methods mainly differ in how the gradient is computed. The works of Adebayo et al. [5] and Molnar [151] cover the key differences. This work mainly looks at Layer-wise Relevance Propagation (LRP) [12] (Figure 2.3b). Unlike gradient-only methods, LRP governs propagation through the neural network from the output to the input layer with distinct rules, allowing for adjustable explanation properties.

An important limitation of gradient-based methods is that some of them partially ignore the network's parameters while others can be easily manipulated [167, 4, 74, 51, 130, 8, 211]. This *unfaithfulness* to the model has raised concerns about how accurately and truthfully these methods explain predictions.

Occlusion-Based Approaches (e.g. [184, 138, 66]), create a local feature relevance model for a prediction. They create perturbations of the input image where some regions are removed. These regions can be rectangular [262] or pixels with a similar colour, called superpixels [184, 138]. Occlusion-based methods obtain predictions for these perturbations to determine how missing regions affect the model and train a local feature-relevance model to encode feature relevance. LIME [184] uses a linear model for this purpose, whereas

(a) Base Image **(b)** "Unattractive" **(c)** "Attractive"

Figure 2.4. Counterfactual explanation: A base image used by an attractiveness classifier to generate counterfactuals via an invertible neural network [212], altering only learned attractiveness features. This method, faithful to the model's logic, uncovers biases towards skin colour and hair. Note: Images were generated for a submission of the author that is unpublished [213].

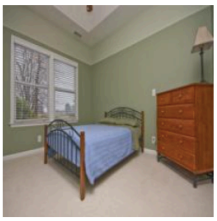
SHAP [138] uses Shapely values. The weight assigned to each region can then be used to create a saliency map, as shown in Figure 2.3d. Surrogate approaches also struggle with unfaithfulness, as they are vulnerable to attacks [214].

2.3.3 Counterfactual Explanations

Counterfactuals are created by minimally adjusting the features of an input to yield a different prediction [245] (Figure 2.4). Such hypothetical *"what-if"* examples can help users understand which features are relevant to a model. While various methods exist for simpler data types, image counterfactuals are most commonly created with a Generative Adversarial Network (GAN) [79, 142, 195, 209, 135, 15, 38]. However, GANs may not be faithful to the model they explain, as they train a separate model to generate explanations, potentially overlooking relevant features and introducing irrelevant ones in the process. Invertible neural networks (INN) are a promising alternative for interoperability (e.g. [96, 189, 62, 140, 97, 113]). INNs create counterfactual images by interpolating a sample in classifier space to change its prediction and then translating this adjustment back to the input space. This procedure is faithful to the model because the generative function synthesising counterfactual explanations is the analytic inverse of the forward function used for predictions.

2.3.4 Concept-Based Explanations

Concept-based explanations utilise abstract image attributes (concepts) linked to a prediction for model explanations (Figure 2.5). Several approaches introduce concepts into the explanation process in different ways. *TCAV* [108] requires users to manually define concepts after the model's training phase and prior to generating explanations. Concepts are mapped



prediction: bedroom
explanation: + 4.2 bed - 1.5 coffee-table + 1.3 sky - 1.3 sofa
- 1.0 drinking-glass - 0.9 television - 0.9 sconce
- 0.8 chair + 0.8 windowpane + 0.7 blind + 0.7 fan - 0.6 armchair
- 0.6 sink - 0.6 switch + 0.5 box - 0.5 plate - 0.5 ottoman
- 0.5 paper + 0.4 cushion - 0.4 tray + 0.4 bottle - 0.4 bowl - 0.4 bag
+ 0.3 chest-of-drawers - 0.3 curtain - 0.3 chandelier - 0.3 table + 0.3 clock
+ 0.3 pot - 0.3 wall - 0.2 telephone - 0.2 fireplace + 0.2 ceiling + 0.2 carpet
+ 0.2 book - 0.1 spotlight - 0.1 flower

Figure 2.5. Concept activation explanation: These explanations show which concepts were involved in a prediction. In this example, taken from Ramaswamy et al. [181], the prediction is explained as a linear combination of 37 concepts. The authors found that the number of presented concepts should be limited to a maximum of 32 to avoid overwhelming users.

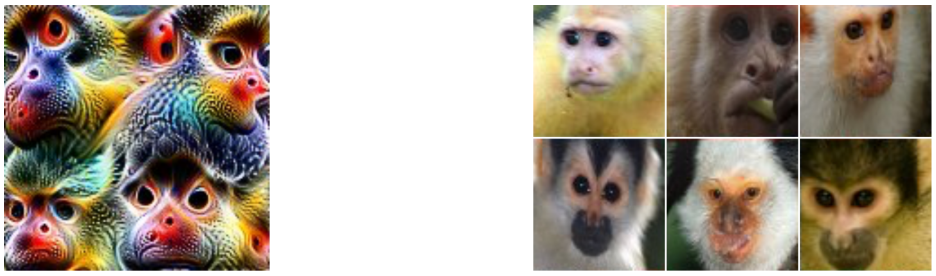
to high-level feature activations to generate explanations, and their similarity is determined by the dot product of the network's internal activations with a concept vector. Although this yields semantically meaningful concepts, it does not ensure the model's reliance on these for predictions, leading to potentially unfaithful explanations.

Later research used k-means [75] and non-negative matrix factorisation [263] to *automatically discover concepts* in trained models, which were presented as superpixels or prototypes, visualising them as superpixels [75] or prototypes [263]. This automation raises questions about the semantic value of such concepts. Although Ghorbani et al. [75] affirmed the consistency of discovered concepts, it is unproven that they are more meaningful than simpler explanation methods.

Concept bottleneck models (CBMs) [119] are an alternative that introduces concepts during training, requiring labelled concepts in training data, but only two datasets currently meet this requirement [246, 14]. CBMs also need training and input data of similar distribution to function effectively.

2.3.5 Feature Visualisations

Feature visualisations generate synthetic images to illustrate what activates a specific neural network unit (Figure 2.6). Generation methods use gradient ascent to refine a synthetic image, maximising targeted unit activation [165]. They try to isolate and highlight the elements that cause a unit's response [173], potentially revealing both local and global model insights. However, a complete understanding of the model is unattainable because of complex unit interactions and the tendency of users to only interpret units that express human-understandable concepts [24].



(a) Channel Visualisation **(b)** Example images

Figure 2.6. Feature visualisation: Visualisations of the patterns (a) that positively activate unit 400 in layer 40 of GoogLeNet [227] accompanied by maximally activating example images (b). Images are taken from the appendix of Olah et al. [173].

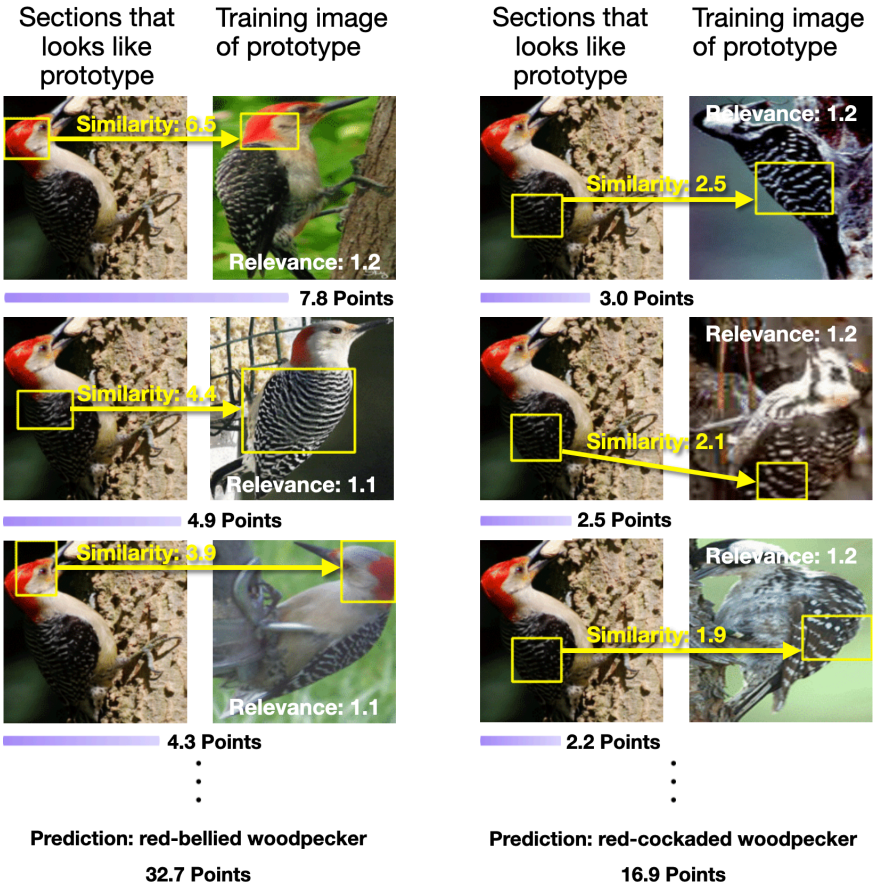


Figure 2.7. Prototypical explanation: ProtoPNet [39] compares input images to learned prototypes, calculating a label's relevance score based on similarity and prototype-label relevance. Summing these scores gives a prediction score, transparently combining prediction and explanation. The images and scores are sourced from the original study and rearranged for clarity.

2.3.6 Prototypical Explanations

Following the taxonomy of Guidotti et al. [82], the methods presented thus far fall into the category of "explaining black-box classifiers" (Figure 2.1). Another approach is to create more interpretable classifiers. For images, this can be achieved using image patches as prototypes. For example, BagNet [27] processes image patches separately and applies a linear classifier to each patch to obtain class evidence. Averaging the evidence produces the final class probabilities. ProtoPNet [39] dissects images to identify prototypical parts and combines evidence from these prototypes to make a final classification decision (Figure 2.7). Prototype-based explanations operate on a "this looks like that" analogy, ensuring faithfulness as the same prototypes underpin prediction and explanation—although this does not extend to their saliency maps [253]. However, a semantic gap between latent and input space similarity may limit their usefulness, as prototypes and input often diverge in appearance [111].

2.4 Necessity of Explanation Evaluation in User Studies

In recent years, scholarly discourse has increasingly noted a gap in rigorous evaluations of explanation methods [52, 240, 3, 170, 162]. Adadi and Berrada [3] found that only 5% of reviewed papers evaluated XAI methods and quantified their relevance. Similarly, Nunes and Jannach [170] observed that a substantial 78% of papers on explanations in decision support systems omitted structured evaluations, frequently relying on anecdotal "toy examples." This trend appears persistent, as recent findings by Nauta et al. [162] showed that from 2016–2020, approximately a third of papers relied solely on anecdotal evidence for evaluations, while only one-fifth involved user evaluations. Most evaluations are functional, using benchmark data or metrics to prove generalisability. This approach is effective for showing technical feasibility but lacks a formal definition of a correct or best explanation [170].

Kim et al. [111] showed a weak link between functional interpretability metrics and participant model understanding. This discrepancy highlights that even a robust formal foundation does not inherently translate to any practical utility or an increase in user understanding [52]. Currently, user studies stand as the most dependable method to incorporate human behaviour in the evaluation of explanations.

To date, they are rare, possibly due to the difficulty in creating realistic settings while avoiding overburdening participants [52, 240, 3, 170]. Their value is undeniable: Some prior studies showed that explanations may help find bugs, biases, or spurious correlations [123, 184, 5]. Others raised questions about the effectiveness of explanations in enhancing model understanding [5, 108], trust calibration [32, 103, 31, 41], and error under-

standing [203, 5], highlighting the need for human experiments to validate intelligibility claims.

2.5 A Taxonomy for XAI User Studies

We employed the *conceptual-to-empirical* approach [166] to systematically identify suitable study designs for evaluating explanations, avoiding the need for an exhaustive review of each study, and derived a taxonomy. This taxonomy was refined through an analysis of 52 publications, selected via a structured literature review using the empirical-to-conceptual method (search criteria and references detailed in Appendix B). We synthesised categories from previous XAI studies and integrated them with key empirical study literature. The final taxonomy, illustrated in Figure 2.8, covers the main features of different study designs. Of course, *"there is no standard design for user studies that evaluate forms of explanations"* [170]. However, this taxonomy provides guidance for researchers (including ourselves) to systematically plan studies. In the following sections, we categorise key aspects of explanation evaluation with human subjects into *task-related*, *participant-related*, and *study design-related* dimensions. (Note that this categorisation builds upon and extends the terminology introduced on pages 3-5.)

2.5.1 Participant Dimensions

Participant Types Mohseni et al. [150] distinguish between *AI novices* (typically end-users), data and *domain experts*, and *AI experts*. This distinction is vital as it impacts other task-related and participant-related dimensions. For example, Doshi-Velez and Kim [52], referencing the work of Neath and Surprenant [163], note that participants' expertise influences their cognitive strategies. However, studies by Kaur et al. [103] and Borowski et al. [25] suggest that AI expertise may not significantly affect the ability to understand and evaluate explanations. Yet, Feng and Boyd-Graber [66] found notable differences between experts and novices in their trust and use of explanations.

Recruiting Method and Number of Participants Recruiting difficulty tends to increase with the level of participant expertise needed [52], affecting both the suitable recruitment methods and the number of participants feasible for the study. While *crowdsourcing* platforms are effective for enlisting large numbers of novices, domain or AI experts are harder to find here. They need to be individually contacted and invited to, typically smaller, *online*, *lab*, or *field studies*.

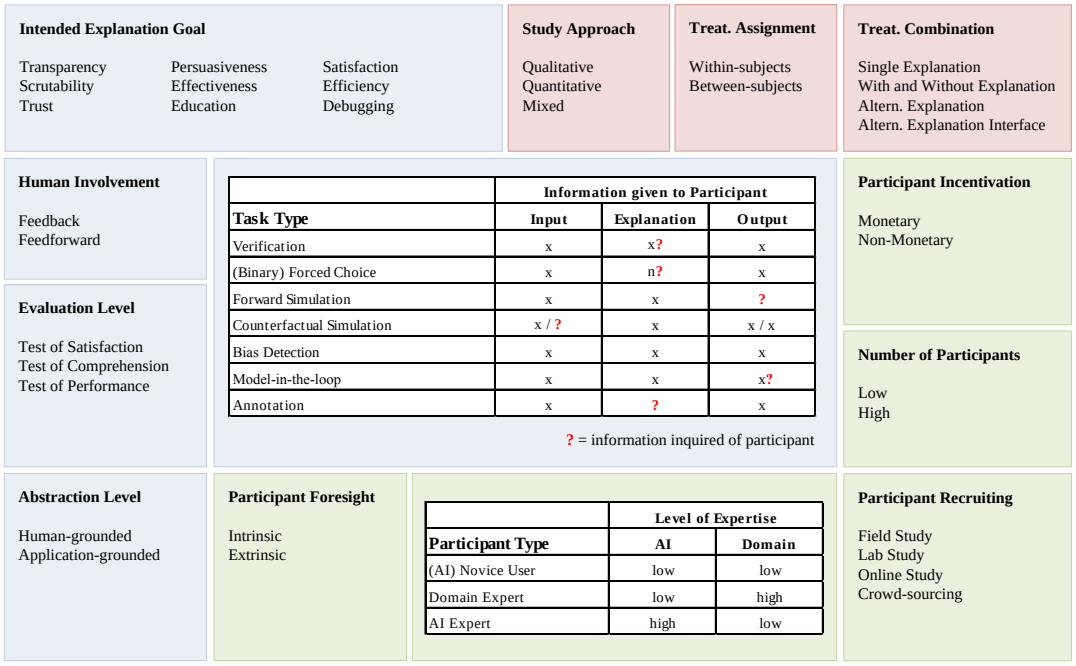


Figure 2.8. XAI User Study Taxonomy: Preliminary taxonomy of human subject evaluation in XAI based on the conceptual-to-empirical approach.

Participant Foresight Narayanan et al. [161], suggest two settings based on the anticipated impact of participant foresight on study outcomes. In an *intrinsic* setting, participants, generally novices, have similar levels of knowledge and rely solely on the information provided in the study. This greatly benefits internal validity. In contrast, an *extrinsic* setting allows for the use of additional, external knowledge, such as data domain or machine learning expertise. This can help participants to evaluate explanation quality or identify model inaccuracies, but it poses a risk to internal validity.

Incentivisation According to Sova and Nielsen [216], choosing the right incentives for a study depends on study length, task difficulty, and required participant expertise. Stadtmüller and Porst [221] suggest using a *monetary incentive* [221], but *non-monetary incentives*, such as gifts for paid employees are also effective [216, 178]. Porst and von Briel [178] found that participants might join a study out of altruistic motives, due to interest, or personal incentives, such as a promise made to the experimenter. Esser [61] points out that the combined benefits of all incentives should outweigh the perceived costs. Special incentivisation for high-quality answers is needed in crowdsourcing experiments, to prevent participants from rushing through them, potentially leading to less useful answers.

2.5.2 Study Design Dimensions

Study Approach and Treatment Combinations Study designs may adopt a *qualitative*, *quantitative*, or *mixed-methods* approach. In such studies, experimenters assign explanation methods as treatments to groups of participants. Based on our review and the work of Nunes and Jannach [170], four combinations are common. The first, *single treatment* provides no baseline treatment. The second, *with and without explanation*, establishes a baseline treatment were no explanations are provided. The third, *alternative explanation*, presents different explanations while keeping the user interface consistent. Finally, the *alternative explanation interface* method changes the user interface across treatments. It is important to note that the choice of explanation methods is influenced by the underlying model's input and output data, the specific machine learning problem it addresses, and its architectural design.

Treatment Assignment Study designs can also differ in their method of assigning treatments. *Between-groups designs* study the differences in understanding between groups of participants, each usually assigned to one treatment. In contrast, *within-subjects designs* study differences among individual participants who are assigned to multiple treatments.

2.5.3 Task Dimensions

Human Involvement Mohseni et al. [150] distinguish two main ways in which participants take part in testing explanations. The first, a *feedback* approach, involves participants assessing given explanations. Their feedback can be explicit, such as direct comments, or implicit, such as selecting one explanation over another. The quality of explanations is then assessed based on this feedback. The second, a less common *feed-forward* approach, does not involve presenting existing explanations. Here, participants create examples of what they consider reasonable explanations, which then act as benchmarks for algorithmic explanations.

Abstraction Level Doshi-Velez and Kim [52] split human evaluations into two types. *Application-grounded* evaluations happen in real-world settings and usually require participants with high expertise. Here, the value of explanations is measured by how well they fit the task at hand and key performance indicators, such as trust in decisions made or the count of cases handled. This setting offers high external validity but makes it difficult to ensure internal validity because of the many factors affecting the outcome. Obtaining field access is the biggest challenge for these experiments, making them less common.

Human-grounded evaluations, involve simplified or abstracted versions of application

scenarios, naturally reducing external validity. Internal validity is improved in this setting because it allows researchers to control more of the factors that could change the outcome.

Explanation Goals Explanations have various intended purposes, and the study scenario must match them to ensure the experiment's validity. Participants must receive appropriate guidance and incentives within the scenario to align with these objectives [239]. Drawing from earlier work [232, 170, 247], we identify nine primary explanation goals:

- *Transparency*: Make the system's workings clear.
- *Scrutability*: Allow users to challenge the system upon errors.
- *Trust*: Enhance or calibrate users' trust in the system.
- *Persuasiveness*: Convince users to act.
- *Satisfaction*: Improve user experience and ease of use.
- *Effectiveness*: Aid informed decision-making.
- *Efficiency*: Accelerate decision-making.
- *Education*: Enable user learning and extrapolation.
- *Debugging*: Assist in identifying and fixing system flaws.

In the case of multiple goals, their dependencies may be complementary, contradictory, or even unknown. For example, studies have shown that higher transparency can lead to misaligned trust [32], reduced trust [220], and lower satisfaction [31]. Random explanations, on the other hand, which are harmful to transparency, have been found to increase users' trust in a model (text-classier: [219, 124], image-classifier: [41]).

Evaluation Level A study requires a measurement to determine whether the explanation goals have been achieved. Hoffman et al. [92] describe three levels and corresponding metrics for explanation evaluation. *Tests of performance* measure the resulting human-XAI system performance. *Tests of satisfaction* focus on participants' self-reported explanation satisfaction and their perceived understanding of the system but cannot confirm the actual depth of understanding. *Tests of comprehension* evaluate participants' mental models and understanding of the system, typically through prediction tasks and generative activities. In this work, we set our focus on the latter since human understanding is the central element of definitions of interpretability found in the literature:

> "To interpret means to give or provide the meaning or to explain and present in *understandable* terms some concepts" — Gilpin et al. [78].

> "In the context of ML systems, we define interpretability as the ability to explain or to present in *understandable* terms to a human" — Doshi-Velez and Kim [52]

> "Interpretability of a model: the degree to which an observer can *understand* the cause of a decision" — Miller [146] based on Biran and Cotton [20].

In line with our HCI perspective, we define our metric of interpretability as:

> **Model Understanding**: The degree of alignment of the user's mental model with the machine learning model's operational processes. — Own definition

This definition addresses the explanation goals of *transparency*, *scrutability*, and *effectiveness* but does not consider the potentially conflicting goals of *trust*, *persuasiveness*, and *satisfaction*. This strict definition allowed us to create study designs with high validity, reproducibility, and adaptability while avoiding complex interactions between factors.

Task Type

We base our categorisation of tasks on the information provided to participants [52, 150] and discuss their ability to measure model understanding in more detail.

Verification Tasks involve participants reviewing the input, explanation, and output and then rating their satisfaction with the explanation.

Forced choice tasks build on this by having participants select the best explanation from multiple options [33, 31]. Both verification and forced choice mainly measure satisfaction with explanations, not model understanding.

Forward Simulation Tasks assess participants' ability to predict a system's output from given inputs and explanations [134]. This method was initially used to test the explainability of search engines and is based on the idea that a high model understanding should enable accurate predictions of its behaviour [160]. In *direct forward simulation* (direct fwd. sim.), participants predict a system's output without explanations for the current input, relying simultaneously on their understanding of relevant and irrelevant features and their combination [142]. Of the tasks presented here, it is the most comprehensive measure of model understanding but also the most challenging. *Proxy forward simulation* (proxy fwd. sim.), allows access to the explanation of the current input [31, 203]. This leaks some information about the prediction, but the task often remains challenging given the complexity of the non-linear feature processing of neural networks. In classification models, forward simulation can be compounded by a confirmation bias when the correct label appears obvious and

explanations are provided, leading participants to anticipate that the model is going to yield the correct prediction [111].

Annotation Tasks require participants to create a suitable explanation for a given input and output of a model. Assessing model understanding involves verifying the correctness of the provided explanation, which is particularly challenging in the case of images.

Counterfactual Simulation Tasks offer participants an input, an explanation, an output, and an alternate output and ask them to generate the necessary input changes to obtain the alternate outcome. This requires identifying which features could influence the model's prediction when added or removed. These tasks assess similar aspects of a participant's mental model as proxy forward simulation tasks. Manual creation of image counterfactuals is time-consuming and has not yet been applied in studies focusing on the understanding of image classification models.

Model-in-the-Loop Tasks show users the model's prediction for an input, requiring them to follow it or reject it as untrustworthy [31, 41]. They provide strong external validity by testing participants' ability to spot errors. However, there are several internal validity issues. Results for participants with high data domain knowledge provide no insights, as they can select the correct answer without relying on a prediction or explanation. Participants may also misjudge their domain expertise due to the Dunning-Kruger effect, influencing which prediction they accept [196]. Lastly, Kim et al. [111] showed that explanations introduce a confirmation bias, convoluting the measurement of "model understanding" and "trust" in one task.

Bias Detection Tasks ask participants to determine whether a model has biases by analysing inputs, explanations, and outputs [128, 184, 5, 108]. For simplicity, we refer to each relevant feature as a *bias* in this work [1]. A correct mental model of feature relevance is needed to fulfil this task. However, whether users understand how feature combinations influence the prediction is not assessed. This task is easier than model-in-the-loop and forward-simulation tasks but has lower external validity, as it provides a less comprehensive measure for model understanding. In *proxy bias detection* (proxy bias det.), model understanding is measured by asking participants whether a model is fit for deployment [4]. The underlying assumption is that participants would not deploy a biased model. However, in our pilot studies, we found that this assumption may not be warranted as some participants did judge a model "good enough" for deployment even though they were aware it was biased. In *direct bias*

[1] Note that bias has a more complex yet overlapping meaning in ethical and responsible AI.

detection, participants need to directly identify biases, for example, by rating feature relevance [108]. This task poses a challenge for internal validity since it requires controlling relevant and irrelevant features to allow accurate validation of participants' answers.

2.5.4 Dataset

In image classification, certain tasks may not be a good fit for every dataset.

Unmodified Images Natural images [108, 31, 203, 25, 111, 181], especially those depicting animals [107, 5, 79, 7], are the most common choice for user studies. Model-in-the-loop tasks require a dataset where assigning the correct label is challenging without the help of a model, even for participants with domain knowledge. For images of everyday objects or animals, this may not be the case. Domain-specific datasets, such as medical imagery [142], can raise the level of difficulty but may limit the availability of participants. Hence, they are less commonly used.

Modified Images Direct bias detection requires a model with ground-truth feature importance [257], which effectively excludes unmodified images because they contain too many relevant features. A common approach is to deliberately bias natural image datasets to introduce a single artificial bias, such as a watermark [108] or biased background [184]. The BAM dataset by Yang and Kim [257] is designed to help with this process. It overlays a labelled foreground object on top of a natural background image while introducing correlations between the foreground and background. However, such modified images appear artificial, eliminating a major advantage of natural images for user studies. For example, when a dog is randomly placed on a bamboo forest background participants can easily infer that the background is a relevant feature. Furthermore, this modification approach is limited to introducing a single feature with known relevance, leaving other relevant and irrelevant features unaddressed. This is problematic since when participants identify irrelevant features as relevant, it indicates a lower level of model understanding.

Synthetic Images Synthetic datasets offer better control over features, improving internal validity. CLEVER [101] and CLEVER-XAI [11] are examples of such datasets, but they still introduce too many features with control over their presence but not their relevance. This makes them ill-suited for bias detection user studies. We created the Two4Two dataset to fill this gap. Its data-generating factors can be correlated with the target class, thereby creating arbitrarily strong biases. After model training, we can verify feature relevance by removing each bias individually and quantifying prediction changes. The dataset and its use in one of our studies are detailed in Chapter 4.

2.6 Our Studies in Relation to the Evolving State of the Art

In this section, we systematically review studies, including our own, that measure the impact of explanation methods on users' model understanding of image classifiers. A structured overview in Figure 2.9 organises studies by their publication date (after peer review). We show the contribution of each study to the field in chronological order. This allows us to draw a broader conclusion about the most effective explanation methods and critically examine common methodological limitations to identify future best practices for research (Section 5). We acknowledge the relevance of studies that assess explanations on other data types but our focus is exclusively on images, as we assert that results do not generalise between data types.

2016 — Evaluation of Surrogate-Based Feature Attribution Ribeiro et al. conducted a user study (N = 27) with a binary classifier distinguishing wolves from huskies, biased towards snow in backgrounds [184]. In the within-subjects study, an example-based explanation, which was always shown first, served as the baseline, showing eight correct and two incorrect predictions. In the treatment condition, the same examples were supplemented with LIME saliency maps. The measure of success was whether participants mentioned the background as a feature (direct bias detection task, chance of random success approx. 50%). While only 12 of 27 participants mentioned the bias in the baseline, 25 mentioned it after the treatment, indicating LIME benefited users' model understanding. The study design could be strengthened by randomising the sequence of explanations shown, reducing the risk of ordering effects. Furthermore, ensuring that participants are not repeatedly exposed to the same images could help isolate the impact of the explanation technique from increased familiarity with the images. Finally, testing the results for statistical significance, such as using a McNemar test, would enhance the robustness of the findings.

2018 — Evaluation of Gradient-Based Feature Attribution Kim et al. [108] conducted a similar direct bias detection study (N = 50) with two gradient-based feature attribution methods [226, 215]. They intentionally introduced bias into the data by imprinting the class label on each image, enabling them to train two models: one that exploited this bias and the other that disregarded it. The study employed a mixed design where participants interacted with one model as a between-groups factor and engaged with both explanation techniques as a within-groups factor. They assessed the importance of the biased feature versus the rest of the image on a 10-point scale (main dependent variable) and their confidence on a 5-point scale. Rating the bias higher than the rest for the biased model indicated correct model understanding. The reverse was true for the unbiased model. Participants performed marginally but not significantly better than the calculated chance level of success (50%),

Publication (Study Number)	Explanation techniques (Conditions)	Baseline (Conditions)	Task (Random Chance)	Dep. var.	N / Cond. (Design)	Result
Ribeiro et al. [184] (III)	Saliency Maps w/ Pred. [184] (1)	Pred. (1)	direct bias det. (≈ 1/2)	N correct	27/2 (within)	SM > Base.
Kim et al. [108]	Saliency Maps [226, 215] (2)	None (0)	direct bias det. (50/100)	Relevance rating	50/4 (mixed)	SM not helpful
This work (I)	Saliency Maps w. Pred. & Label w/ and w/o. Scores [12] (2)	Examples w/ Label & Pred. (2)	direct fwd. sim. (1/2)	% correct	64/4 (between)	SM > EXPLS
Buçinca et al. [31] (I)	Examples w/ Label (2)	None (0)	proxy fwd. sim. (1/2)	% correct	183/2 (within)	EXPLS ≈ EXPLS
Buçinca et al. [31] (II)	Examples w/ Pred. (2)	Input w/ and w/o. Label (2)	model-in-loop (1/2)	% correct	102/3 (between)	EXPLS > Base.
Shen and Huang [203] (II)	Saliency Maps [68, 202, 215] (1)	Input w/ Label (1)	proxy fwd. sim. (1/5)	% correct	105/2 (between)	Base. > SM
Adebayo et al. [5]	Saliency Maps w/ Label [208, 215, 226] (3)	None (0)	proxy bias det. (—)	Certainty rating	54/5 (within)	SM not helpful
Borowski et al. [25] (I)	Feature visualisation [173] (1)	Guessing (1) & Examples (2)	direct fwd. sim. (1/2)	% correct	10/3 (within)	EXPLS > Vis.
Borowski et al. [25] (II)	Feature visualisation [173] (1)	Examples of feature (1)	direct fwd. sim. (1/2)	% correct	23/2 (within)	EXPLS > Vis.
This work (II)	Counterfactuals & Concepts (2)	Examples w/ Scores (1)	direct bias det. (1/5)	N correct	240/3 (between)	(EXPLS ≈ CF) > CON
Mertes et al. [142]	Counterfactuals [142] (1)	Saliency Maps [184, 12]	proxy fwd. sim. (1/2)	% correct	118/3 (between)	CF > SM
Ramaswamy et al. [181]	Concepts [119] (3)	Examples w/ Pred. (1)	mixed fwd. sim. (1/4)	% correct	125/4 (within)	CON ≈ EXPLS

Figure 2.9. Extensive comparison of reviewed studies: Example-based explanations, often regarded as mere baselines, are surprisingly difficult to surpass.

suggesting that explanations barely increased model understanding. No significant difference in confidence between correct and incorrect answers was observed, indicating that saliency maps may be more confusing than helpful.

2020 — Study I: Evaluation of Gradient-Based Feature Attribution Both prior studies focused on bias detection using classifiers with simple biases, such as background [184] or image captions [108]. Although insightful, these studies were secondary to the main objectives of their publications. Our first study (N = 64), detailed in Chapter 3, set the evaluation of an explanation technique as its main objective. It focused on saliency maps generated by LRP [12] and pioneered the challenging task of direct forward simulation in interpretable vision. Participants predicted whether a CNN would recognise a specific object in an image (50% chance of random success.) The proportion of correct answers served as the main dependent variable. Performance was compared to an example-based baseline. In our design, the main factor was whether example images were supplemented with saliency maps. It also included classification score visibility as a second factor in a $2x2$ full-factorial between-groups design. Consequently, half of the participants were shown supplementary saliency maps, while others were not, and half of each received detailed CNN classification scores. Counterbalancing task image sequences avoided ordering effects. Overall, attribution methods did not substantially increase participants' model understanding. However, they gave rise to a very small, but significant, performance improvement. Classification scores did not increase model understanding. We asked participants to identify relevant and irrelevant features to justify their predictions. Upon analysis, we found that the saliency maps caused the participants to concentrate on the highlighted features. However, it is uncertain whether this caused them to overlook other important attributes not emphasised by the maps. This observation suggested that it would be beneficial to enhance our study design using a model with known but subtle feature importance, enabling direct comparison with actual biases. This was implemented in Study II.

2020 — Evaluation of Example-Based Explanations Buçinca et al. [31] conducted two studies comparing inductive and deductive example-based explanations. The first study (N = 10) had a within-subjects design where participants used the input, the ground truth, and an explanation for a proxy forward simulation task (50% chance of random success.) The proportion of correct answers served as their main dependent variable. No significant differences in model understanding were measured. The second study (N = 23) used a between-groups design and a model-in-the-loop task (50% chance of random success), with explanations supplementing the model output instead of the ground-truth label. In addition, the study introduced two additional baseline conditions. In the initial baseline, participants

viewed only the input, assessing their performance without the model's help. The second baseline added the model's prediction, evaluating their performance with model assistance. This design effectively quantifies the impact of model assistance on performance and the extent to which explanations contribute to model understanding, as measured by the resulting incremental performance increase. Displaying predictions resulted in a significant performance improvement. Both explanation techniques increased performance significantly over both baselines. Since there was no significant difference between the two techniques, this study indicated that both example-based explanation techniques equally increased users' model understanding.

2020 — Evaluation of Perturbation-Based and Gradient-Based Feature Attribution Techniques The second study[2] by Shen and Huang [203] (N = 105) used a variation of a proxy forward simulation task and a between-groups design. Participants were told that the model's prediction was incorrect and were given five possible labels that the model might have falsely predicted (chance of random success 20%). The proportion of correct answers served as their main dependent variable. In the baseline condition, participants were only shown the input image. The three treatment conditions each displayed a different saliency map, generated using either extremal perturbations [68], Grad-Cam [202] or SmoothGrad [215]. These explanations were detrimental to participants' performance. Consequently, this study concluded that feature attribution explanations were harmful to model understanding.

2020 — Evaluation of Bug Detection Capabilities of Gradient-Based Feature Attribution The study by Adebayo et al. [5] (N = 54) used a proxy bias detection task. Participants used a 5-point scale to rate whether they would recommend selling a model to an external customer. Since the rating scale's midpoint was marked as "unsure/maybe", it measured confidence but not success [3]. This rating was the main dependent variable. In this within-subjects design, participants were shown saliency maps generated with the vanilla gradient [208], Integrated Gradients [226] or SmoothGrad [215]. They viewed these explanations for five models, four of which had bugs. The baseline model was correctly trained and received standard input images. The first faulty model was trained on random labels, leading to the learning of incorrect features. The second was biased by spurious correlations in its training data, akin to the approach by Ribeiro et al. [184]. The final layer of the third faulty model was randomised, hindering proper feature combination. The fourth faulty model was correctly trained on out-of-distribution images, resulting in a lack of relevant fea-

[2] We do not review the first study as its design is less elaborate and showed similar results.

[3] Ratings could be converted to a measure of success, but the results were not analysed this way.

tures for classification. Significance in differences of ratings across conditions was assessed using a Wilcoxon signed-rank test. It was complemented by paired t-tests assessing differences within participants under different model conditions [4]. Participants were significantly less likely to recommend the faulty models, except for the out-of-distribution model, which received similar ratings to the correct model. Observations during the study showed that participants paid little attention to saliency maps and instead focused on predictions. Furthermore, an analysis of saliency maps showed that they were similar for both correct and incorrect predictions, which led the authors to question the usefulness of saliency maps. A baseline condition without saliency maps could improve this study design and substantiate this conclusion.

2021 — Evaluation of Exemplary-Based and Synthesised Feature Visualisations In both studies by Borowski et al. [25], participants compared two images to determine which was more likely to activate the network (chance of random success 50%). The proportion of correct answers served as the main dependent variable. Although the authors classify their task as a direct-forward prediction task, it is important to clarify that participants predicted the activation of a hidden layer, not the output layer. In the first study (N = 10), three conditions were tested: random guessing with no information provided, example-based explanations displaying minimally and maximally activating images for a feature, and synthetic feature visualisations by Olah et al. [173]. This rigorous design included random guessing to ensure that task images did not have a bias leading to a success rate other than random chance [5]. Example-based baseline explanations improved model understanding compared to random guessing and synthetic visualisations. The result was unexpected as example images were included as a baseline, not a treatment. In the second study (N = 23), random guessing could be removed safely, and three factors were added: participants' machine learning expertise (Lay, Expert), number of examples, and presentation scheme (Max1, Max9, Min+Max1, Min+Max9). Schemes varied in the number and types of examples that were presented: one maximally activating example (Max1); one minimally and one maximally activating example (Min+Max1); nine maximally activating examples (Max9); and nine minimally and nine maximally activating examples. This refined design confirmed that example-based explanations lead to higher model understanding than feature visualisation, even across various levels of participant expertise. This result aligns with the findings of Kaur et al. [103], which suggest that expertise does not significantly influence the effectiveness of explanations, contrary to common belief. Additionally, the study demonstrated

[4] A Friedman test could provide a more robust and comprehensive analysis, as it is better suited to handling the repeated measures design and allows for the analysis of multiple factors and their interactions.

[5] The main author clarified this upon request.

that providing more examples correlates with improved model understanding. From our perspective, this study further indicates that example-based explanations must be included as a baseline whenever possible.

2022 — Study II: Evaluation of Example-Based, Concept-Based, and Counterfactual Explanations Our second work, detailed in Chapter 4 uses a direct bias detection task. Participants evaluated five attributes as either relevant or irrelevant (with a 50% chance of random success). The number of correct answers served as the main dependent variable. To address the need for a model with known feature importance, we propose a synthetic dataset, Two4Two, depicting two abstract animals. It allowed us to bias an invertible neural network (INN) arbitrarily. We designed a custom example-based baseline explanation technique that allows users to inspect all attributes that potentially predict the target class. An initial study (N = 50) with a within-subjects design used only this technique. It confirmed that identifying the model's main feature and the shape bias is relatively easy while finding the colour bias is difficult. The main study (N = 240) used a between-groups design. It compared our baseline against two state-of-the-art explanations: automatically discovered concepts [263] and faithful counterfactual interpolations generated with an invertible neural network. Surprisingly, no method outperformed our baseline, while concept-based explanations performed significantly worse. Even though counterfactuals were more helpful in discovering the strongest bias in the model, some participants rated the relevance of the background incorrectly, as slight changes in the interpolations were still sufficiently salient to be considered relevant.

This work builds on the methodological improvements of Study I and uses an even more rigorous study design. In addition to these findings, this work has contributed a benchmark dataset, benchmark model, and benchmark explanation technique. Section 4 outlines these findings and contributions in greater detail.

2022 — Evaluation of Counterfactuals and Feature Attribution The study (N = 118) conducted by Mertes et al. [142] used a proxy forward simulation task and a between-groups design. Participants predicted whether a model would assign the label Pneumonia to an X-ray image (the chance of random success was 50%). The proportion of correct answers served as the main dependent variable. Task images were supplemented with a slider. In the first condition, moving the slider created a linearly interpolated counterfactual image. The two feature attribution conditions overlaid the image with a saliency map either generated with LIME [184] or LRP [12]. The results show that model understanding was significantly higher with counterfactuals than with the feature attribution methods. This study did not consider an example-based explanation but chose saliency maps as their baseline.

2023 – Evaluation of Concept-Based and Example-Based Explanations The study (N = 125) conducted by Ramaswamy et al. [181] evaluated concept-based explanations in three variations, using 8, 12, or 32 concepts and example-based explanations as the baseline. Concepts were generated using the Concept Bottleneck Model by Koh et al. [119]. Participants used each explanation technique to predict which of four abstract labels would be assigned by the model (within-subjects design, chance of random success: 25%.) This task is a crossover between direct and proxy forward simulations. Participants first had to select the concepts present in the task image. If this selection was correct, the concept-based explanation would reveal the correct answer by assigning the highest score to the correct label. The proportion of correct answers served as the main dependent variable. The results show that model understanding did not differ between example-based explanations and concepts, while participants were slower in using concepts. However, no assessment of statistical significance is reported.

2.7 Best Practices for XAI Human Evaluations

As demonstrated in the previous section, study designs in computer vision have significantly advanced over time. We propose several best practices based on our experiments and extensive literature review, to further enhance the rigour and reliability of future studies, ensuring that they are aligned more closely with accepted practices in HCI and the empirical sciences.

Clear Explanation Goal Researchers should clearly define the explanation goal their experiential hypothesis targets from the outset of their studies. All design decisions need to align with this focus. All methods should be treated equally and fairly. For example, they must be presented in a similar fashion and with a comparable amount of visual information.

Focus on Actual Challenges We recommend focusing on model understanding or building appropriate trust in a model, which are themes for grand research. While increased user satisfaction and a method's persuasiveness are important, they are no measure of model interpretability. As such, they may not advance the goal of making AI more fair or ethical.

Participants Participants should be placed in a context where the evaluated explanation goal matters and is relevant to them. Application-grounded evaluations often naturally provide such a context. Human-grounded evaluations are a valuable alternative for researchers, such as ourselves, who have limited field access or concerns about internal validity. However, careful attention is crucial when setting the context, particularly when preparing participants for the experiment. Click workers trying to finish another 2$ task with unclear

instructions may choose to just move on to the next task quickly rather than caring how an image classifier works. The Appendix D.1 and D.6 provide details on our instructions and screening mechanisms, which may serve as a source of inspiration.

Preregistration and Statistical Rigour A statistical analysis plan should be finalised before any participant engagement and, ideally, automated for execution after data collection. This pre-emptive clarity prevents methodological shortcomings, such as inappropriate measurement scales, conditions, or significance tests. Universities often offer statistical consultation to vet study designs and analyses in advance. Additionally, and whenever possible, finalising and preregistering[6] your study protocol is highly recommended.

Mixed Methods Gathering qualitative data along with quantitative data offers considerable benefits. This aids the researcher's understanding of the participants throughout the experiment and can be used as an additional validity check. Moreover, it also provides deeper insight into participants' reasoning with explanations.

Baseline Comparison We recommend a comparative study design with several alternative explanations. As we have shown in our review, omission of a baseline may lead to limited empirical insight ([108, 5]), while including one can lead to unexpected but insightful results (Study I–II, [25].) One evaluated method should serve as a reasonable baseline; we suggest example-based explanations for this purpose. We encourage the community to take on the challenge of surpassing our simple example-based method presented in Study II.

[6]Example of a preregistration: `https://aspredicted.org/blind.php?x=/62X_15J`

Chapter 3

Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study

3.1 Research Questions and Publication

In this chapter, we report in detail on our Study I, an online between-group user study designed to evaluate saliency maps (generated with LRP [12]). We measured model understanding by asking participants whether the model would correctly predict labels for 14 images (direct forward simulation). In a $2x2$ design, the presence and absence of saliency maps and CNN prediction scores served as our two independent variables.

Research Questions

- **RQ 6** — Do saliency maps allow users to develop a better understanding of how the CNN model classifies a class of images?
- **RQ 7** — Do scores influence the participants' ability to predict the system outcome on the task images?
- **RQ 8** — When saliency maps are present, do users pay attention to features differently?

We answer all research questions with the user study. Our results indicate that when saliency maps were available, participants answered correctly more frequently than when they were absent (60.7% vs. 55.1%, p = 0.045). However, the overall performance was generally low even with the presence of saliency maps. Our data also indicates that saliency maps influenced people to notice saliency-maps-features. However, it is unclear whether

such explanations deter them from considering other attributes that are usually not highlighted by saliency maps.

Contributions

- The first empirical evaluation that compares example-based to feature-attribution explanations.
- A more rigorous comparative study design in comparison to prior work.
- Demonstrating that forward simulation is a comprehensive but challenging task for measuring model understanding.
- The finding that feature-attribution methods provide little benefit for model understanding over simple example-based explanations.
- Findings about the limitations of saliency maps and local explanations in general.
- The finding that prediction scores do not benefit model understanding.
- The finding that saliency maps can help participants to notice localised saliency-maps-features, but also draw them away from considering general attributes that are usually not highlighted.

Publication This study has been published as a full paper at the 2020 Annual ACM Conference on Intelligent User Interfaces (IUI)[1]. At the time of writing, it has been cited 219 times in the four years since its publication.

> Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (IUI). ACM, 263–274. doi: 10.1145/3377325.3377519

The sections of this chapter are taken from the original publication. However, to avoid repetitions, the abstract, introduction, related work section, and conclusion have been removed. Their contents are covered in significantly more detail in Chapter 1–2 and 5.

Author Contribution The study apparatus is the contribution of Ahmed Alqaraawi (interface implementation, model training, and provisioning of explanation techniques). Once aware that Ahmed Alqaraawi, and Martin Schuessler each separately planned a user study

[1] A paper presentation had been scheduled, but the conference fell victim to the COVID-19 Pandemic: https://iui.acm.org/2020/index.html

evaluating saliency maps, Enrico Costanza suggested that they join their efforts. He closely supervised the entire process and contributed to all aspects of this work. Martin Schuessler refined the study design and interface. He also conducted the data analysis.

In a series of pilot studies, Martin Schuessler compared whether saliency maps generated with LIME [184] or LRP [12] are more informative to users, identifying LRP as the evaluation candidate for this study. Martin Schuessler and Ahmed Alqaraawi designed an example-based baseline explanation method that showed correct and incorrect predictions. During the study, Ahmed Alqaraawi and Enrico Costanza handled participant recruitment and data collection. Martin Schuessler and Ahmed Alqaraawi worked on the qualitative analysis. Paper writing was a shared effort.

3.2 User Study Design

We designed a between-group online study to evaluate whether saliency maps can help users understanding of a highly complex CNN used for multi-label image classification. In the multi-label image classification problem, an image can contain multiple objects. For example, the assignment of the labels "*horse, train*" is considered correct if both, a horse and a train are visible in the image. We choose this problem because in this context, saliency maps have the potential to highlight specific parts of the image that correspond to one label, as well as parts that correspond to alternative labels.

The study included two independent variables that varied between groups, with a full factorial design. Both were related to the amount of information shown to participants: *presence of saliency maps* and *presence of classification scores*.

A screenshot of the experimental setup is shown in Figure 3.1. In the following sections, we lay out a more elaborate description of the study. At this point, it is essential to point out that we needed to strike a balance between the number of participants, the duration of the study and the variation of experimental factors.

3.2.1 Model and Evaluated Methods

Dataset, CNN Model Architecture and Training

Various public datasets, algorithms and configuration options exist for the multi-class image classification problem. We used the PASCAL Visual Object Classes (VOC) dataset (19714 images), because of its popularity, and its limited number of classes (20).

Additionally, we used the Keras library for Python, starting from an existing Keras model trained on the ImageNet dataset [45], utilizing the VGG16 architecture [208].We then fine-tuned the model on the train-val part of the PASCAL VOC 2012 dataset [64], achieving an

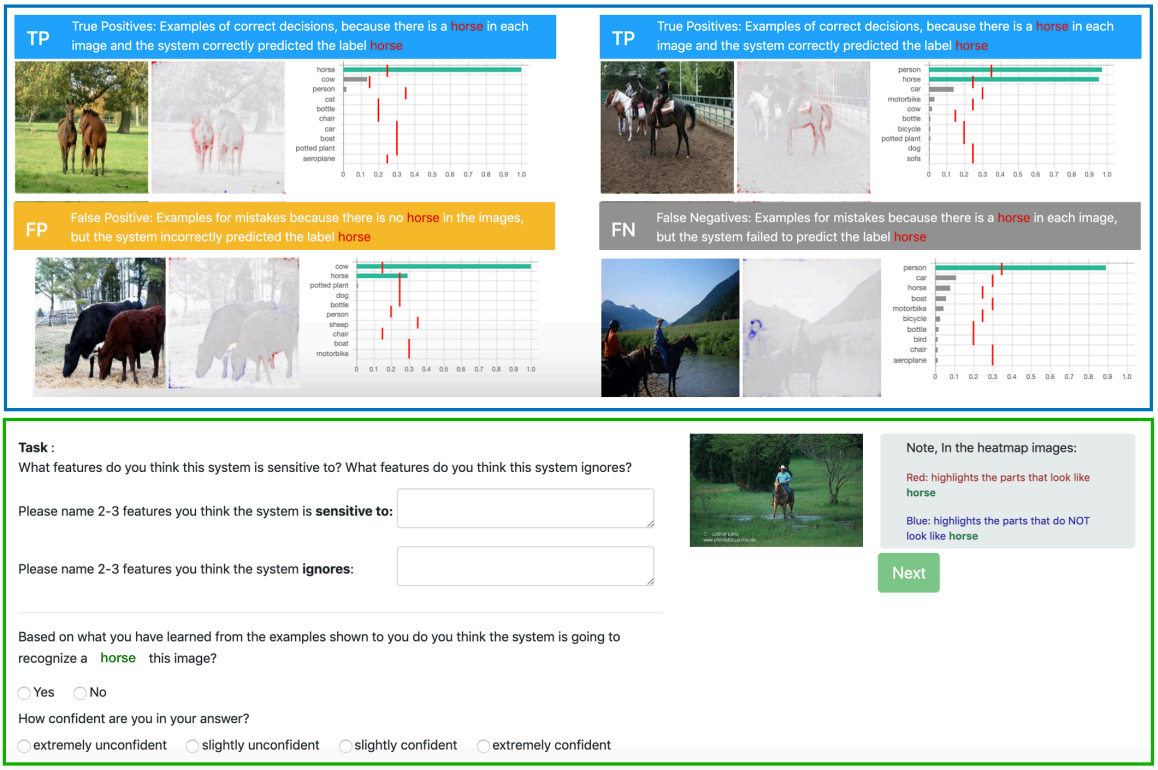


Figure 3.1. The interface: A demonstration of examples and questions are displayed in the *Saliency-map and Score* condition

Average Precision (AP) score of 0.91 on the training-set and 0.73 on a the validation-set. On a hold-out test-set (the PASCAL VOC 2007 test data [63]), the AP was 0.74. We did not train the model to reach state of the art performance. This was an intentional design choice to understand how explanation techniques could facilitate user understanding about the strengths and limitations of the model.

Saliency Maps and Scores Generation

A variety of algorithms have been proposed for generating saliency maps (cf. Section 2.3.2). In our pilot studies, we investigated two popular implementations: LIME [184] and LRP [12]. With LRP, saliency maps are not restricted to super-pixel patches but highlight contours of objects, which was preferred by most of our pilot study participants. For this reason and to simplify our setting, we chose to focus on the LRP algorithm only. Concretely, we used the α-β propagation rule [12] with $\alpha = 2$ and $\beta = 1$. Figure 3.2a shows a **true positive (TP)** example, where the model correctly predicts a train. The saliency map suggests that the red part of the image containing the rail supports the classification of this image as a train. Figure 3.2b shows a **false positive (FP)** example where the system *falsely* predicts a train. The red part of the image contains what *looks like* a rail. They support the classification of this image as a train. The blue parts are against this classification.

(a) True Positive (True Positive) image for *"train"*

(b) False Positive (False Positive) image for *"train"*

Figure 3.2. Examples of a saliency map explanations for the label *"train"*: The TP on the left highlights the contours of the lines below the train. A possible interpretation is that the CNN has learned to recognise trains when rails are present. A possible interpretation of the FP on the right is that edges in the lower part appeared similar to rails, which could explain this error.

Score Thresholds Since an image in the PASCAL VOC dataset can contain multiple objects, for each object class, the CNN computes a classification score between 0 and 1. Hence, a threshold needs to be defined so that the score can be translated into an outcome: *detected* when the score is above the threshold, or *missed* otherwise. We calculated threshold values for each class (e.g. horse, cat) because the CNN performs differently across classes. In particular, we obtained each threshold by maximising the F1-score for the class on the training dataset. In Figure 3.1, the small vertical red lines represent these selected thresholds.

Presentation

The interface of the study (Figure 3.1) was implemented as a Web application, using HTML5 and Python with the Django framework. We served the application from a standard Web server. The view-port of the participant browser window needed to be at least a 1000px wide and 600px high during the study.

3.2.2 Tasks

We gave our participants the task to predict the classification outcome of the CNN described above for a fixed set of 14 task images from the hold-out test set. More specifically, for each task image, we asked them to list 2-3 features they believe the system is sensitive to and 2-3 features the system ignores. We then asked participants to predict whether the system will recognise an object of interest ('cat' or 'horse') in the given task image (strict forward simulation - cf. Section 2.5.3). We also asked them to rate their confidence in their forecast on a 4-point forced Likert item. Figure 3.1 depicts the interface for one task

image (with a reduced number of example images). Half of the participants started with images of *horses*, while the other half, began with images of *cats*. To increase participants engagement in the study, in addition to an £8 payment for their time, participants received an additional performance-based bonus of £0.5 for each correct answer as an incentive. Seven task images were concerned with the class *"cat"* and another seven with the class *"horse"*. For each task image, participants were shown 12 example images from the CNN training set to inform their judgement. All participants worked on the same task images and were shown the same example images.

Selection of Example Images

We selected the example images for every task image from the PASCAL training set, based on their cosine distance from the task image in the embeddings space generated from the penultimate layer of the network. The assumption was that user understanding might benefit from looking at visually similar images. Showing the outcome of the classifier (i.e. TP, FN and FP) for the examples has been found to be important for the utility of explanation techniques [124]. For this reason, we sampled examples of different outcomes for each task image:

- 6 examples of True Positives (TP), where a label had been correctly assigned;
- 3 examples of False Negatives (FN), where the CNN had failed to assign the label;
- 3 examples of False Positives (FP), where the CNN had incorrectly assigned the label.

We also based our decision, regarding the number of shown examples, on experience from pilot studies. We had noticed that if we presented too many examples, participants were likely to only look at a random subset of them. At the same time, if the number was too low, there was a risk that not enough information was made available to participants. For this study, we selected 12 as a compromise. We also noticed that the saliency maps of TP examples are more informative than FN and FP. Thus we decided to show more TP than FN or FP examples.

Selection of Task Images

We intended our study to be no longer than 40 minutes to avoid fatigue effects. This design choice limited the possible number of task images. Consequently, we had to choose between sampling from a variety of classes or sampling from a subset of classes. In our pilot studies, participants found predicting model behaviour very confusing when the class in question was continually switching. Furthermore, the more classes they had to reason about the more challenging the tasks became, because they were not able to "learn" much about the model's

behaviour regarding a specific class. We also wanted to capture a variety of cases where the model had given correct as well as incorrect output. For these reasons, we decided to limit our experiment to two classes but included three TP, two FN and two FP for each class.

We drew task images randomly from the hold-out test dataset, with the constraint of having a mid-range classification score. In our pilot studies we had found that images with a low classification score (close to the threshold) were almost unpredictable for participants, while images with a high score were easily predictable. Consequently, we chose to sample from the middle, as we expect to see the most performance variation this way.

3.2.3 Conditions

The study included the following two independent variables:

Presence of Saliency Maps This factor had two levels: shown or omitted. When shown, the saliency map for the relevant class was displayed next to each example image. It is important to note that saliency maps were not shown for the task image but only for the examples (strict forward simulation task).

Presence of Classification Scores This factor also had two levels: shown or omitted. When shown, a bar chart of the top 10 classification scores was displayed next to each example image. Classification scores produced by the CNN are the default sources of explanatory information on the instance level. Hence, we aimed to investigate whether visualising this additional numerical information would outperform, compliment or interact with the presence of saliency maps.

The two independent variables were combined in a full factorial design, resulting in the following four conditions:

- Saliency maps not shown and scores not shown (Baseline)
- Saliency maps not shown and **scores shown**
- **Saliency maps shown** and scores not shown
- **Saliency maps shown** and **scores shown**.

Figure 3.1 illustrates the **saliency maps shown** and **scores shown** condition. In other conditions, the interface looked the same, except not showing the saliency maps or scores.

3.2.4 Participants

We recruited 64 participants (16 per condition) through Prolific [2], an online crowdsourcing platform. For the sake of data quality, we required participants to have an approval rate above 95% on the Prolific Academic platform, have normal or corrected to normal vision, and to be fluent in English. Moreover, we also made it mandatory for participants to be above 18 years of age and to have a technical background (i.e. a degree in computing or engineering), because of the technical concepts used in our study (i.e. neural networks, classification outcomes, scores, image pixels).

3.2.5 Procedure

After providing informed consent, each participant went through a short tutorial providing the necessary background about the experiment as well as clear instructions for using the system. The tutorial included examples of how the model classified a specific image and clear definitions of TP, FN and FP. We presented participants who belonged to conditions that would show saliency maps with additional information and examples that described this explanation technique and how they can be interpreted (See Appendix C.2). Similarly, participants assigned to a condition showing scores received additional advice on their interpretation.

Upon completion of the introduction, participants commenced completing their 14 tasks. At the end of the study, we gave them feedback for each task images and showed them their earned bonus.

3.3 Results

3.3.1 Outcome Prediction Accuracy

We were interested in investigating the effect that the presence of saliency maps and scores has on the ability of participants to forecast the CNN classification outcomes of images. We based our performance assessment on the percentage of correct forecasts per participant. We summarized the data in Figure 3.3. A Shapiro-Wilk test revealed that the percentage of correct forecasts within groups were approximately normally distributed ($W = 0.957, p = 0.027$). A Levene's Test showed performance variances between groups were similar ($F_{(3,60)} = 0.156, p = 0.925$).

A two-way independent ANOVA revealed a statistically significant main effect of the presence of saliency maps on the performance ($F_{(1,60)} = 4.191, p = 0.045, \eta^2 = 0.063$).

[2] https://prolific.ac/

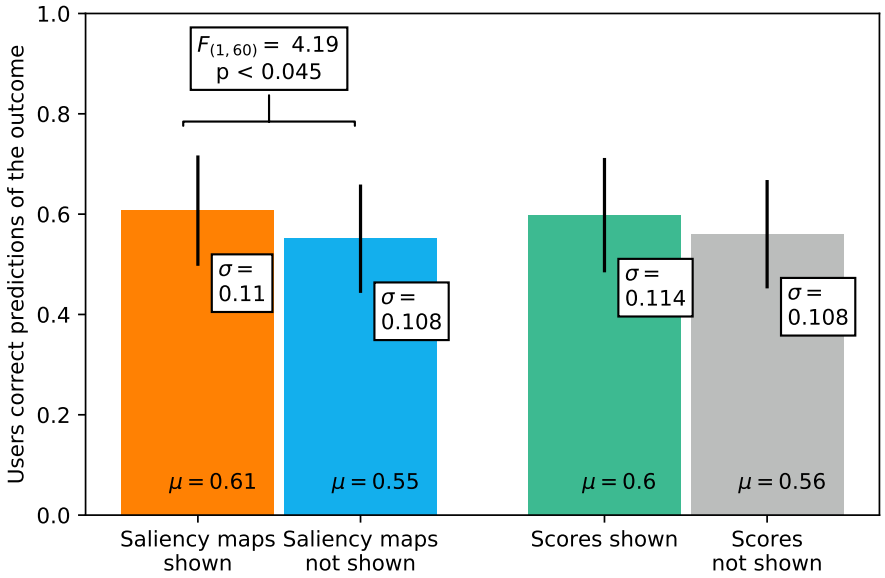
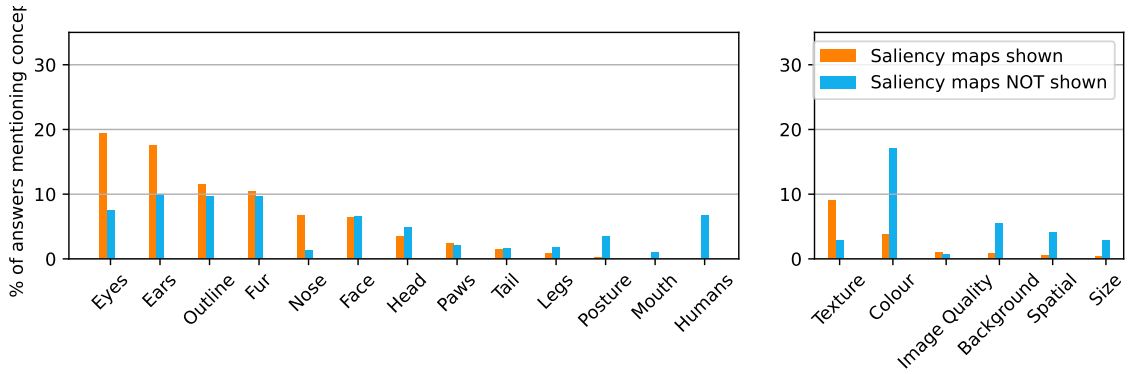


Figure 3.3. Performance results: When saliency maps were shown, participants were significantly more accurate in predicting the outcome of the classifier (left). Scores did not significantly influence the participant's prediction performance (right). Success rates were relatively low across conditions, showing that tasks were very challenging.

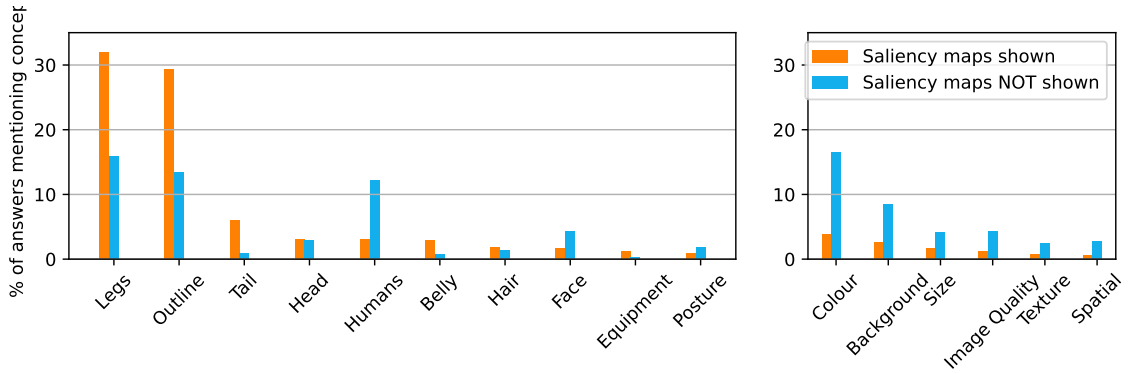
In the presence of saliency maps participants were more accurate in predicting the outcome of the classifier ($M = 60.7\%, SD = 11.0\%$ vs. $M = 55.1\%, SD = 10.8\%$). There was no significant main effect of the presences of scores on performance ($F_{(1,60)} = 1.938, p = 0.169, \eta^2 = 0.029$). Furthermore, there was no interaction effect ($F_{(1,60)} = 0.060, p = 0.807, \eta^2 = 0.001$).

3.3.2 Confidence

We also asked participants to rate their confidence in their forecast on a 4-point forced Likert item. Answers were coded by numbers 1-4 and summed up per participant. A one-way independent Kruskal-Wallis test showed that confidence was similar across conditions (H(3) = 1.130, p = 0.770). On average participants tended to be *"slightly confident"* in their answers (Median = 3.000). We also consider participants' accuracy on the subsets of images corresponding to different outcomes (i.e. TP, FP, FN). Overall the accuracy was higher for TP images, on average 79.4%, it was lower for FP, on average 46.9%, and even lower for FN, on average 36.7%.



(a) The normalised frequencies of individual features mentioned by participants for images of cats.



(b) The normalised frequencies of individual features mentioned by participants for images of horses.

Figure 3.4. Frequencies of individual features mentioned by participants: The top shows frequencies for images of cats, the bottom for horses. The left side shows features belonging to the Saliency-Features. The right side shows features belonging to the General-Attributes (frequencies were normalised for each participant).

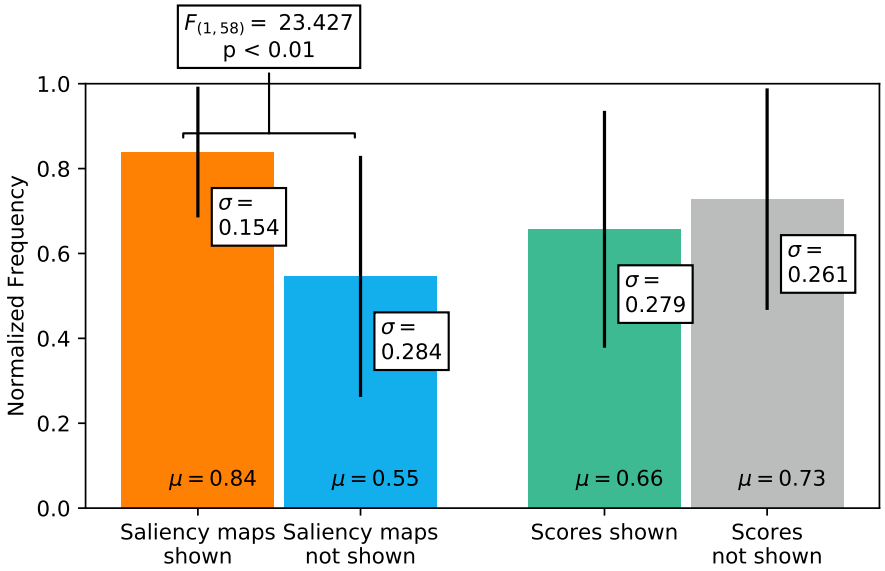


Figure 3.5. The ratio of mentioned Saliency-Features: It summaries the share of saliency-features participants mentioned per task. They mentioned significantly more such features when saliency maps were present (Left). Scores did not have an influence (Right).

3.3.3 Mentioned Saliency Maps Features

Besides making a prediction, we asked participants what features they think the classifier is sensitive to and what features it ignored.

Excluded Data

An analysis of the qualitative data revealed that two participants misunderstood these tasks. Consequently, they were excluded from this analysis. It also became apparent that many of the remaining participants misinterpreted the question about the features the system *ignored*. Therefore, we focused only on replies participants gave regarding the sensitivity of the classifier to features.

Mixed-Method Analysis of Answers

We carried out a qualitative content analysis [141] on the free text replies. In the first pass, two of the authors coded the answers inductively. Each response could be assigned several open codes based on the features or concepts it addressed. Subsequently, coders discussed their individually established codes and agreed on a shared and simplified codebook. We decided to assign each code to one of two mayor code groups: *Saliency-Features* and *General-Attributes*.

Saliency-Features This group included codes referring to features, which could be localized to pixels in the proximity of the object of interest and that saliency maps *could* highlight. The rationale for this was that we aimed to compare how frequently participants mentioned concepts related features that saliency maps *could potentially* highlight. Besides the somewhat obvious feature codes such as *Ears* and *Legs*, this group also included: *Equipment* - which applied to all objects associated with domestication such as *"leash"or "saddle"*, *Outline* which applied to answers referring to the *"shape"* or *"contour"* of the object of interest and *"Fur"* which was used for utterances referring explicitly to the *"fur", "skin"* or texture pattern *on* the animal.

General-Attributes This group included codes that refer to utterances of generic properties of the image. An example is the code *Background* - which applied to answers referring generically to *"surroundings"* or *"context"* but also objects in the background such as *"trees"*. Another example is *Image Quality* which was used for replies addressing issues of *"contrast"*, *"blur"* , *"lighting condition"* or *"occlusion"*. The code *Texture* was assigned when answers referred to images *"texture"* generically (i.e. "Fur patterns" are considered as a Saliency-Features).

Saliency-Features Ratio For the quantitative analysis, we counted the number of Saliency-Features codes and General-Attributes codes. We noticed that some participants wrote a lot in the qualitative response and therefore mentioned a lot of features, while others did not. To prevent this from skewing the results, we calculated a ratio. We obtained the Saliency-Features ratio for each participant by dividing the number of Saliency-Features codes by the total number of Saliency-Features and General-Attribute codes that we had assigned to their answers. Therefore a ratio of 0.6 means that 60% of the features that a participant mentioned were Saliency-Features. In the same fashion, we calculated ratios for all codes. The top of Figure 3.4b shows the ratios for the answers participants gave for images of cats, while the bottom of Figure 3.4b shows them for images of horses.

Ratio Analysis The Saliency-Features ratio was subjected to a statistical analysis. The data is summarized in Figure 3.5. A Shapiro-Wilk test revealed that the rate of Saliency-Features within groups were approximately normally distributed ($W = 0.900, p < 0.01$). A Levene's Test showed that the variances between groups were significantly different ($F_{(3,58)} = 3.749, p = 0.016$). To account for heteroscedasticity we ran a two-way independent measures ANOVA using white-corrected coefficient covariance matrix [251]. It revealed a statistically significant main effect of the presence of saliency maps on the rate of mentioned Saliency-Features ($F_{(1,58)} = 23.427, p < 0.01, \eta^2 = 0.295$). Partic-

ipants mentioned a larger share of Saliency-Features when saliency maps were present ($M = 83.9\%, SD = 15.4\%$ *vs.* $M = 54.6\%, SD = 28.4\%$). There was no significant main effect for the presences of scores ($F_{(1,58)} = 1.384, p = 0.244, \eta^2 = 0.013$) and no interaction effect ($F_{(1,58)} = 0.004, p = 0.948, \eta^2 = 0.001$).

3.4 Discussion

Through a combination of quantitative and qualitative analysis, the results of our study highlight the potential to use saliency maps as an explanatory tool for non-expert AI users, as well as their limitations. In the following subsections, we reflect on the key issues and highlight implications for design and further research.

3.4.1 The Utility of Saliency Maps exists, but It Is Limited

Our results show that when saliency maps were shown, participants predicted the outcome of the classifier significantly more accurately. Scores, instead, did not have a statistically significant effect. However, even with the presence of saliency maps, success rates were still relatively low (60.7%). Hence, the task of estimating the system's predictions on a new image remained challenging. This is also reflected by our participant's self-reported confidence in their answers, which was not affected by the presence of saliency maps or scores, and was on average still quite low. To explain this moderate outcome, we investigated participants' performance in more detail on subsets of images corresponding to different outcomes. Participants across conditions seemed to be better in predicting the system's outcome when it was correct (*true positives*: 79.4%). They were mainly struggling with the prediction of errors, performing worse than chance (*false postives*: 46.9% and *false negatives*: 36.7%). An interpretation of this result is that participants are possibly inclined to over-estimate the performance of the systems on challenging cases. Such cases are represented by FP and FN images. In fact, in 67.3% of all cases, participants predicted that the system would be correct, whereas it was only correct in 42.9% of the cases. One of the envisioned applications of explanations is aiding users in building appropriate trust into a system [56, 32]. Unexpected and unpredictable failures of a system affect trust more negatively than those that can be understood and anticipated [131, 56]. Therefore, it is important that users can understand when the system will fail. As detecting errors is a claimed utility of instance-level explanations [184, 128], we suggest that **future work** should evaluate this empirically in more detail. Our study design did not allow to draw conclusions in this regard because we did not fully counterbalance the order of tasks and True Negatives (TN) were not part of the task set.

3.4.2 Reasoning on Examples

In our study, we based the sampling strategy on the similarity distance between the task image and the training set. The rationale behind this was that people might learn more effectively from examples that are similar in appearance to the task image. It might help them to reflect upon the *visually similar* images that the system had successfully classified (i.e. TPs) and images the system had classified incorrectly (i.e FN, FP). We hypothesised that such contrasting reasoning can help users to understand the system's causes of successes and failures. However, when considering the examples presented to participants, we noticed that the usefulness of FN saliency maps is negligible. They usually highlight very little evidence (see i.e. the FN example in Figure 3.1). For FN examples, the actual image and the other saliency maps (TP, FN) become the only source of information for understanding why an example has not been recognised by the system. This insight suggests that the utility of saliency maps varies according to the classification score. In other words, a saliency map may highlight what supports the prediction of some class, but it will fail to provide counter-factual evidence, namely, the absence of evidence.

We would like to emphasise that for a human, it is easy to spot and point to the absence of a feature concept, while it is not for a CNN. Humans can easily break down an image into meaningful regions (semantics) [65]. In contrast, CNNs look for patterns in a sub-symbolic fashion that lead to an outcome [21, 134]. Because CNNs do not process data in a 'semantic' fashion, other patterns in an image (which may not belong to the concept) can contribute towards a classification outcome in unexpected ways [128]. **An implication** for the design is that we need to develop explanation algorithms that bridge the gap between humans and machines by leading the user to understand that the system is not basing its classification decision on higher-level 'semantics' of the image. Furthermore, we would like to emphasise that choosing representative examples with their corresponding saliency maps, which summarise the behaviour of the system well, is an under-explored topic. New approaches for generating saliency maps and for applying them to various machine learning problems are presented (see review [3]). However, very little work exists that investigates for which instances users should examine salience maps. Researchers have acknowledged that users can only inspect a limited number of saliency maps [184], but to the best of our knowledge, only two works explore sampling strategies [184, 128] - none of which where applicable for this work. An important implication, then, is that **further research** needs to characterise the effect of different sampling strategies of saliency map examples on users interpretation of the system operation.

3.4.3 Saliency Maps Can Help Participants Notice Features

Our results clearly indicate that saliency maps influenced our participants to notice the highlighted saliency features and to suggest that such features are important for the classification outcome. The ratio of mentioned Saliency-Features (e.g. *legs, outline*) compared to General-Attributes (e.g. *color, image quality*) was significantly higher when saliency maps were present while scores had no influence (Figure 3.5).

This effect can be explored in more detail in Figure 3.4b. It shows that saliency maps seem to lead people to pay attention to specific parts of the object of interest. For example, Figure 3.4b depicts the share of mentioned features for images of horses. It is evident that some features such as *legs*, *outline*, *tail* and *belly* were mentioned much more frequently by participants exposed to saliency maps, while general-attributes such as *background* and *colour* are mentioned more often when the saliency maps are not shown.

3.4.4 Facilitating Global Model Understanding by Explaining Local Features

It is worth emphasising that even when users notice features, this does not necessarily imply that they will perform better in predicting the outcome of the CNN or reach a global understanding of the model. Saliency maps provide only a visualisation of the importance of pixels in a single image. Transferring knowledge about potential features to new images, where they are presented in different orientations, scales, forms and perspectives, is very challenging. Furthermore, it is hard to get a quantifiable measure of the importance of individual features in an image. Again complexity increases if one attempts to quantify the importance of a feature on new images. In other words, it is difficult to estimate how the classification score would change if a feature would be absent. Would the score go down by a factor of 0.1, 0.2 or 0.6? Moreover, does the presence of different features cause an interaction effect? It is challenging for users to reason about this, especially when considering that CNNs process the input data in a non-linear fashion [21].

An implication for the design of explanation systems, then, is that saliency maps should be complemented by a global measure that explains how sensitive the presence of a feature is to the prediction of some class. For example, how sensitive the presence of *nose* is to the prediction of *cat*? In that regard, complementing saliency maps with this additional information could be valuable for users to build quantifiable measures of saliency maps, and perhaps avoid biases that might arise from exploring an unrepresentative subset of the dataset. Concept activation explanation, prototypical explanations and feature visualisations are global explanation techniques that could compliment saliency maps (cf. Section 2.3).

3.4.5 The Importance of General Attributes

Another reason why noticing Saliency-features does not necessarily facilitate a better understanding of a model is that general-attributes (e.g. colour, contrast) might influence the classification outcome. However, these general-attributes are usually not directly highlighted by saliency maps, because as a more general image property, they can not be localised to individual pixels. This points to the previously stated limitation of the expressive capabilities of saliency maps [198]. In fact, saliency maps might even prime participants to primarily consider only highlighted features, and give less weight to other attributes that are not highlighted but important. In contrast, users preconceptions may cause them to focus on attributes such as the *brightness* of the image, even if it is not a major cause of failure. **An implication** for design is to develop explanations that convey the right expectation to users. We suggest that saliency maps should be complemented by more global representations of the image features. For example, saliency information could be related to global descriptors of the images, such as overall contrast or brightness measures.

3.5 Limitations

The design space for the study we presented was vast. Our design choices outlined in Section 3.2 introduced some limitations, which we make explicit in this section.

The first limitation is the small number of image classes we considered. We decided for this compromise considering the limited time for each session, and the limited knowledge participant would have been able to obtain about class-specific behaviour. Future work should run a long-term evaluation (i.e. lasting several days or weeks) to allow participants to explore a large dataset with multiple classes in more depth. Another limitation of our design is that we used one specific network architecture (VGG16 [208]) and one specific technique to generate saliency maps (LRP [12]). With a series of pilot studies, we have tried to identify a combination of both techniques which provided saliency maps that participants found to be informative. However, this also means that results might be different with a different combination of techniques.

A limitation of our analysis is that the study design did not allow us to draw conclusions about users performance for different outcomes types (e.g. TP, FN, FP). The reason for this was that we did fully counterbalanced tasks, and True Negatives (TN) were not part of the task set. Future studies should address this limitation and study this aspect in more detail.

Finally, our participants were required to have a technical background, but we did not control for ML expertise. We see potential to repeat our study with different participant populations, such as ML-experts, or lay users.

Chapter 4

Do Users Benefit from Interpretable Vision? A User Study, Baseline, and Dataset

4.1 Research Questions and Publication

This chapter reports on our synthetic dataset and Study II, an online between-groups user study to evaluate counterfactual and concept-based explanations against an example-based baseline. After conducting Study I, we realised that forward simulation assesses two important aspects simultaneously: whether users understand which features are relevant to the model and how the model combines these features to make a prediction. Although it provides a comprehensive and effective measure of model understanding, it does not allow us to determine which aspects users struggle with. Our Study I and our technical review showed that different explanation techniques draw users' attention to different features. Whether they draw attention to the model's used features is unclear. Our next experiment was designed to answer this question. We had to use a different task, direct bias detection, which requires a model with known feature importance, as explained in Section 2.5.4 We aimed to create a model with at least one hidden bias that could not be easily spotted by example-based explanations, which were again chosen as our baseline. We decided to omit feature attribution methods from Study II due to their unfaithfulness and issues with spatially overlapping features (Cf. Section 2.3.2). This time, counterfactuals generated with an invertible neural network that did not have these issues were chosen as our primary evaluation candidate.

Research questions

- **RQ 9** — Can a synthetic dataset be used to bias a model arbitrarily?
- **RQ 10** — Can we create a model with multiple biases where at least one bias is hard to detect with example-based explanations?
- **RQ 11** — Do counterfactuals allow users to develop a better model understanding than concept-based or example-based explanations?

We answer RQ9, by creating Two4Two: a synthetic dataset depicting two abstract animals. Its data-generating factors can be correlated with the binary target class, thereby creating arbitrarily strong biases. A biased dataset does not guarantee a biased model. We empirically confirmed that, despite the default main feature (the leg position of the abstract animals), the model also uses the animal's shape and colour for its prediction. A biased model does not guarantee that its biases are difficult to detect. An initial user study (N = 43), answered RQ 10 by validating that participants were struggling to find both biases contained in our model using this technique. We improved our baseline explanation method used in Study I by optimising it for bias discovery. Example images are arranged in a grid grouped by the model's logit predictions. This design allows users to inspect all attributes that potentially predict the target class.

The main study (N = 192), answered RQ 11, where we compared the baseline against two state-of-the-art explanations: automatically discovered [263] concepts and counterfactual interpolations generated with an invertible neural network. We found that none of these explanations outperformed the baseline, even though some features were identified more accurately with counterfactuals. We qualitatively analysed participants' textual justifications and obtained insights into their use of explanations.

Contributions

- The Two4Two dataset with full control over the biases it contains.
- The dataset was specifically designed for user studies and to challenge existing interpretability approaches.
- A model with three relevant features, one of which is difficult to detect.
- A method to verify ground-truth feature importance for models trained on Two4Two;
- We provide a carefully crafted adoptable study design as a template for future empirical evaluations of interpretable vision using the task of bias detection.

- Our design is suitable for lay users and integrates several measures to ensure high-quality crowdsourced responses, including professionally produced instruction videos and extensive screening with multiple-choice tests.
- Our design includes a simple, yet powerful baseline technique that relies on the model's outputs only, allowing participants to scan easily through different model outputs, which we propose as a benchmark for future studies;
- We provide open access to our dataset, explanation techniques, model, and study design, including instructions and videos, to support the replication of our results and the adaptation of our design to other explanation techniques.
- The finding that example-based explanations outperform concept-based explanations.
- The finding that counterfactuals performed similarly to example-based explanations.
- Findings on mental heuristics that influence users' model understanding and usability issues encountered with all methods.

Publications An early version of the dataset, of which Martin Schuessler is the main author, was published as a worksop poster and was presented at the 2021 International Conference on Learning Representations (ICLR):

> Martin Schuessler, Philipp Weiß, and Leon Sixt. 2021. Two4Two: Evaluating Interpretable Machine Learning - A Synthetic Dataset For Controlled Experiments. In *Responsible AI – Workshop* (ICLR). doi: 10.48550/arXiv.2105.02825

The refined dataset and the studies have been jointly published as a full paper (Shared first author). It was presented at the 2022 International Conference on Learning Representations (ICLR):

> Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. 2022. Do Users Benefit From Interpretable Vision? A User Study, Baseline, And Dataset. In *Proceedings of the International Conference on Learning Representations* (ICLR). https://openreview.net/forum?id=v6s3HVjPerv. doi: 10.48550/arXiv.2204.11642

The sections of this chapter are taken from the original publication. Once again, the abstract, introduction, and related work section have been removed. Their contents have been covered in significantly more detail in Chapters 1–2 and 5.

Author Contributions Initially, the effort to develop Two4Two and to create a new counterfactual explanation method were two separate projects. Based on the insights from Study I, Two4Two was jointly conceptualised by Martin Schuessler and Phillip Weiß. Under the supervision of Martin Schuessler, Phillip Weiß implemented the first version of the synthetic dataset generator. Leon Sixt invented and built the method for generating counterfactuals. Martin Schuessler designed and conducted a study to evaluate this method. We used the CelebA dataset for this study but encountered difficulties as participants' reasoning was influenced by their prior knowledge of human faces. Our initial submission did not pass peer review [213]. As a result, the two projects, Two4Two and INN Counterfactuals, were combined to conduct a more rigorous evaluation. Many refinements were needed to adapt Two4Two for this study and to make it available under an open-source license for future research. Leon Sixt and Martin Schuessler implemented this. Most contributions for the accepted paper, on which this chapter is based, are a collaborative effort. Due to the highly collaborative nature of our work, we influenced each other's ideas and shared responsibilities. Leon Sixt created the method to quantify the importance of the ground-truth feature. It is inspired and enabled by the Two4Two data generator. Martin Schuessler revised Leon Sixt's method for generating counterfactuals based on the results of pilot studies. Martin Schuessler conceptualised and refined the layout of all explanation techniques in pilot studies. All aspects of the study designs and conduction were the responsibility and contribution of Martin Schuessler, with the following exceptions: A pilot run that Oana-Iuliana Popescu conducted considerably influenced the design, and Leon Sixt implemented a revision to the statistical analysis (prior to conducting the study). Martin Schuessler iteratively created a study design, which was a significant improvement over the design of Study I and the designs used in Sixt et al. [213] and Schuessler et al. [199]. The design ensured that participants had a full understanding of their task using professionally created voiced instruction videos and extensive participant screening involving several multiple-choice tests. Paper writing was a shared effort.

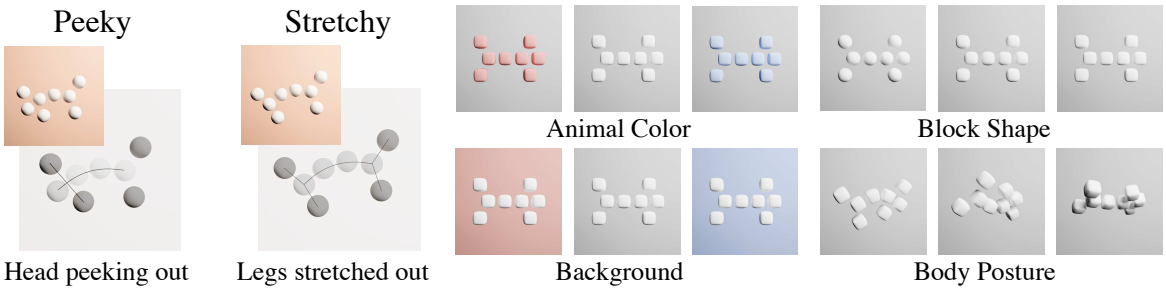


Figure 4.1. The left panel depicts the main difference between *Peeky* and *Stretchy*: the legs' position. While *Peeky* shows one pair of legs moved inwards, *Stretchy's* legs are moved outwards. Two4Two offers different attributes: animal color, background color, the shape of the blocks and the animal's body posture. All of which can be controlled and biased separately.

4.2 Two4Two: Datasets with Known Feature Importance

We developed an open source library that allows researchers to render synthetic image data suitable for human-subject evaluations. It was engineered for the task of bias detection (cf. Section 2.5.3.

4.2.1 Dataset Description

Datasets generated with Two4Two consist of two abstract *animal* classes, called *Peeky* and *Stretchy*. Both consist of eight blocks: four for the spine and four for the legs – Two4Two. For both animals, one pair of legs is always at an extended position. The other pair moves parallel to the spine inward and outward. The attribute *legs' position*, a scalar in [0,1], controls the position. At a value of 0.5, the pair of legs are at the same vertical position as the last block of the spine. Peekies have a leg position ≤ 0.52 which means legs are moved mostly inwards to the body centre. In the same fashion, Stretchies are extended outwards, legs' position ≥ 0.48. We added some ambiguity to ensure a model has an incentive to use possible biases. Therefore, Peekies and Stretchies are equally likely for a legs' position between 0.48 and 0.52. It is also difficult for humans to tell if the legs are outward or inwards in this range. Besides the legs' position, the dataset has the following parameters which can be changed arbitrarily and continuously:

- *body posture* (bending and three rotation angles)
- *position*
- *animal color* (e.g. from red to blue)
- *blocks' shape* (from cubes to spheres)

- *background color* (e.g. from red to blue)

When designing the dataset, we wanted to ensure that

1. Participants can become experts within a few minutes of training.
2. It allows for the creation of multiple biases that are difficult to find.
3. It provides a challenge for existing interpretability methods.

Goal (1) is met as participants can be instructed using only a few examples (see the tutorial video in Appendix D.1). The high number of controllable attributes achieve Goal (2). Biases can be introduced by correlating different attributes with an animal type. Figure 4.3 on page 60 shows a code snippet that creates a rotation bias. We validate in our first study that this can be used to bias a model in such a way that its biases do not stand out. Goal (3) is met by spatially overlapping attributes and long-range image dependencies. Spatially overlapping attributes, like colour and shape, directly challenge saliency map explanations. Long-range image dependencies, like the legs' positions relative to the spine, can not be explained when analyzing patches separately as done in Chen et al. [39] and Brendel and Bethge [27]. Both properties are common in real-world datasets: For example, race and gender in facial datasets are encoded by spatially overlapping features. Long-range image dependencies are particularly relevant for pose estimation and visual reasoning [101].

4.2.2 Introducing Biases

For our studies' dataset, we sampled the block's shape in a non-predictive biased fashion. This means that for legs' positions that clearly showed a Peeky [0, 0.45] most blocks were rather cubic, while for legs' positions that clearly showed a Stretchy [0.55, 1] most blocks were rather round. However, for the legs' positions between [0.45, 0.55] the blocks shape was uniformly distributed. In particular, in the even narrower interval [0.48, 0.52] where a classifier can only be as good as random guessing, the block's shape does not provide any additional information about the target class. In Figure 4.4, we show the joint distribution of the block's shape and legs' position.

We sampled the animals' color to be predictive for the target class. At the small interval where the legs overlap [0.48; 0.52], we distributed the animal color to provide additional class information. Stretchies were more likely to be red, and Peekies were more likely to be blue. Outside of this centered interval, the color gradually became uniformly distributed (see Figure 4.4). Hence, color was more equally distributed than the shape, making the color bias harder to detect visually. The remaining attributes, background color and body posture, were sampled independently of the class, and we expected our model to ignore them.

Factor	Range	Distribution	Biased	Class Information
Legs' Position	[0, 1]	Uniform with overlap	Yes	-
Color	[0, 1]	See Figure 4.4	Yes	Yes
Shape	[0, 1]	See Figure 4.4	Yes	No
Position Y	[-0.8, 0]	Uniform	No	No
Position X	[-0.8, 0]	Uniform	No	No
Background	$[0.05, 0.95]$	Uniform	No	No
Rotation Yaw	$[0, 2\pi]$	Uniform	No	No
Rotation Roll	$[-\pi/4, \pi/4]$	Truncated Normal(0, $0.03\pi/4$)	No	No
Rotation Pitch	$[-\pi/6, \pi/6]$	Truncated Normal(0, π / 8)	No	No
Bending	$[-\pi/10, \pi/10]$	Truncated Normal(0, π / 20)	No	No

Figure 4.2. Distribution of each attribute in the study's dataset: *Biased* denotes whether an attribute is unequally distributed for the two classes. *Additional Class Information* show if an attribute contains any additional information about the target class not already given by the legs' position.

4.2.3 Measuring Ground-Truth Feature Importance

Even if a dataset contains biases, it is unclear how relevant they will be to a neural network after training. Feature importance also depends on the network architecture, the optimization process, and even the weight initialization. As Two4Two allows us to change any parameter in isolation, we can directly compare the model prediction between two images that differ in only one parameter. For these two images, we measured both the median absolute logit change and also for how many samples the predicted class was flipped. Both measures quantify how influential each parameter is (see Figure 4.5).

As expected, the legs' position had a strong influence on the prediction. The model relied more on animal color than on the blocks' shape, which is expected as the color contains additional information about the class. Surprisingly, the prediction flip for unrelated attributes such as background was only slightly lower than for blocks' shape.

To analyze this further, we calculated a linear fit for each parameter change to the logit change. We reported the coefficient of determination R^2, which indicates how much of the variance in the prediction can be explained linearly by the analyzed property. While the unrelated properties sometimes flip a prediction, the direction of that flip is random ($R^2 \approx 0$). In contrast, the biased parameters influence predictions in a directed fashion, with animal color (R^2=0.751) being clearly more directed than blocks' shape (R^2=0.307).

```python
import dataclasses
import numpy as np
import matplotlib.pyplot as plt

from two4two.blender import render
from two4two.bias import Sampler, Continouos
from two4two.scene_parameters import SceneParameters

@dataclasses.dataclass
class RotationBiasSampler(Sampler):
    """A rotation-biased sampler.

    The rotation is sampled conditionally depending on the object type.
    Positive rotations for peaky and negative rotations for stretchy.
    """

    obj_rotation_yaw: Continouos = dataclasses.field(
        default_factory=lambda: {
            'peaky': np.random.uniform(-np.pi / 4, 0),
            'stretchy': np.random.uniform(0, np.pi / 4),
        })

# sample a 4 images
sampler = RotationBiasSampler()
params = [sampler.sample() for _ in range(4)]
for img, mask, param in render(params):
    plt.imshow(img)
    plt.title(f"{param.obj_name}: {param.obj_rotation_yaw}")
    plt.show()
```

Figure 4.3. Source code example to create a biased sampler with Two4Two: High positive rotations are predictive of Stretchy and low negative rotations of Peaky. This illustrative example differs from the sampler used in the study.

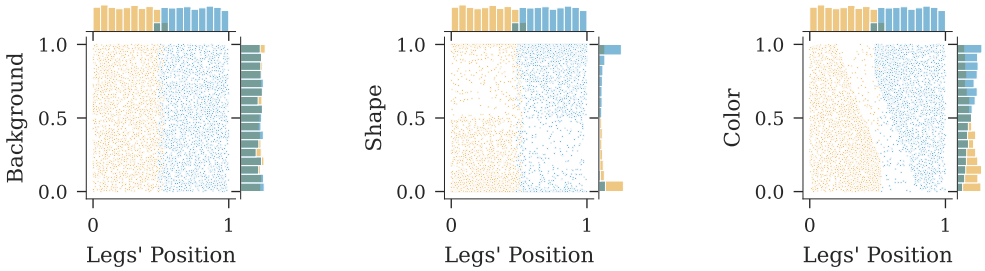


Figure 4.4. The joint distributions of legs' position and three other attributes: background (left), shape (middle), and color (right). Datapoints are yellow for Peekies and blue for Stretchies. The background is not biased. The shape is biased for legs' position lower than (0.45) or greater (0.55), but is uniform in the centre. The colour contains additional predictive information about the target class, as it allow to discriminate between Peeky and Stretchy where the legs' position overlaps. However, for more extreme arms' positions the colour is uniform and not biased.

4.3 INN Model and Evaluated Methods

As discussed in Section 4.2, Two4Two was designed to challenge existing interpretability methods, e.g., saliency map explanations and patch-based models. We selected two methods that might provide the user with the necessary information: counterfactuals generated with an invertible neural network (INN) and concept-based explanations [263] (C.f. Section 2.3.3-2.3.4).

4.3.1 INN Counterfactuals

We trained an INN using both a supervised and an unsupervised objective [48, 47]. To predict the target class, the model first applies the forward function φ to map a data point $\boldsymbol{x}$ to a feature vector $\boldsymbol{z} = \varphi(\boldsymbol{x})$. Then, a linear classifier takes those features $\boldsymbol{z}$ and predicts the logit score $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{z} + b$. Any input can be reconstructed from the feature vector by applying the inverse function $\boldsymbol{x} = \varphi^{-1}(\boldsymbol{z})$. The model has a test accuracy of 96.7%. Further details can be found in Appendix D.2. The baseline and concept techniques are also applied to this model. To create a counterfactual example $\tilde{\boldsymbol{x}}$ for a data point $\boldsymbol{x}$, we can exploit the linearity of the classifier. Moving along the weight vector $\boldsymbol{w}$, i.e., adding $\boldsymbol{w}$ to the features $\boldsymbol{z}$, changes the model's prediction. By controlling the step size with a scalar α, we can directly quantify the change in the logit value $\Delta y = \alpha\boldsymbol{w}^T\boldsymbol{w}$. The modified feature vector $\boldsymbol{z} + \alpha\boldsymbol{w}$ can be inverted back to the input domain, resulting in a counterfactual $\tilde{\boldsymbol{x}} = \varphi^{-1}(\boldsymbol{z} + \alpha\boldsymbol{w})$ which visualizes the changes introduced by a step $\alpha\boldsymbol{w}$ in $\boldsymbol{z}$-space. The INN's explanations are visualized in a grid where each row shows a single counterfactual interpolation (see Figure 4.6b).

Factor	Prediciton Flip [%]	Median Logit Change	R^2
Legs' Position	41.680	2.493	0.933
Color	7.080	0.886	0.751
Shape	3.920	0.577	0.307
Position Y	2.960	0.597	0.007
Background	2.640	0.523	0.006
Rotation Yaw	3.480	0.669	0.001
Rotation Roll	2.260	0.413	0.001
Bending	3.640	0.605	0.000
Rotation Pitch	3.500	0.627	0.000
Position X	3.380	0.581	0.000

Figure 4.5. Importance of the data generating factors to the model's prediction: The *Mean Logit Change* reports the median of the absolute change in logit values. The *Prediction Flip* column quantifies how often the model's prediction changed the sign when changing the attribute. For the R^2 *score*, we fitted an ordinary least squares from the factors' deltas to the deltas of the model's logits and then report the coefficient of determination (R^2).

4.3.2 Automatically-Discovered Concepts

We adapted the NMF approach of Zhang et al. [263] to our specific network architecture. Because the network's internal representations also contain negative values, we used matrix factorization instead of NMF. We generated the concepts using layer 342 (from a total of 641 layers). The layer has a feature map resolution of 8x8. This choice represents a trade-off between enough spatial resolution and high-level information. We ran the matrix factorization with 10 components and selected the five components that correlated most with the logit score (r is in the range [0.21, 0.34]).

Our presentation of concept-based explanations was very similar to Zhang et al. [263]: we visualized concepts with five exemplary images per row and highlighted regions corresponding to a concept. Since our classifier is binary, a negative contribution for Stretchy actually means a positive contribution for Peeky. Hence, we could have characterized a concept as *more Peeky* and *more Stretchy*, to make the design similar to the other two explanation techniques. However, as the concepts did not strongly correlate with the model's output, presenting them as class-related could confuse participants: a *more Peeky* column would have contained some images showing Stretchies and vice versa. Thus, we presented them separately in two consecutive rows (See Figure 4.6c). Presenting concepts in this fashion gives them a fair chance in the study because participants rated the relevance of each attribute for the model rather than for each class separately.

4.4 Human Subject Study

We share the view of Doshi-Velez and Kim [52] and Vaughan and Wallach [240] that user-testing of explanation techniques is a crucial but challenging endeavor. As our second main contribution, we propose and conduct a user study based on the Two4Two dataset which can act as a blue-print for future investigations. Our design has been iterated in over ten pilot studies and proposes solutions to common problems that arise when evaluating explanation techniques on crowd-sourcing platforms with lay participants.

4.4.1 Design Considerations

Data without Prior Domain Knowlege

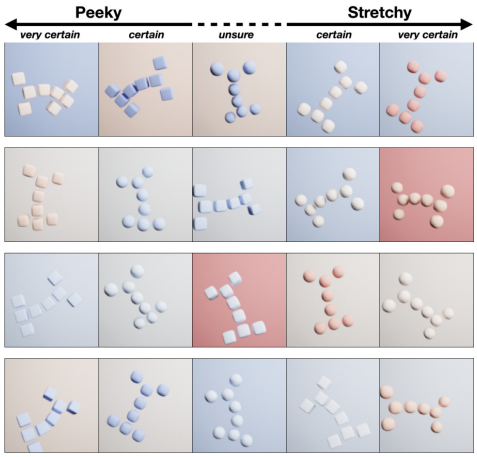
We specifically designed the Two4Two dataset to avoid overburdening participants, as might be the case with other types of data. Within a few minutes, participants can easily become domain experts. Since the data is unknown to them prior to the study, we avoid introducing any prior domain knowledge as a confounding factor.

Manageable but not Oversimplified Tasks

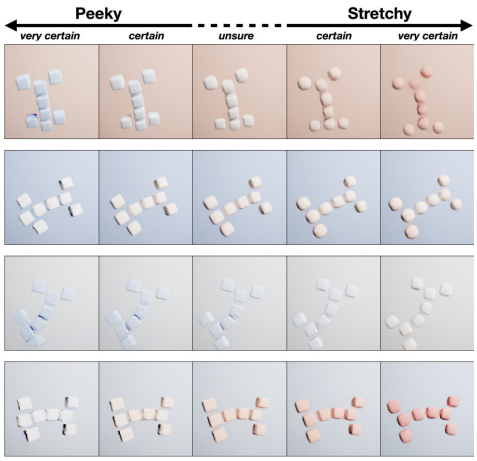
We use the task of *direct bias detection*: participants had to rate features as either *relevant* or *irrelevant* to a model. The task directly reflects users' perception of feature importance. Furthermore, it has the advantage of being suitable for lay participants. At the same time, it is also grounded in the model's behaviour. This is an advantage over tasks used in several previous studies, which only evaluated whether explanations were *accessible* to users, e.g. by identifying the target property *smiling* using image interpolations [209] or assigning images to a concept class [263, 75]. However, these tasks are an oversimplification and cannot measure any insights the users gained about the model. In contrast, the task of *forward prediction* requires substantial model understanding and is very challenging, as reflected by the participants' low accuracy in our previous study (Section 3).

Baseline Explanation Technique

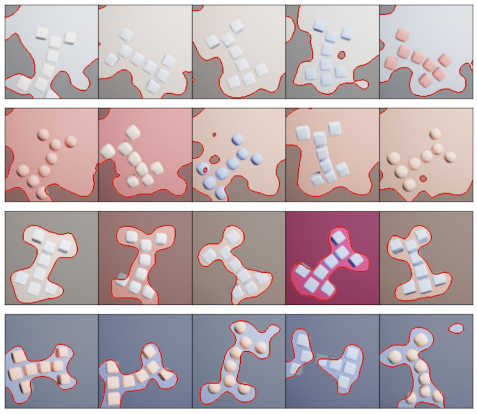
To quantify whether an explanation is beneficial for users, it must be compared to an alternative explanation. In this work, we argue that a very simple and reasonable alternative for users is to inspect the model's logits assigned to a set of input images. Others studies already discovered that users predominantly rely on predictions rather than on sophisticated complimentary explanations when reasoning about a model [41, 5]. Simple example-based explanations have surfaced as a strong baseline in studies by [25]. We designed such a baseline explanation as shown in Figure 4.6a. After several design iterations, we settled for a



(a) Baseline



(b) Invertible Neural Networks



(c) Concepts [263]

Figure 4.6. Examples of the explanations used in the study: We tested whether users can identify the class-relevant features of images showing two types of animals. We biased attributes like the animal's color to be predictive of the class and investigated whether explanation techniques enabled users to discover these biases. We tested a simple baseline **(a)** which shows random samples grouped by the model's output logit, counterfactual samples generated by an invertible neural network **(b)**, and automatically discovered concepts **(c)**. A participant viewed only one of the above conditions.

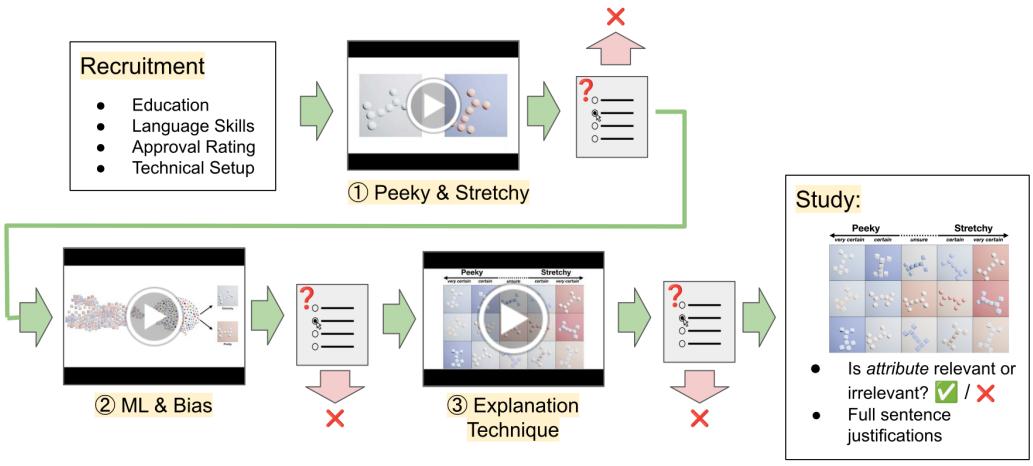


Figure 4.7. Participant flow and screening: We recruited proficient English-speaking participants from Prolific with a high approval rate. They viewed three tutorial videos, followed by a written summary and multiple-choice comprehension questions. Participants who failed the test twice were excluded from the study. See Appendix D.1 for details.

visually dense image grid with 5 columns sorted by the logit score, each column covering 20% of the logit values. The columns were labeled very certain for Peeky/Stretchy, certain for Peeky/Stretchy, and as unsure. Pilot studies showed that participants' attention is limited. We thus decided to display a total of 50 images, i.e. an image grid of 10 rows. The number of images was held constant between explanation techniques to ensure the same amount of visual information and a fair comparison. In this work, we focused on binary classifications. For a multi-class setting, one could adapt the baseline by contrasting one class verses another class.

High Response Quality

We took extensive measures to ensure participants understood their task and the explanation techniques as illustrated in Figure 4.7. Participants were required to watch three professionally-spoken tutorial videos, each under four minutes long. The videos explained, on a high level, the Two4Two dataset, machine learning and how to use an assigned explanation technique to discover relevant features. To avoid influencing participants, we prototyped idealized explanations using images from Two4Two. The explanations showed different biases than those in the study. Each video was followed by a written summary and set of multiple choice comprehension questions. After failing such a test once, participants could study the video and summary again. When failing a test for a second time, participants were excluded from the study. We also excluded participants if their written answers reflected a serious misunderstanding of the task, indicated by very short answers copied for all attributes or reasoning that is very different from the tutorial. We recruited participants

from Prolific who are fluent in English, hold an academic degree and have an approval rate of $\geq 90\%$. To ensure they are also motivated, we compensated them with an average hourly pay of £11.45 which included a bonus of £0.40 per correct answer.

4.4.2 Experimental Design

We conducted two online user studies. Before starting the data collection, we formulated our hypotheses, chose appropriate statistical tests, and pre-registered our studies (see Appendix D.5-D.6). This way, we follow the gold-standard of defining the statistical analysis before the data collection, thus ensuring that our statistical results are reliable [42]. The first study (N=50) analyzed whether the task was challenging enough that other methods could potentially improve over the baseline. We tested if at least one bias in our model (either the animal's color or the blocks' shape) was difficult to find using the baseline technique. Consequently, we used a within-subjects design.

In the second study (N=240), we evaluated the two explanation techniques described in Section 4.3 against the baseline using a between-groups design. Participants were randomly, but equally assigned to one of the explanation techniques. We specified two directed hypotheses. We expected participants in the INN condition to perform better than those in baseline, because the baseline does not clearly highlight relevant features, whereas interpolations highlight features in isolation. We expected participants viewing concepts to perform worse than those in the baseline, due to their inability to highlight spatially overlapping features.

For both studies, participants completed a tutorial phase first. Using their assigned explanations, they then assessed the relevance of five attributes: legs' position relative to the spine, animal color, background, rotation or bending, and blocks' shape. The questions were formulated as: *"How relevant is <attribute> for the system?"*, and participants had to choose between *irrelevant* or *relevant*. The percentage of correct answers (accuracy) served as our primary metric. Participants also had to write a short, fully-sentenced justification for their answers.

4.5 Results

4.5.1 Data Exclusions

As stated in the preregistration, we automatically excluded all participants who withdrew their consent, failed one of the comprehension questions twice, skipped a video, or exceeded Prolific's time limit for completion. If a participant was excluded, a new participant's place was made available until the pre-registered number of completed responses was reached. We

Condition	$N_{\text{collected}}$	N_{filtered}	Overall	Legs	Color	Backgr.	Shape	Posture
Study 1	50	43	73.4	86.0	48.8	86.0	74.4	72.1
Study 2	240	192	67.0	78.2	58.9	66.8	73.1	59.1
INN	80	62	**84.5**	***100.0**	*82.3	*79.0	90.3	**71.0**
Baseline	80	71	80.8	85.9	59.2	**95.8**	**93.0**	70.4
Concepts	80	59	32.2	45.8	32.2	18.6	32.2	32.2

Figure 4.8. Performance results per attribute: The mean accuracy for each attribute by condition. $N_{\text{collected}}$ provide the number of participants collected and N_{filtered} the number of remaining participants after the filtering. Stars mark statistical significance.

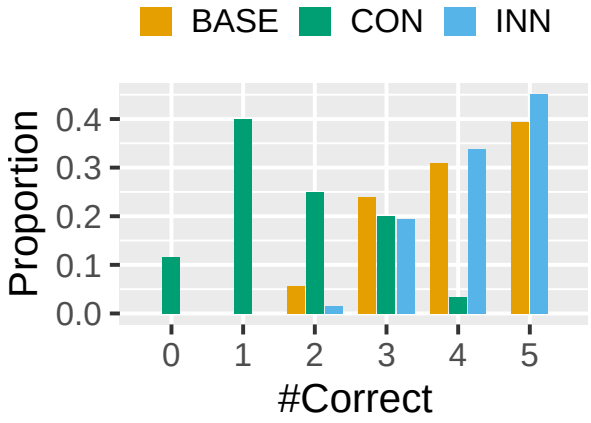


Figure 4.9. Performance results per condition: The proportion of correct answers for baseline (Baseline), concepts (CON), and INN.

excluded 63 study respondents for the first study, and 145 for the second study in this fashion. We ensured that all participants were naive about the dataset. Once they participated in a study, they were blacklisted for future studies.

For completed studies, two annotators independently marked the participants' written answers and excluded those with copy and paste answers or indications of grave misunderstandings of the instructions. Participants were labeled as: *include*, *unsure*, or *exclude*. Both anotators had an agreement of $\kappa = 0.545$ for the first study and $\kappa = 0.643$ for the second (measured *include* vs. *unsure* and *exclude*). Disagreements were solved by discussion. In total, we excluded 7 participants from the first study (14%) and 48 participants from the second study (20%).

4.5.2 First Study

For the accepted 43 participants, we used two-sided exact McNemar tests on their answers about the relevance of the *legs position* compared with *animal color* (first test) and *back-*

ground (second test). Participants found the color bias less often than the legs' positions (p <0.0001). The success rate for the color attribute was 49% vs. 86% for legs. The shape bias was not significantly harder to find than the legs' positions and was identified correctly with 74% accuracy (p=0.3036). Hence, we confirmed our hypothesis and concluded that other methods still have room for improvement over the baseline.

4.5.3 Second Study

In the second study, we evaluated 192 valid participant responses (62 INN, 71 BASE, 59 CON). We expected data to be different from the normal distribution, and a Shapiro-Wilk test for all conditions confirmed this ($p < 0.001$). We depict the number of correct answers per condition in Figure 4.9. A Kruskal-Wallis test showed a significant differences in accuracy scores between conditions ($p < 0.001$). For focused comparisons, we used two Wilcoxon-rank-sum tests with Bonferroni correction to correct for multiple comparisons. The accuracy scores differed significantly between the baseline and concept conditions ($p < 0.001$, r=0.778). The performance of participants using concepts was rather poor, with only 31.7% accuracy, considering that random answers would yield a score of 50%. For concepts, not a single attribute surpassed the 50% barrier. We found no significant difference when comparing the baseline and counterfactuals (p=0.441, r=0.091). Their mean accuracies are close, with 80.8% for baseline and 84.5% for counterfactuals. INN counterfactuals helped users to discover the main attribute, legs' position, (p <0.001) and color bias (p=0.033) more reliably.[1] However, counterfactuals performed significantly worse for the background attribute (p=0.033), while for blocks' shape and position we found no significant difference (for both, p=1).

4.5.4 Qualitative Results

To understand how participants integrated the explanation techniques into their reasoning, we analyzed the textual answers of each feature qualitatively. Two annotators first applied open coding to the answers. They performed another pass of closed coding after agreeing on a subset of the relevant codes, on which the following analysis is based. Overall, the participants perceived the task as challenging, as they expressed being unsure about their answers (N=71).

We designed our image grid to show both possible classes and provide information about the model's certainty. We found that many participants integrated this additional source of

[1] The statistical analysis of the attributes for INN vs. baseline was not pre-registered. The reported p-values for the attributes were corrected for eight tests (including the pre-registered tests) using the Holm–Bonferroni method.

information into their reasoning. This was especially prevalent in the baseline condition (N=51). Participants particularly focused on the columns 'very certain Peeky' and 'very certain Stretchy', as well as on the column 'unsure'. While this may have helped confirm or reject their own hypotheses, it sometimes led to confusion; for example, when an image that exhibited a pronounced leg position, and therefore could easily be identified as Peeky or Stretchy, was classified by the model as 'unsure' (N=14).

Across conditions, we also observed that participants expect that all images needed to support a hypothesis. *"The animals are in different colors, there are blue stretchy and also blue peeky animals, If the color was relevant peeky/stretchy would be in one color etc"* (P73, BASE). Across conditions, most participants that applied such deterministic reasoning failed to find the color bias. In contrast, other participants applied more probabilistic reasoning, which helped them deal with such contradictions: *"Peeky is more likely to be blue in colour, whereas Stretchy is more likely to be pink. This is not always true (e.g. the shapes can be white in colour at either ends of the spectrum) but it might be somewhat relevant to help the system decide"* (P197, INN).

Another observed strategy of participants was to reference how often they saw evidence for the relevance of a feature (N=35), which was very prevalent in the concepts condition (N=20). Especially concepts were rather difficult for participants to interpret. A common issue was that they expected a relevant feature to be highlighted completely and consistently (N=38). Several instances show that participants struggled to interpret how a highlighted region can explain the relevance of a feature, *"If this [the legs position] were relevant I would have expected the system to highlight only the portion of the image that contains the legs and spine. (e.g. only the legs and one block of the spine at each end). Instead, every image had minimally the entire animal highlighted"* (P82, CON). Furthermore, spatially overlaping features were another cause of confusion: *"there are rows in which the animal is highlighted but not the background so it could be because of color, shape or rotation"* (P157, CON)

Participants erred more often for the background in the INN condition than for the baseline. We conducted an analysis to investigate this issue. We found that 29 participants stated that they perceived no changes in the background of the counterfactuals and hence considered this feature irrelevant. Another 21 participants noted that they saw such a change, which let 12 of them to believe its a relevant feature. *"The background color changes in every case, it's also a little subtle but it does"* (P205). Another 9 participants decided that the changes were too subtle to be relevant. *"The background colour does not change an awful lot along each row, maybe in a couple of rows it changes slightly but I do not feel the change is significant enough that this is a relevant factor in the machine decision"* (P184).

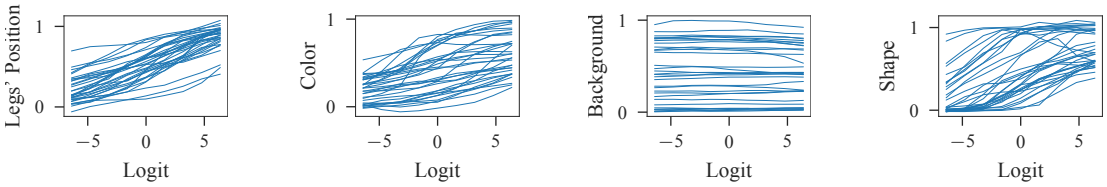


Figure 4.10. Attribute changes along counterfactual interpolations: Changes were measured by an observer convnet. Each line corresponds to a single sample whose logit score is modified through linear interpolations in the classifier space.

4.5.5 Do Counterfactuals Highlight Irrelevant Features?

Indeed, subtle perceptual changes in background color are present (Figure 4.6b). To quantify these changes, we decided to use an objective observer: a convolutional neural network. We trained a MobileNetV2 [193] to predict the parameter values of individual attributes of an image (e.g., background color, object color, etc.) using a completely unbiased version of Two4Two. After training, the model could predict the parameter values almost exactly (MSE < 0.0022, for details, see Figure D.1 in the Appendix). We then used this model to evaluate the parameter values of counterfactual INN interpolations, each spanning 99% of the logit distribution. We visualize the predictions of MobileNetV2 in Figure 4.10. All predictive properties (legs' position, body color, blocks' shape) are changed by the counterfactuals consistently. For the background, the changes are subtle but present. We also quantified the change in parameters using the difference between the maximum and minimum predicted value per individual interpolation which is shown in Figure 4.11. This supports the finding that relevant attributes change the most and the background changes just slightly like other irrelevant attribute. However, this seems enough to give some participants a false impression about its relevance.

4.6 Limitations

Synthetic Data We presented a user study on a synthetic dataset. We believe that the results also have implications for natural image data. When we created Two4Two, our objective was to translate challenges faced on "real" computer vision data (like spatially overlapping features) into an abstract domain. Although some properties of photorealistic datasets are lost in this abstraction, a method performing poorly on Two4Two would likely not perform well on a natural dataset with spatially overlapping features.

Limited Factors and Resolution Due to budget constraints, we limited the number of factors in our experimental design (external vs. internal validity trade-off). Our study intro-

Attribute	Mean Maximal Change	SD
Legs' Position *	**0.662**	0.140
Color *	**0.440**	0.190
Shapes *	**0.624**	0.208
Bending	0.059	0.042
Background	0.045	0.044
Rotation Pitch	0.186	0.126
Rotation Yaw	0.102	0.182
Rotation Roll	0.003	0.001
Position X	0.105	0.078
Position Y	0.103	0.078

Figure 4.11. Influence of all attributes on the model's predictions: Differences for each attribute between the maximum and minimum predicted value encountered when interpolating along the weight vector. Relevant attributes (*) change the most (values in bold). The background shows no considerable difference to the other irrelevant attributes.

duced a predictive bias for the animal's color and a non-predictive bias for the blocks' shape. It remains unclear how our results may have changed for a different dataset configuration: certain biases could exhibit different visual saliency. It remains also left for future work to determine which visual interface design is optimal for a given method. Furthermore, our study design restricted participants to make binary choices and provide textual justifications – limiting our understanding of the participants issues.

Chapter 5

Conclusions of Part I

At the outset of our research, there was limited empirical evidence on the effectiveness of novel explanation techniques to increase user understanding of computer vision models, and it was unclear which types were most beneficial. This gap stemmed from the initial separation of the AI/ML and HCI research communities [1]. Human-centric insights and best practices for experimental design were often not considered in machine learning research. HCI researchers, on the other hand, were limited by their engagement with simpler models or mockups for explainability studies [146]. Our work bridges this divide by rigorously evaluating advanced explanation techniques against complex models, thereby enhancing methodological standards in interpretable vision and fostering insights that are valuable to both domains. This section aims to distil essential insights and implications for future research and practical applications, reflecting on the evolution of our understanding throughout this endeavour.

5.1 Example-Based Explanations Are More Than a Baseline

Initially, our expectations differed significantly from the outcomes observed in our research, leading to an unforeseen revelation: *Example-based explanations are not only a baseline but a robust competitor to all other explanation methods*. As our review showed, this surprising observation was validated by subsequent studies [25, 41, 181], underscoring the reliability of this conclusion.

While it is a sobering realisation for XAI research that they have not been surpassed by more sophisticated methods for over a decade now, it allows us to make a clear **recommendation for practitioners**: Since example-based explanations are very easy to implement, they should be used whenever possible. From a practical point of view, they stand out

among other methods because they are just as effective for model understanding but offer model-agnostic flexibility and avoid the need for model-specific architectures or access to model internals. They also sidestep the accuracy-explainability trade-off seen with custom models, such as invertible networks or concept bottleneck models.

Their main limitation is that they are local explanations, placing the interpretive burden on users to deduce influential features and concepts. We argue that reaching a solid model understanding is impossible with the sole use of local explanations. However, example-based explanations are often preferred by users [100] and perceived as intuitive [112]. They may appear uninformative at times [112], but avoid visually overstating their ability to explain a model, unlike unfaithful methods such as feature attribution, GANs, or manually annotated concepts. When implementing them, it is crucial to carefully choose the images used as examples and how they will be presented. This is because showing examples of accurate classifications can increase users' trust in the model while highlighting examples of incorrect classifications can decrease it [19, 220]. We suggest presenting both types of examples to help users understand the capabilities and limitations of the model. The interface we designed for Study II was adequate for detecting bias and explaining binary classification. Correct and incorrect classifications were arranged in a grid based on the model score, which is an improvement over the design used in Study I, where they were visually separated. Study I showed that showing plain scores was not helpful. In Study II, users could vertically scan the image grid for relevant features, and they primarily focused on high-certainty examples and those at the class boundary. So, we suggest focusing on displaying examples from these certainty classes. Our latest design can be adapted for multi-class classification to compare one class to another. However, since this requires inspecting more examples, the effectiveness of this variation is uncertain.

5.2 Three Explanation Methods Are Less Effective for Images Than Claimed

Saliency Maps

Even though feature-attribution methods have been shown to be effective for text classifiers [123, 121, 184, 124, 66], we cannot confirm that generating saliency maps using these methods [184, 12] is equally effective for image classifiers. The results of Study I indicate that saliency maps can help users learn about some specific image features the system is sensitive to and slightly enhance their ability to predict the network's outcome for new images. However, even with saliency maps present, the CNN model remained largely unpredictable for participants (60.7% prediction accuracy). For misclassified images, prediction accuracy

remained well below chance level (43.8% for False Negatives and 49.2% for False Positives). That feature-attribution methods perform worse on the images than on text might be due to the higher complexity of models and input space. However, throughout our work, we became aware of the numerous limitations of saliency maps.

Limited to Local Features Even with very informative examples, saliency maps can only highlight the importance of features that are localisable to pixel regions, a limitation shared with local concepts. Our studies revealed that when a region contains multiple features, it may lead to difficulties for users. Researchers can use our Two4Two dataset, which has several spatially overlapping features, as a challenge to address this. According to Study I, saliency maps distract users from global features. Therefore, we recommend using them with concept explanations focusing only on global features.

Potentially Misleading As discussed in Section 2.3.2, saliency maps are unfaithful, which may be acceptable if explanations are helpful. However, our Study I and other studies found little benefit in using saliency maps (Section 2.6 and [108, 5, 142, 41]). The study of Chu et al. [41] even found them harmful to model understanding. Their presence influences how users think about a model [220]. Even if randomly generated, they increase users' trust in a model [41]. As an **Implication for practitioners**, we recommend using other faithful explanation techniques, such as example-based explanations or INN counterfactuals, to avoid misleading users with unfaithful saliency maps.

Limited to Feature Presence Saliency maps only convey that a feature is present and relevant. They do not easily convey its relevancy or the effect of its absence on the outcome. In contrast, counterfactuals give clearer insights by altering only relevant features, often adding missing ones during creation.

Counterfactuals

Our motivation for including INN counterfactuals in Study II was to support contrastive reasoning and faithfully explain the model. Although they calculate feature relevance accurately, they fail to communicate this to users. In the counterfactual images, the model's main features were changed significantly more than the background, which was irrelevant. However, every pixel or feature is marginally relevant to the model and can cause slight changes in the resulting image. In our study, participants perceived that the background, which constitutes a significant portion of the image, was a relevant feature due to subtle changes.

Automatically Discovered Concepts

In study II, the use of concepts was detrimental to participants' model understanding as their performed was worse than chance (31.7%) and significantly worse than when using counterfactuals and examples. The main issue was that concepts failed to completely and consistently highlight relevant features, providing evidence that while discovered concepts may been consistent [75], they are not guaranteed to be semantically meaningful. A recent study by Ramaswamy et al. [181] partially replicated our result, finding that concepts provide no benefit over example-based explanations. Manually annotated concepts are no alternative either, as they have faithfulness issues. We suggest that **future studies** investigate concept bottleneck models as an alternative.

5.3 Cognitive Heuristics Limit Explanation Understanding

The analysis of the qualitative answers of participants in both of our studies revealed several misconceptions that influence users' interpretation of explanations (Section 4.5.4, 3.4.2 and 3.4.4).

Semantic Reasoning Many users assume AI systems semantically process input. Therefore, it is important to communicate to users that neural networks *primarily* look for patterns in a sub-symbolic fashion and do not process the higher-level semantics of an image [21, 134]. We believe that prototypical explanation and feature visualisation are most suitable for conveying this to users.

Deterministic Reasoning Users often create hypotheses based on the relevance of features and expect them to be true for every classification in a deterministic manner. However, this demonstrates a significant limitation of any local explanation method because neural networks process features non-linearly [21], which is more akin to probabilistic reasoning. This contradicts users' intuition. Even if they detect which features are important for the network, this knowledge only contributes to a limited increase in the overall understanding of the model.

5.4 Open Challenges to Enhance Model Understanding

Our studies provide a critical evaluation of four interpretability methods. When considered alongside the studies reviewed in section 2.6, we conclude that users continue to struggle

with understanding image classifiers, regardless of the explanation technique employed. These findings serve as a reminder that making AI explainable is still very much an open technical challenge. We emphasise that *interpretability for computer vision should remain an active field of enquiry*. We recommend less emphasis on algorithmic-centred contributions, which simply derive new methods. Instead, we suggest addressing some of the following **technical challenges**:

- Which sampling method finds images for example-based explanations that increase users' model understanding the most? (Section 3.2.2)
- How do we generate faithful counterfactuals that highlight only the most relevant features to benefit from their contrastive abilities? (Section 4.5.5)
- How can feature-attribution methods become faithful to the model, and which faithful saliency maps increase users' model understanding the most? (Section 2.3.2 and 3.4.2)
- How do we automatically discover semantically meaningful concepts faithful to the model?
- How do we visualise spatially overlapping features or concepts?
- How can we create a dataset as controllable as Two4Two but with natural images?
- Can we design an explanation method that is aligned with or rectifies users' cognitive heuristics? (Section 5.3)
- Can we design a global explanation method that uses a linear combination of a limited number[1] of semantically meaningful features, concepts or prototypes?

5.5 More and Better User Studies Are Needed

As AI models become increasingly complex, further studies are necessary. Recall that, according to Miller [146], interpretability is *"the degree to which an observer can understand the cause of a decision"*. Hence, any claims of advancing interpretability without empirical evidence must be considered unverified. There is now sufficient insight that automatic metrics, cherry-picked anecdotal evidence, and measurements of explanation satisfaction do not reliably provide such evidence (Section 2.4.) User studies are laborious and expensive but crucial for future interpretability research. In addition to validating explanation methods, these studies can guide technical innovations by identifying areas where users struggle. Such studies must be conducted with rigour. Our taxonomy and identified best practices can serve as guidelines for planning future studies (Section 2.5 and 2.7). We echo the verdict of Doshi-Velez and Kim [52] that XAI researchers and reviewers need to *"respect the time*

[1] Based on the findings of Ramaswamy et al. [181] no more than 32.

and effort involved to do such evaluations". We argue that the HCI community is well-placed to collaborate with machine learning researchers to conduct better evaluations. We hope the work here can serve as an example. We open-source our experimental guidelines, videos, study designs, and code, and we encourage the community to use them.

Part II

Investigating Gaze Perception in Video-Mediated Communication

Chapter 6

Investigating Gaze Perception in Video-Mediated Communication

6.1 Research Questions and Publication

In this chapter, we shift our focus from improving human-AI interaction in the context of interpretable vision to improving human-to-human interaction mediated by computer vision applications. While previous chapters scrutinised users' comprehension and trust in image classifiers, here we extend our exploration to how computer vision can enrich digital conversations. This shift in research focus happened amid the COVID pandemic, which elevated video conferences to a critical communication tool. Millions of users working from home face challenges in virtual collaboration, such as Zoom fatigue. Leveraging our user-centric design experience covered in the first part of this work, we tackled one fatigue factor: the lack of gaze information and eye contact.

We explore an enhanced system concept, Gazing Heads, with the capability of conveying gaze and attention. It is a round-table virtual meeting approach that enables direct eye contact and signals gaze via controlled head rotation. Similar to the explanation techniques for image classifiers, the technology to achieve Gazing Heads is not quite mature. We built a camera-based simulation of Gazing Heads for four simultaneous video conference users to investigate the potential benefits of our proposed concepts and inform the development of new synthetisation algorithms.

Research Questions

- **RQ 12** — Do synthesised head rotations using view transition convey users' gaze and attention?
- **RQ 13** — How does the presence of perceivable gaze cues in video conferencing

systems affect users' experience and communication behaviour?

- **RQ 14** — Does Gazing Heads increase social presence?
- **RQ 15** — Does Gazing Heads influence how often and how long users gaze at one another?
- **RQ 16** — What technical challenges and user experience issues need to be addressed to implement gaze-aware video conferencing on standard laptop computers?

We conducted another rigorously designed user study comparing Gazing Heads (GH) with our baseline condition, the conventional "Tiled View" (TV) video conferencing system, for 20 groups of 4 people, on each of two tasks. The study found that head rotation clearly conveys gaze and strongly enhances the perception of attention. Measurements of turn-taking behaviour did not differ decisively between the two systems (though there were significant differences between the two tasks). A novel insight in comparison to prior studies is that there was a significant increase in mutual eye contact with Gazing Heads, and that users clearly felt more engaged, encouraged to participate, and more socially present.

Overall, participants expressed a clear preference for Gazing Heads. These results suggest that fully implementing the Gazing Heads concept, using modern computer vision technology as it matures, could significantly enhance the experience of video conferencing.

Contributions

- We conceptualised Gazing Heads which provides higher levels of gaze realism and is anticipated to be feasibly implementable on standard laptop computers in the near future;
- We built a four-party experimental rig to test the hypothesis that synthetic head rotation enhances video conferencing;
- We ran a study with more participants ($N = 80$) and using a wider range of measures than earlier work;
- Making gaze perceivable (Gazing Heads) is found to improve the perception of attention in video conferences;
- Control over head rotation significantly enhances mutual eye contact, social presence and user engagement;
- Interviews and questionnaires show that users prefer Gazing Heads over Tiled View. This suggests that by addressing missing gaze cues one can alleviate a major factor of Zoom fatigue, once the technicalities of real-time synthesised head rotation are resolved.

Publication This work has been accepted as a journal paper in ACM Transactions on Computer-Human Interaction and will be presented at the 2025 ACM Conference on Human Factors in Computing Systems (CHI):

> Martin Schuessler, Luca Hormann, Raimund Dachselt, Andrew Blake, and Carsten Rother. 2024. Gazing Heads: Investigating Gaze Perception in Video-Mediated Communication. *ACM Transactions on Computer-Human Interaction*. TOCHI 31, 3. doi: 10.1145/3660343

The sections presented are an extended version of the original publication and include additional details.

Author Contribution Carsten Rother and Andrew Blake conceived the idea of synthesising head rotations using an advanced computer vision method and eye-tracking. Martin Schuessler designed, and conducted all aspects of the study, supervised by Raimund Dachselt. Inspired by Hydra [201], Martin Schuessler suggested the simulation of the concept based on cameras. Martin Schuessler and Lucas Horman built the Gazing Heads Simulation. Luca Horman assisted in conducting the studies and provided technical supervision of the experimental rig during these studies. All authors contributed iterative refinements of the experimental rig and interaction concept. Martin Schuessler and Luca Horman conducted the qualitative analysis. Martin Schuessler conducted the quantitative analysis. Paper writing was primarily shared between Martin Schuessler, and Andrew Blake. Raimund Dachsel and Carsten Rother made revisions.

6.2 Introduction

Since the beginning of the Covid-19 pandemic, video-conferencing has seen an unprecedented scale of adoption. Despite the benefits, "zoom fatigue" has become a major concern. One cause is the lack of non-verbal communication cues [13, 186], including gaze. Gaze cues serves crucial functions in regulating turn-taking, providing feedback, signalling attention, and conveying intimacy and emotions [117, 9, 105]. They increase group engagement, collective performance, and creativity [23, 187]. In today's typical video conferencing systems, each user views the other users only frontally, confined to a small screen. We refer to this kind of layout as a "Tiled View". Because of gaze misalignment, and because all users receive the same view, users cannot perceive who is gazing at whom. Conversation is measurably and palpably different from a face-to-face encounter [201, 136]. Alternative communication cues are needed [13] to help perceive others' communicative acts and avoid misunderstandings [187]. Augmented (AR) and Virtual Reality (VR) systems might

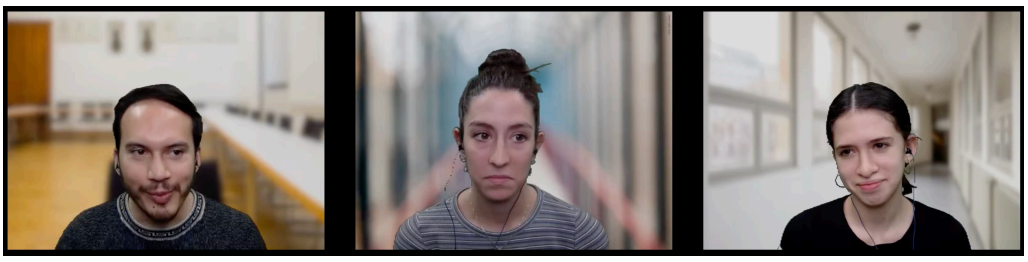
address this problem by representing users as gaze-aware avatars in an immersive environment (e.g. [177]). However, a challenge for these systems is to achieve graphical and behavioural realism, both of which are subject to ongoing research [188]. Existing AR/VR systems heavily rely on head-mounted displays which introduces a major inconvenience and a barrier to adoption. Meeting systems using cameras, on the other hand, can capture natural and realistic images of users without wearables. Two major directions have been taken to correctly convey interlocutors' gaze using cameras. The first is to use multiple screens in each workstation so that every interlocutor that is shown to the user is assigned their own, physically separated, screen and camera, preserving gaze and head rotation as non-verbal cues (e.g. [201, 172]). However, separate screens and cameras constitute a barrier to adoption and a potential usability issue. They require users to actively turn their heads to follow the conversation and they place interlocutors in peripheral vision. It has been hypothesised that this could attenuate the effect of non-verbal communication cues [201]. The second direction is to artificially modify the captured image to convey gaze [89, 243]. However, the resulting modified images often appear unnatural, and no empirical evidence exists that this approach provides any benefit over the well-established Tiled View layout.

We envision a system, *Gazing Heads*, in which each user sees the others displayed on their single-screen display, with gaze and attention conveyed through synthesised head rotation. Gazing Heads would use hardware which all video-conference users already own — a single screen, single camera, and microphone, with no need for additional displays or wearables. We anticipate that head rotation could soon be synthesisable in real-time, with sufficient realism, using software only [248, 80], once issues such as graphical realism [88], the uncanny valley effect, delays and synchronisation issues in high-quality video transmission are solved.

We built a simulation of Gazing Heads for four users using seven cameras placed around each user. The illusion of head rotation is created by transitioning between cameras. Our within-groups user study (N=80) compared Gazing Heads with Tiled View (Figure 6.1), with 20 groups of 4 participants, all tackling two different tasks. The first task was a group discussion about a controversial topic; the second was a game where participants were assigned specific roles with conflicting objectives. We used a wider variety of measures than prior gaze studies [201, 89, 183, 225, 115, 242] to gain more detailed insights. Objective measures of gazing behaviour and turn-taking were recorded, together with subjective ones, via questionnaires inspired by previous work. There were interviews at the end of each session to obtain qualitative insights.

Results show that simulated head rotation in Gazing Heads indeed conveys gaze and attention. Participants knew better when they or others were being addressed. Compared with Tiled View, users experienced a higher degree of social presence and engagement. We

(a) Gazing Heads: Round-table discussion where gaze between all interlocutors is present due to rotation of heads.



(b) Tiled View: Traditional video conferencing where gaze cues are absent.

Figure 6.1. Snapshots of the Gazing Heads (a) and Tiled View (b) simulations taken during our user study: They show the view of the fourth participant in a virtual discussion. We found that with Gazing Heads it is effortlessly apparent who is looking at whom. Gazing Heads proved beneficial for social presence and user engagement.

observed a significant increase in mutual gaze but there was no significant difference in turn-taking.

6.3 Background and Related Work

Gazing Heads would operate on a single screen without additional hardware while providing live, gaze-aware video, including third-party gaze. But what are the alternatives? There are four classes of telepresence systems that make gaze perceivable. Group-to-group systems (e.g. [165, 176, 175]) enable video-mediated communication for spatially separated groups of people. One-to-many systems (e.g. [102, 210, 110]) represent one remotely located interlocutor to a group. One-to-one systems (e.g. [129]) focus on the video-mediated conversation between two interlocutors. Virtual meeting room systems (e.g. [201, 243]) have each spatially separated interlocutor joining individually, and Gazing Heads is in this class. In standard virtual meeting room systems, the same view of each user, taken from that user's single camera, usually placed above the screen, is transmitted to all other participants. This introduces misalignments hindering gaze-awareness. *Direct eye contact misalignment* occurs when a user is being looked at but that gaze is misaligned. *Third-party gaze mis-*

alignment occurs when one user is being looked at by another user, but the observing third party does not perceive this gaze. Many methods only focus on direct eye contact but ignor third-party gaze [99, 115, 89, 71, 94, 254]. Gazing Heads addresses both issues.

6.3.1 Technology for Virtual Meeting Rooms

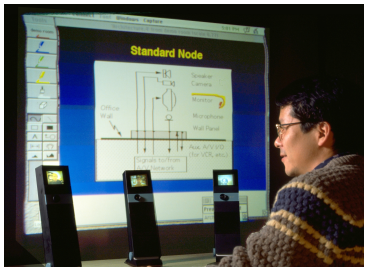
Gaze-aware Virtual meetings can also be created using Virtual Reality (VR) (e.g. [222]) or Augmented Reality (AR) (e.g. [177]) but they demand substantial hardware: tracking devices, head-mounted displays, and powerful GPUs. Moreover, avatars lack realism [18, 35, 188, 17], obscure social cues [174, 156], and uncanny valley effects are evident [174, 156].

Given that AR/VR is expensive and unrealistic, an alternative is camera-based systems, and we found five precedents, as depicted in Figure 6.2. Their technical capabilities are compared with standard video-conferencing in Figure 6.3.

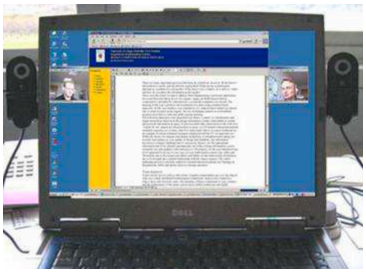
The first three systems use ante-hoc correction — avoiding gaze misalignment before images are recorded. A 1:1 physical-virtual space mapping, with dedicated cameras and displays for each interlocutor, is arranged to coincide with the virtual mapping and foster active head turning. The *MAJIC* system by Okada et al. [172] places cameras behind two life-sized projections of interlocutors (Figure 6.2a) — three users in all. The *Hydra* system by Sellen [201] uses three small screens with integrated cameras instead of projections (Figure 6.2b), supporting four users. Both systems require special hardware, and are hard to extend for more users. The *IC3* system by Sun and Regenbrecht [225] addresses some of these issues by using a single display. The three-party video conference system places the two interlocutors to the far left and far right of a screen with a camera mounted next to them (Figure 6.2c). The setup is simple and compact but cannot be extended to more than three users, nor does it address the third-party gaze problem, given that it has only two views.
The four-party system *GAZE-2* by Vertegaal et al. [243] uses a single semi-transparent display with three cameras behind it (Figure 6.2d). In principle, more cameras could be added to support further users. The 2D mages are rotated in a 3D virtual space to attempt to convey third party gaze direction, which largely fails because of the well-known “Mona Lisa effect”: the eyes of Mona Lisa gaze towards the observer, rather than rotating with the 2D display.

GazeChat by He et al. [89] is the only system which can be used on a conventional laptop without additional hardware. However it transmits animated 3D profile photos (created by neural rendering) (Figure 6.2e), without head rotations or live video. It animates users’ gaze but other parts of the face are inanimate — verbal cues and facial expressions are lost.

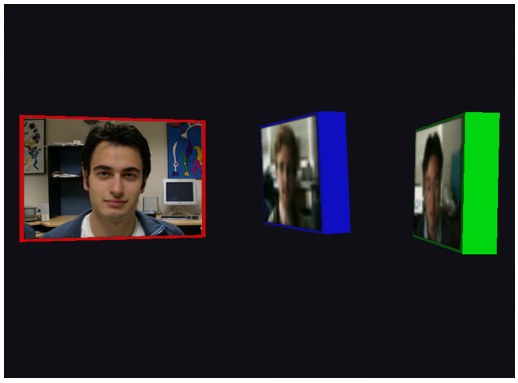
(a) MAJIC [172]: Two cameras are placed behind the life-sized projections of interlocutors on half-transparent film. The spatial arrangements of displays and cameras preserve direct eye contact and third-party-directed gaze. This setup introduces considerable hardware requirements.



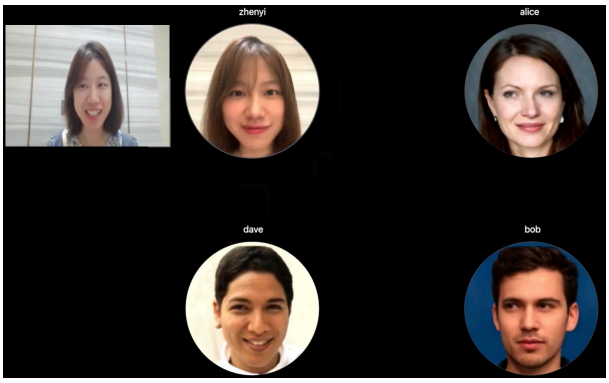
(b) Hydra [201]: Each interlocutor is assigned their own, physically separated screen and camera, preserving gaze and head rotation as non-verbal cues. The small separated displays encourage active head turning but also introduce usability issues.



(c) IC3 [225]: Using a camera for each interlocutor shown on screen, this compact three-party system allows for direct eye-contact. Observers can recognise whether they are being looked at, but third party-directed gaze is not conveyed accurately.



(d) GAZE-2 [243]: Three cameras are placed behind interlocutor video tiles. Third-party-directed gaze is conveyed by rotating 2D video tiles in 3D, which is inferior to any actual 3D rotation of the head due to distorted visual gaze cues.



(e) GazeChat [89]: A single camera per user is used for eye-tracking. Gaze information is used to animate the user profile image. In this screenshot, all interlocutors are gazing at user zhenyi. Besides the modification of gaze, the images remain inanimate. They convey fewer cues than live video. Zhenyi is smiling (top left mirror view) but that is not conveyed by her avatar.

Figure 6.2. Screenshots of five gaze-aware meeting room systems: different approaches to preserve gaze in video conferencing have been pursued, each introducing their own issues with usability or barriers to adoption. All images reproduced with the kind permission of their respective authors.

	Direct eye contact method	Third-party gaze method	Users	Displays per User	Cameras per User	Live video
Zoom	None	None	N	1	1	Yes
MAJIC [172]	ante-hoc (camera behind)	ante-hoc (camera behind)	3	2	2	Yes
Hydra [201]	ante-hoc (camera close)	ante-hoc (camera close)	4	3	3	Yes
Gaze-2 [243]	ante-hoc (camera behind)	post-hoc (approximating gaze)	4	1	1	Yes
IC3 [225]	ante-hoc (camera close)	None	3	1	2	Yes
GazeChat [89]	post-hoc (synthesised gaze)	post-hoc (synthesised gaze)	N	1	1	No

Figure 6.3. Key properties of prior gaze-aware systems in relation to standard video conferencing: Systems that use a single display (Zoom, Gaze-2, IC3 and GazeChat) have limited or no gaze-awareness support or lack live video. Systems that offer full gaze correction and live video (Hydra and MAJIC) rely on multiple displays and cameras. Note that Zoom was included as a representative of common video conferencing solutions, of which it has the highest market share [28].

6.3.2 Related Prior User Studies

Only two of the five related systems mentioned in the previous section have been evaluated in an extensive user study. GazeChat [89] was evaluated by four groups of four people (N=12), each having a group discussion. Each group tested four conditions: two variations of GazeChat, an audio-only meeting, and Tiled View with live video (within-group design). Questionnaires were employed to measure social presence, user engagement, and general user experience. GazeChat proved superior to the audio-only interface in some ways but, compared with TileView, GazeChat was worse at signalling direct eye contact and did not provide any other improvements – presumably because there was no live video. Turn-taking and gaze behaviour were not investigated in that study. It was not the objective of GazeChat to compete with live video conferencing but rather for use where live video is not an option due to privacy concerns or bandwidth limitations.

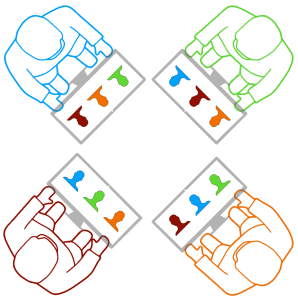
The user study conducted with the Hydra system [201] is the most relevant for our work. Twelve groups of four people (N=48) had a discussion in three conditions: face-to-face, Tiled View, and the Hydra (within-group design). A questionnaire was used to measure user experience and some aspects of social presence. Turn-taking behaviour was measured by processing participants' voice recordings but no significant differences were found between Hydra and Tiled View. Nonetheless Hydra was preferred by participants and was rated as superior for perceiving gaze and attention. The study also investigated the difference between the two video-mediated systems and face-to-face. It found that video-mediated conversations were significantly less dynamic. Hydra did show that gaze-awareness and head rotation could make attention perceivable in video conferencing and served as an inspiration for our study.

Our Gazing Heads study aims to determine whether future systems using synthetic head rotation, once the technology is mature, are actually likely to improve the video-conferencing experience. Given current technical limitations, we have built a simulation of Gazing Heads using additional cameras. The study provides more detailed insights than the prior work. It represents a substantial update in the light of modern developments in hardware and processing. It tested 4 users together, where some of the prior studies had only 3. It included a larger number of participants (N=80), employed a more comprehensive questionnaire, and also analysed eye-tracking data.

6.4 Simulating Gazing Heads

Gazing Heads was developed through a sequence of pilot studies, and is illustrated in Figures 6.1a and 6.4. Since technology is not yet mature enough and the goal was to conduct a

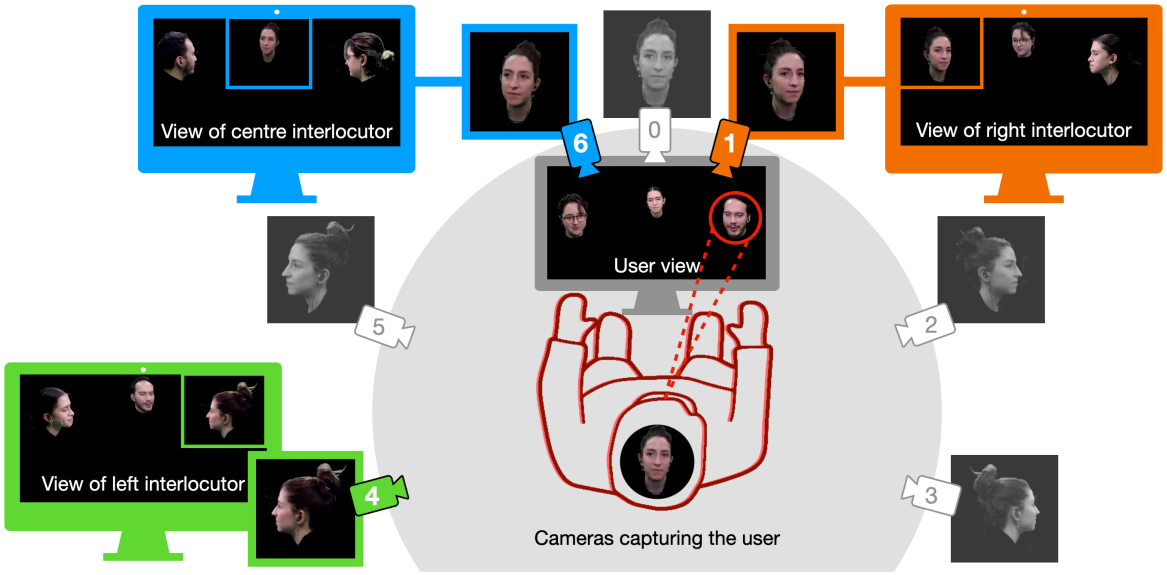
(a) *Workstation equipment*: 7 webcams, 3 lights, 4K display and eye-tracker. We built four such workstations for our study.



(b) *Virtual arrangement*: Users form a virtual circle among workstations.

User gaze at	Send view to left IL	centre IL	right IL
Left IL	Cam 6	Cam 1	Cam 3
Centre IL	Cam 5	Cam 0	Cam 2
Right IL	**Cam 4**	**Cam 6**	**Cam 1**

(c) *Gaze mapping*: We send separate camera views to each interlocutor (IL) depending on the users gaze. Figure 6.4d shows an example.



(d) *Camera angles and mapping example*: This user gazes at her right interlocutor (IL). Based on the mapping shown in Figure 6.4c, we send the following camera views: Direct eye contact via camera 1 to right IL, slight head rotation via camera 6 to centre IL and strong head rotation via camera 4 to left IL.

Figure 6.4. Setup of Gazing Heads: We use eye-tracking to convey gaze by sending different camera views to each interlocutor. Note that in all subfigures, cameras are consistently numbered $(0-6)$, and colours denote interlocutors and their respective views.

user study to evaluate the concept, rather than providing a full implementation, we did not yet build a system that uses only a single camera.

6.4.1 Sitting in a Circle

Interlocutors (IL) were arranged on a single screen to enable ready perception of gaze. We created the illusion of a circular arrangement which encourages the exchange of gazes [98] by displaying the middle one of three interlocutors slightly smaller and higher than the others (Figure 6.1 top). On each user's workstation, the other users are displayed on the left, centre, and right, in a way that is consistent with four users around a table. The user in each of these three positions can turn to any of the other three. Hence, nine rotation angles for every user are needed: three positions on the screen, each with three unique head rotations. Note that we did not include a self-view as they are believed to have numerous negative effects [13], such as absorbing visual attention [73] and reducing the perception of others' emotional responses [204].

Available computer vision methods to synthesise rotation angles from a single camera perspective face substantial technical challenges: low realism [207], artefacts [109, 70, 260], and limited rotation angles [248, 70]. Some more advanced methods do not work with participants that wear glasses or have long hair [35]. We therefore used several cameras to obtain the required views (ante-hoc gaze correction) and switch between them. In principle nine cameras are needed — one for each of three screens, then one for each of three head rotations. By careful selection of viewing angles, we reduced nine cameras to seven, with two doing double duty — cameras 1 and 6 in the illustration of Figure 6.4d.

6.4.2 Gaze Switching

Dedicated hardware (Tobii Eye Tracker 5) tracks gaze on a 4k 27-inch display (for details on accuracy, see Appendix E.6). Each user's screen splits into focus areas to map gaze to camera views (Figure 6.5). When a user changes gaze, a camera transition is triggered appropriately. Users typically switch gaze about every second, often just glancing briefly which often goes unnoticed in physical settings [9]. When a gaze switch is detected, two criteria are continuously assessed: gaze duration on the same user for at least $750ms$ (dwell time) and a $2000ms$ lapse since the last transition (refractory period). The moment both are met, a new transition is initiated. This approach avoids over frequent transitions, yet allows quicker gaze switches following a period without transitions. Both values were determined through pilot studies. Instant transitions between camera views are visually distracting, so scaled alpha-blending was used to blend the current view with the subsequent view:

$$g(x) = (1 - \alpha)f_0(x) + \alpha f_1(x), \tag{6.1}$$

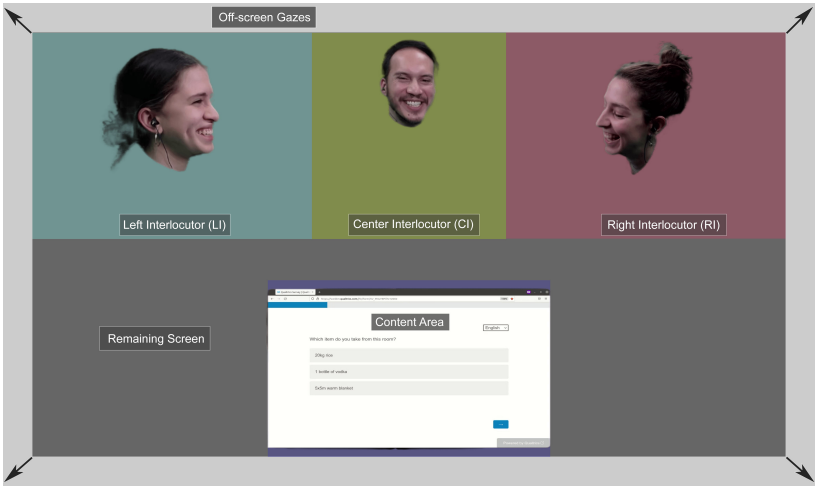


Figure 6.5. The six focus areas of the Gazing Heads prototype in the game task: We used these areas to decide which camera view to send to the other interlocutors. We also used them in the eye-tracking analysis of our user study. Participants changed their gaze from one area to another roughly every second. Note that in the discussion task, the content area was removed.

where g is the new image and f_0 and f_1 are the images that are blended. Here α is increased from 0 to 1 as time t increases from 0 to T:

$$\alpha = (e^{\frac{3t}{T}} - 1)/(e^3 - 1). \tag{6.2}$$

This scaled exponential function for α gives a strong visual hint initially of the transition and then smoothly fades out. We tested several blending functions and found this one to be an effective compromise between rapid initial signalling of change in attention and maintaining an illusion of smooth head rotation.

Consistency Between Camera Views We configured exposure time, focus, and white balance to capture visually consistent images from every view at a constant frame rate for seamless transitions. Three cameras were mounted above the screen on a DIY rack. Four cameras capturing the side views were mounted on microphone stands of equal height. Consistent diffuse lighting from all angles was obtained from two softboxes (85W compact fluorescent lamp light bulb, 5500K) mounted overhead and an LED ring for facial illumination (35W, 5500K). Each camera captures a different background so chroma-keying with U-shaped green screens removes the backgrounds. Chroma-keying removes all regions of an image matching a specific colour range. We used a U-shaped green background screen but faced the problem that the cameras recording the side views needed to face the green screen. To also remove them, we processed the image so that only the largest continuous region, which was always the user, would be left in the final image. The use of a green screen

is unrealistic in a commodity implementation but acceptable in a simulation. Note that we also explored *background matting* [133] using prior knowledge learned via a deep neural network to separate images into foreground and background. Even with a GPU the method introduced considerable delay. Sometimes the background would "leak" through glasses or hair. This method did not satisfy our requirements for our simulation. A black background strengthens the illusion of 3D head rotations since heads have a rough ellipsoidal shape. It also masks small errors in background segmentation.

View Stabilisation Interlocutors are centred and scaled to similar size in their respective views by software that tracks faces after background removal, and crops them, with temporal filtering to suppress jitter. This is done gracefully over time to allow for some margin for natural head movements (e.g. leaning into a conversation).

Shoulder Removal A typical user shifts gaze by rotating the head and changing eye-gaze but alters upper body posture only slightly. Switching between camera views creates a "stiff-necked" illusion of substantial upper body rotation, and that looks unnatural. Therefore participants wore green turtlenecks so the chroma-keying background segmentation omitted neck and shoulders. This enhanced the illusion of head rotation since heads have a rough ellipsoidal shape.

Low Latency Architecture We explored several design alternatives to reduce latency and skew while maintaining a frame rate of 30 Hz. Ultimately we chose a client-to-client architecture, sending 12 separate audio and 12 separate H264-encoded video streams over the local network without synchronising them. Measured latency was 80.0 ± 0.2 ms for audio and 133.33 ± 33.33 ms for video (see Appendix E.7 for details). Consequently, skew is 53.33 ± 33.53 ms, falling within ITU recommendations G.114 and BT.1359-1 [69, 30]. It also outperforms common video conferencing solutions for which Xu et al. [256] reported delays of 130 to 270 ms for audio and 230 to 270 ms for video latency, for common multiparty video conferencing solutions. (For details on latency measurement see Appendix E.7.)

High Quality Spatial Audio To ensure high-quality audio, we used Røde Lavalier GO microphones and In-Ear headphones with a feedback loop from the microphone to the headphones ensure high quality spatial audio, matching the direction of perceived audio approximately to speaker location.

6.5 User Study Methodology

We conducted a user study to test our concept and investigate whether head rotation can be used for conveying gaze and attention in video conferencing. We also wanted to understand how Gazing Heads influences users' experience and communication behaviour. Concretely, we expected Gazing heads to influence turn-taking behaviour due to the regulatory function of gaze as a turn-taking cue [105, 54, 55, 117]. We also expect it to create a more personal, intimate and immersive experience [117] leading to an increase in social presence [136, 37, 206]. Lastly, we were investigating whether Gazing Heads influences users engagement [36] and gazing behaviour. Two tasks — discussion and game — were used with a wider variety of measures than previous gaze-awareness studies [201, 89, 183, 225, 115, 242], and with the largest number of participants ($N = 80$) used in studies of this kind to date. Participants used the two different systems (Gazing Heads and Tiled View) in a within-group design.

6.5.1 Experimental Setup

The Gazing Heads simulation as described above served as the *treatment system*, with the Tiled View as a comparative *baseline system*. We did not compare against other gaze-aware solutions as no empirical evidence exists that they would outperform the Tiled View (e.g. [201]) but instead may perform worse (e.g. [89]). In Tiled View, three equally sized video tiles were placed at screen locations similar to those used in Gazing Heads. Head rotation was disabled and only central cameras were used, ignoring the 6 off-centre cameras. Background segmentation remained active but without the green turtlenecks, so the upper body was not segmented out — see Figure 6.1 on page 85.

We recruited 42 male, 35 female, and 3 non-binary or non-conforming gender subjects, and relatively young with $76\,\%$ younger than 25 and only $6\,\%$ 35 years or older. Participants were accustomed to video conferencing; $96\,\%$ of them used it at least once a month and $81\,\%$ at least every week. One group conducted the experiment in Russian, two groups in Spanish and the remainder in German (all native speakers). Thereby we ensured proficiency in their respective languages. All participants received a 30€ Amazon voucher for their participation.

6.5.2 The Group Discussion and the Survival Game

Prior studies have mostly used group discussion [201, 89, 175, 225, 183, 115] – though a few used collaborative problem-solving [242, 241, 238] – and it is evident that task type can significantly affect behaviour. Group discussion as in Sellen [201] resulted in higher turn-taking frequencies ($3.9 - 4.3\,\mathrm{min}^{-1}$) than the problem-solving task used by Vertegaal

et al. [242] ($1.0-1.3\,\mathrm{min}^{-1}$). In a problem-solving task there is reduced need for eye contact which may well reduce the likelihood of a measurable effect [225].

Group discussion has good external validity because it represents a real-world situation. One drawback is reduced internal validity as participants' prior knowledge, individual traits such as extroversion, and group dynamics, can influence the conversation. In extreme cases, pilot studies showed one or two participants holding the floor most of the time, masking the effects being measured. We nonetheless included group discussion, for comparability with other studies [201, 89, 175, 225, 183, 115]. We also included the problem-solving task for increased internal validity.

Controversial Group Discussion

Five controversial statements were tested:

- Research and development of brain-machine interfaces, such as Elon Musk's Neuralink, should be prohibited or at least placed under strict regulation, as reading one's thoughts has dramatic ethical implications.
- Covid vaccination should be compulsory for all those who are not expected to suffer long-term adverse health effects from vaccination.
- Industrial livestock farming should be progressively banned.
- Short-distance flights should be banned or taxed heavily.
- Physically healthy people should have the right to euthanasia (e.g. by taking a deadly pill under a doctor's supervision) if this is their own explicit wish.

In each group, prior to the main experiment, participants rated agreement with each of five statements. The two statements whose ratings varied the most were selected as the two topics for discussion. Then for each topic, participants were instructed to find consensus as a group within five minutes. They were not stopped dead at five minutes, to avoid lowering engagement for later tasks. Instead, we interrupted the task when the current speaker(s) finished their turn or the group reached an agreement.

Game: Surviving in the Wild

Our second task was designed to:

- incentivise participants to take the floor;
- make turn-taking more dynamic and more evenly distributed;
- encourage participants to pay attention to non-verbal communication and understand others' intentions.

Players travelling to a remote island have to reach consensus on the choice of items shown on screen, to bring with them (Figure 6.5). We instructed them that several items would be crucial for survival. At each stage, they picked one item out of three they take with them. Once consensus, based on a majority vote was reached, we present the next set of items. They were given seven minutes to agree on as many items as possible. One player was randomly selected to be a clandestine saboteur so they would be incentivised to focus even more intensely on one another (details in Appendix E.2). Again we allowed players to reach agreement rather than stopping the game dead at seven minutes.

6.5.3 Three Questionnaires: Semantic Differential, UX, and Comparative

We reviewed similar studies for potential questions [259, 201, 23, 158, 238, 206, 86, 259] and categorised them by the concept or property dimension they measure. The most relevant were: presence, turn-taking, engagement, user satisfaction and usability. No existing questionnaire covered all dimension we were interested in so that we selected a few questions for each dimension to create two custom questionnaires. The *comparative questionnaire* asks about seven aspects of system preference relating to turn-taking, perceived attention and naturalness of interaction (see Figures E.3, in Appendix E.3 for details). The *UX (User Experience) questionnaire* asks about direct eye contact, directed third-party gazes, and "off-gazes" directed at no one. In addition a standard *Semantic Differential* questionnaire measures social presence [206]. The UX and Semantic Differential questionnaires were filled out twice, once after using each system.

Presence Lombard and Ditton [136] refer to "presence as social richness" which we call *social presence* which is measured by the semantic differential questionnaire. They term "presence as transportation [to a virtual room]" which we call *virtual presence* and which is measured in the UX questionnaire via a question about the feeling of "being in the same room". Two statements were included about the perceptibility of interlocutors' reactions and becoming acquainted with them. The comparative questionnaire asked which system participants would use for persuading others. These questions are known to be correlated with social presence [37, 206]. We also added a question about which system was considered to be more social.

Engagement Two statements about participants' excitement and the interactiveness of the conversation were included in the UX questionnaire. We also added a question to the comparative questionnaire measuring participants' satisfaction with their contribution to problem-solving [238], and one addressing engagement/excitement.

Usability During pilots, some participants complained that they felt excluded when interlocutors turned away from them. Others found the head rotation and camera transitions distracting. We added two questions addressing these issues to the custom UX questionnaire.

Overall Preference and Willingness to Adopt The comparative questionnaire asked which system allowed for a more natural interaction — this can be seen as a measure of presence [252], and also as a high-level quality metric for video-mediated communication. As indicators of overall user experience, participants were asked which system they would recommend to others and which system they wanted to use for a final interview.

6.5.4 Procedure

Upon arrival, participants were instructed that the goal of the study was to evaluate two video conferencing systems. Each workstation was calibrated to the individual participant. Once set up, participants filled in a demographic questionnaire and rated their agreement to the five controversial group discussion statements. A server gathered live experimental data, administered conditions and questionnaires, supervised the experiment, and channelled communications to participants. Each session was screen-recorded . A full factorial within-group design was used, each group using both Gazing Heads and Tiled View systems for the group discussion, and for the game. The system to use first was distributed evenly across groups. Each session began with the group discussion, followed by the game. Then, switching systems, there was a game session followed by another discussion on a different topic. The discussion lasted about 6 minutes ($M = 5{:}58$, $SD = 1{:}27$) and the game took around 8 minutes ($M = 8{:}22$, $SD = 2{:}16$). The social presence and the UX questionnaire were filled out after using the first system and then again after using the second system, together with the comparative questionnaire. A final interview of roughly 14 minutes ($M = 13{:}37$, $SD = 6{:}49$) was conducted using the system favoured by the majority. In case of a tie, the group was asked to reach a consensus on which system should be used. Altogether the experiment took $90-120$ minutes, which included calibration, answering questionnaires and extensive Covid-19 protection measures.

Covid-19 Protection Measures We took extensive measures to comply with local Covid regulations. The selection of possible participants was restricted to vaccinated or negative tested members or guests of the university. They were required to wear an FFP2 mask until they were alone in their assigned room. We ventilated all rooms prior to any experimental session. Further, we disinfected all materials and surfaces that participants interacted with. Based on the finding of Abraham et al. [2], turtlenecks were ensured to be SARS-CoV-2

virus free by treating them for at least 5 min with 65°C hot air or 20 min with 40° hot water. Further, we used waterproofed in-ear headphones cleaned and disinfected in a 15 min ultrasonic bath using a 4% Instrusol AF+ solution. Our governing authority approved our documented safety measures before conducting the experiment.
(Additional details of the study procedure can be found in Appendix E.1.)

6.5.5 Data Analysis

Speech We recorded participants separately in each station. Quantitative measures for video-mediated communication are generally based on simultaneity of speech and on turn-taking [235]. We used the definitions of terms and measures from Sellen [201], which are sensitive enough to detect difference between a physical and virtual conversation.

> A *turn* consists of the sequence of talk-spurts and pauses by a speaker who "has the floor". A speaker gains the floor when they begin speaking to the exclusion of everyone else and when they are not interrupted by anyone else for at least 1.5 seconds. The duration of a turn begins with the first unilateral sound, and ends when another individual turn or a "group turn" begins. Note that turns therefore include periods of mutual silence at the end of utterances, when no one else has yet taken the floor.
>
> A *group turn* begins the moment an individual turn taker has fallen silent and two or more others are speaking together; the group turn ends the moment any individual is again speaking alone.
>
> *Simultaneous speech* is speech by one or more speakers who do not have the floor. [We] further distinguish between overlaps and simultaneous speech which do not lead to a speaker switch. Simultaneous speech which does not precede a speaker switch is called non-interruptive simultaneous speech.

We converted any absolute measures to frequency or proportion measures as our session duration varied in length. Audio streams were analysed using a Voice Activity Detector [228], based on a deep learning transformer model. Pure laughter, detected by a residual neural network [77, 191], was distinguished from speech activity. Manual annotation separated pure laughter from any utterances made while laughing.

Gaze Times We defined six layered rectangular gaze areas as in Figure 6.5 corresponding to: the three participants, the display area for the game task, the rest of the screen, and off-screen. Once again, dependent variables for duration were converted into frequency of occurrence or proportional duration.

Shared Eye Contact Shared eye contact is perceivable in Gazing Heads (but not in Tiled View) and we investigated whether this affects how long and how frequently participants gaze at one another. A mutual gaze event begins when participant A focuses on the area where participant B is displayed on screen A, and vice-versa. It ends as soon as either is focusing on a different area.

Significance Tests For ordinal self-reported data, we used a single-tailed proportions test ($p_0 = 0.5$) for comparative questions and Wilcoxon signed-rank test for all others. For comparative questions Participants had the option, in the questionnaire, to like both systems equally. In those cases votes were omitted from analysis, treated solely as an indicator of uncertainty about that question. For social presence, the reliability of the semantic differential attributes to measure the underlying concepts was assessed Cronbach α. The social presence score was calculated as the median score of the four attributes. We used a factorial repeated measures ANOVA for the metric speech and gaze measures, assessing the assumptions of normality and sphericity with Shapiro-Wilk and Mauchly's tests, respectively.

Interview Insights Thematic analysis [29] was applied to recordings of group interviews, and captured using a coding scheme. A scheme was devised with three major domains of technology and experience and a total of 11 topics described by 47 codes. Appendix E.5 provides descriptions of the domains, their topics and the codes used. Codes were applied, checked and resolved by two of the authors.

6.6 Results

Here are the main results from the three sources of data: recorded speech, gaze data, and user experience questionnaires.

6.6.1 Speech Activity

Only one significant difference between systems was found in speech activity, that group turns occur 21% more frequently in Tiled View than in Gazing Heads, as Figure 6.6 shows. However they are rare and so are not a strong indicator of difference in turn-taking behaviour. No significant differences are observed between systems for the other 10 turn-taking measures, refuting our assumption that Gazing Heads would make speech activity more akin to physical rather than virtual interaction. Between tasks, there are significant differences in 10 out of the 11 measures, indicating that the game was a more dynamic task:

- Turn switches occurred more frequently.

- Turns were shorter.
- Group turns occurred more frequently.
- Less time was spent with only one person speaking.
- More Simultaneous speech occurred.
- More non-interruptive simultaneous speech took place.
- Turns were more evenly distributed.
- No differences in simultaneous speech taking control.
- Higher overlaps in speaker switches.
- Reduced switching times.

Those differences confirm that the measures are indeed sensitive to changes in communication behaviour.
(Note 1 of the 20 sessions didn't produce speech data usable for analysis due to recording issues.)

6.6.2 Gaze and Eye Contact

Eye contact occurs just a bit more frequently (Figure 6.7) with Gazing Heads, both during the discussion ($+8.9\,\%$ more frequently) , and the game ($+3.6\,\%$) ($p < 0.01$). Such mutual gaze as *did* occur in Tiled View was also not so useful — participants were "somewhat not" able to distinguish whether gaze was actually directed at them (Figure 6.11). Change of focus would be another potentially interesting variable, happening on average every second. However there is no significant difference across tasks or systems. As an aside, there are marked differences between tasks. For example mutual gaze is significantly reduced in the game task, for both systems ($p < 0.001$), because more time is spent in the game gazing at the content area which contains the information needed to play (Appendix E.4 provides a more detailed analysis).

6.6.3 Overall System Preference

Users prefer Gazing Heads for most aspects of interaction, as Figure 6.8 shows. It is more engaging (87%) and makes it easier to sense the attention of others (91%). There is a clear tendency to choose Gazing Heads for the final discussion session (72%). Participants tended to find Gazing Heads more natural (62%) and social (62%), suitable for a persuasive discussion (51%, $p = 0.015$), and to provide a better turn-taking experience (49%, $p = 0.052$). Gazing Heads is favoured for facilitating interactive conversation, getting to know people

		Discussion Task		Game Task		p-Value	
	Hydra [201]*	GH	TV	GH	TV	Task	System
Turn Frequency per Minute	4.29	2.69 (1.2)	2.75 (0.9)	6.31 (1.1)	6.32 (1.2)	$<$ **0.001**	0.869
Turn Duration	16.62 s	24.48 s (9.8)	22.97 s (10.5)	8.04 s (1.8)	7.74 s (2.0)	$<$ **0.001**	0.535
Group Turn Freq. per Minute	0.24	0.22 (0.2)	0.23 (0.2)	1.43 (0.6)	1.76 (0.8)	$<$ **0.001**	**0.037**
Turn Distribution (H)	1.83	1.90 (0.1)	1.90 (0.1)	1.95 (0.0)	1.97 (0.0)	**0.001**	0.620
Time one Person spoke	74.70 %	90.5 % (4.4)	90.3 % (3.3)	73.0 % (6.0)	70.9 % (5.8)	$<$ **0.001**	0.300
Simultaneous Speech	5.40 %	2.5 % (3.0)	3.3 % (2.3)	11.1 % (5.4)	13.2 % (6.5)	$<$ **0.001**	0.062
non-Int. Simult. Speech	10.22 %	2.4 % (2.5)	3.1 % (1.9)	10.9 % (4.7)	11.9 % (5.1)	$<$ **0.001**	0.187
Interruptive Simult. Speech	3.50 %	1.0 % (1.3)	1.1 % (1.0)	4.0 % (1.8)	4.5 % (2.1)	$<$ **0.001**	0.276
Sim. Speech Taking Control	41.60 %	30.9 % (21.5)	22.8 % (14.0)	31.3 % (8.5)	30.5 % (6.9)	0.113	0.212
Speaker Switches Overlaps	43.50 %	23.7 % (14.9)	25.6 % (17.9)	45.0 % (12.2)	44.3 % (9.3)	$<$ **0.001**	0.810
Switching Time	0.25 s	1.00 s (0.8)	0.69 s (0.7)	0.18 s (0.4)	0.14 s (0.4)	$<$ **0.001**	0.171

Figure 6.6. Speech analysis. The middle four columns compare our two systems GH and TV within the two tasks. Differences between systems are significant only for the group turns. Differences between tasks are significant for most measures, indicating that the measures are sensitive. Figure E.4 in Appendix E.3 has additional descriptive statistics along with F values.

* The first column shows the results of Sellen's study with the Hydra system for comparison [201]. Interestingly, even after 30 years, measures are broadly consistent with our measurements, lying between the value for the discussion and the game.

($p = 0.01$) and for making the discussion more exciting. Overall, however, participants were undecided about recommending Gazing Heads, in its current state, over Tiled View. *(All reported differences are significant at $p < 0.001$, unless otherwise stated.)*

	Discussion Task		Game Task		p-Value	
	GH	TV	GH	TV	Task	System
Focus Changes per Minute	58.25 (24.0)	58.34 (22.9)	59.28 (19.9)	62.49 (22.5)	0.109	0.053
Eye Contact per Minute	30.71 (13.2)	28.21 (10.9)	13.60 (6.9)	13.13 (7.0)	$<$ **0.001**	**0.008**
Eye Contact (% of Session)	27.8 % (10.9)	26.0 % (9.6)	10.1 % (5.3)	9.3 % (4.9)	$<$ **0.001**	**0.035**
Eye Contact Duration	0.58 s (0.2)	0.58 s (0.2)	0.46 s (0.1)	0.45 s (0.1)	$<$ **0.001**	0.548

Figure 6.7. Eye-gaze analysis. Four columns comparing the average measures obtained in the two tasks with our two systems. There was a small but significant system effect: participants had more eye contact when using Gazing Heads. The mean and sd were aggregated across tasks for column 3–4 and system for column 6–7

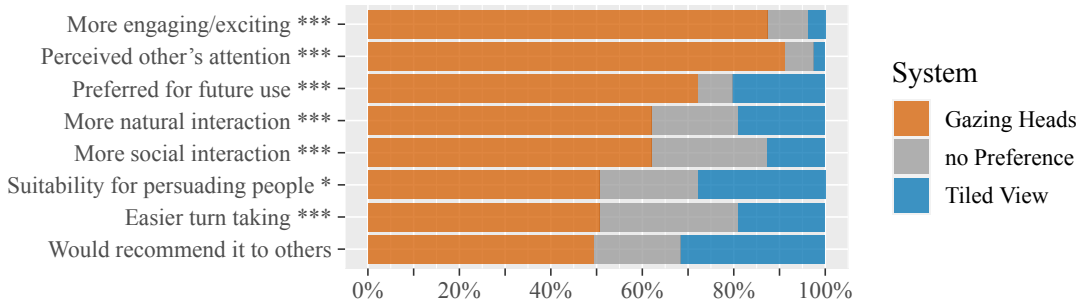


Figure 6.8. Participants' system preference ratings. Participants generally preferred Gazing Heads (significance levels: p < 0.05 *, p<0.01 **, p<0.001 ***).

6.6.4 Social Presence, User Experience, and Awareness

The four attributes used to measure social presence are reliable with Cronbach $\alpha = 0.85$ — they are plotted in Figure 6.9, together with an overall score. The first conclusion is that participants experienced a higher degree of social presence with Gazing Heads ($p = 0.02$), and 62% of participants preferred it over the Tiled View for a more social interaction. Figure 6.10 shows us exactly which aspects of social presence are different: Gazing Heads made it easier to get to know people ($p = 0.01$); and also made them a little more aware of others' presence ($p < 0.01$). With both systems, participants found it exciting to follow the discussion and perceived the conversation as highly interactive; still, they significantly preferred Gazing Heads ($p = 0.04$, $p < 0.001$). Gazing Heads makes participants feel significantly more in the same room as one another ($p < 0.001$), and although it is perceived as "somewhat" safe from distraction, it is not felt to be as safe as Tiled View ($p < 0.001$). The

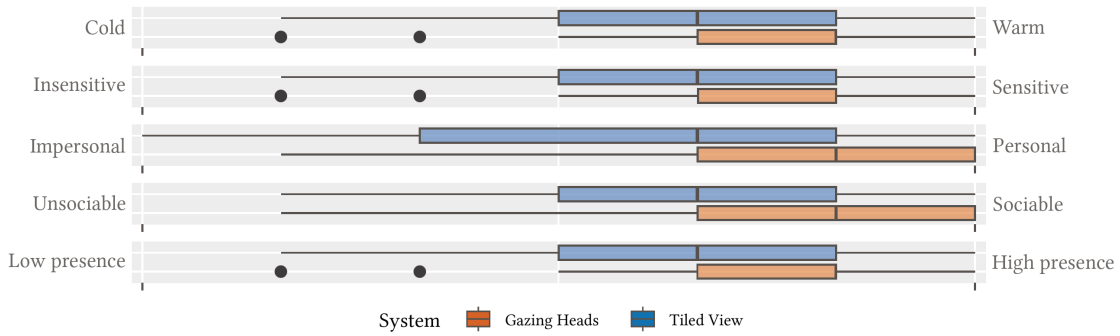


Figure 6.9. Participants' social presence ratings. The social presence score was based on participants' median ratings of four attributes (sociable, personal, sensitive, warm) for the two systems. Scores are significantly higher for Gazing Heads ($p = 0.02$). (Dots outside the whiskers indicate outlier scores.)

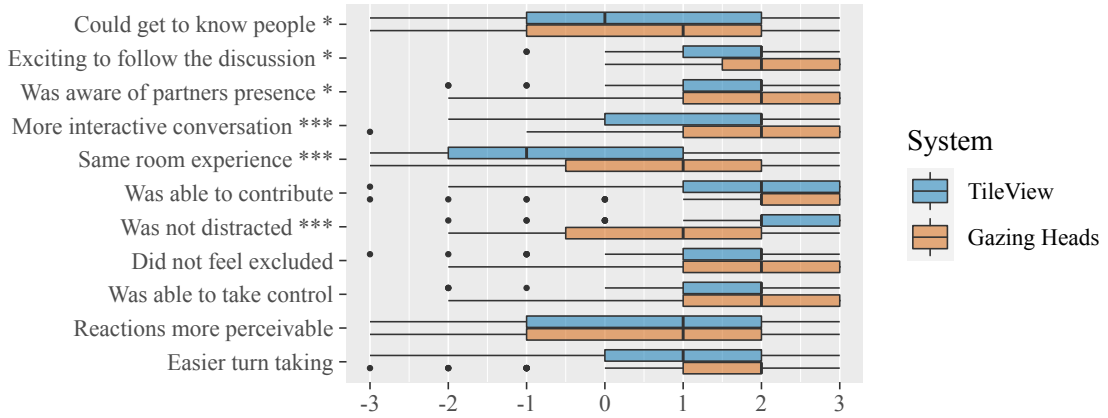


Figure 6.10. Participants' user-experience ratings. Ratings for several aspects of the user experience ratings. The scale ranges from strongly disagree (-3) to strongly agree (3). (Significance levels: p < 0.05 *, p<0.01 **, p<0.001 ***.)

two systems are similar in several respects: the ability they give an individual to contribute to their team's solution ($p = 0.246$); degree of exclusion from the conversation ($p = 0.790$); and ability to take control of it when they want to ($p = 0.176$). The reactions of others are only "somewhat" perceivable ($p = 0.263$) in both systems. Turn-taking is "easy" with Gazing Heads, and "somewhat easy" with Tiled View, but the difference is not significant ($p = 0.084$). Gazing Heads also performs significantly better on awareness of gaze and attention, and awareness of who is being addressed (Figure 6.11), with 91% of participants reporting that they perceived attention more easily (Figure 6.8, $p < 0.001$). They also found it easier to reason about who is being addressed or attended to (Figure 6.11, $p < 0.001$ for all six ratings). It was unclear whether either system could convey disengagement — *i.e.* "gazing at no one".

(Note: The responses of 1 group were excluded from the analysis due to a procedural error.

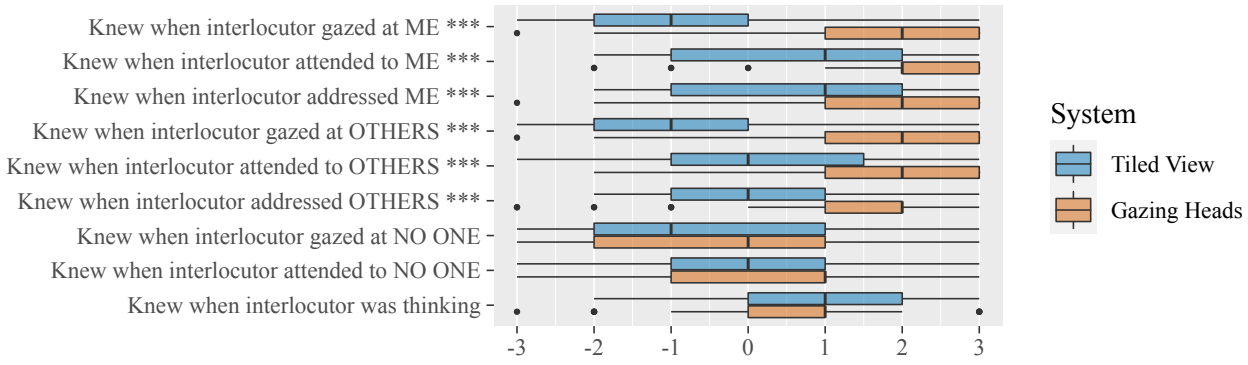


Figure 6.11. Participants' awareness ratings. Ratings for perceiving gaze, perceiving attention, and awareness of being addressed. (Significance levels: $p < 0.05$ *, $p<0.01$ **, $p<0.001$ ***).

One participant lost his replies by accidentally logging out of the system.)

6.7 Discussion

Our study has shown that the Gazing Heads concept improves video conferencing, providing gaze awareness via synthesised head rotation. This section explores the implications of results from the previous section, making particular use of insights gained from the exit interviews.

6.7.1 Conveying Attention

Participants (all but 2) understood intuitively, within the first two minutes of using Gazing Heads, that head rotation conveys visual attention. The questionnaires showed already that they perceived attention more readily in Gazing Heads, and during interviews they often mentioned that Gazing Heads conveyed attention better ($N = 36$), made it more perceivable ($N = 69$), and made it easier to gain attention ($N = 16$). *(Here N denotes the number of participants for which a code was applied at least once.)* One **implication** is that correcting a user's frontal view without adding head rotation as in [99, 115, 89, 71, 94, 254] is insufficient to achieve gaze awareness. The user study for Gaze Chat [89] which only used such a correction showed that it did not improve users ability to perceive attention. Hydra [201] did use head rotation, and did find an improvement. Users' accuracy in estimating the gaze of others may be low [117] so head rotation provides an additional cue for attention [88].

6.7.2 Subjectively Higher Engagement

In the last section, we saw that Gazing Heads felt more engaging and this was a common topic in interviews ($N = 39$). Interviewees occasionally even felt social pressure to participate because the signalling of attention was so clear ($N = 32$):

> "If your goal is that everyone is taking part in a discussion, you HAVE TO use this system. Because you are simply forced to stay engaged." "[Especially] when everyone starts looking at one person."

Our findings contrast with the studies of Hydra [201] (and GazeChat [89]) where no improvement of engagement over Tiled View was found. With Hydra, users needed to actively turn their heads and were reluctant to do that. Gazing Heads translates almost every gaze into a head rotation and makes them salient by placing them in users central vision rather than in peripheral vision on small screens. We believe this increased users engagement.

6.7.3 Conveying Disengagement and Gazing-Away

During interviews, users took issue with the camera selection rule which showed them turned towards another participant, even when they were not gazing at anyone ($N = 11$):

> "The only thing that is missing, perhaps, is that you cannot look at nobody. So even if I am just staring in front of me, then I am still [being displayed as] looking at someone."

They also want to be able to disengage, which Gazing Heads does not facilitate:

> "if people could like watch me very closely ... when I am looking at my phone or something like that [...] in larger groups when you want to disengage from the discussion ... then it's almost creepy."

To address these concerns, a future system could add a neutral object like a table, as suggested in interviews ($N = 10$), similar to the content area which for us was present only during the game. However, this would risk reducing mutual eye contact and attenuating the effects of gaze cues [10] (see also the gaze analysis in Appendix E.4). One might also show disengaged users as greyed, inanimate or separated, allowing them to be listeners only, or to work on something in parallel. Systems could offer different layouts depending on whether high engagement is a major objective, and future work might look into that.

6.7.4 Increased Social Presence in a Virtual Space

We saw in the previous section that Gazing Heads increases the feeling of social presence. Also in the interviews, an increase of social presence was frequently ($N = 58$) mentioned as

an advantage of Gazing Heads. It was usually described as a feeling of "being closer" to the other participants or having a more personal experience. Similarly, interviewees described improved virtual presence ($N = 50$) but were divided as to whether the improvement was substantial enough to feel "in the same room" ($N = 23$), or not ($N = 14$), for example:

> "With heads floating in a black room it is totally unclear in which real distance we are actually located to one another [...] we are all floating in this empty black thing."

or

> "Well I still think the situation ... uhm ... is still video-telephony .. and that's just not so immersive that it completely detaches you [from your current surroundings]."

It was also evident that the concepts of social and virtual presence are considered as related (see also Lombard and Ditton [136]), for example:

> "Here you feel closer, because it's [rather] simulating a room and not just [a] screen."

As for other design aspects that may contribute to physical and social presence, participants frequently praised the consistent placement of interlocutors in a virtual circle ($N = 15$). Participants also commented that the separate backgrounds in Tiled View were a strong visual indicator that they were not in a shared space ($N = 16$). For Gazing Heads our design choice of using a plain black background was a frequent topic. Several participants disliked it ($N = 21$) because it was perceived as cold, providing too little context. Others preferred it in plain black ($N = 8$) for increasing contrast and reducing distraction. Various suggestions for alternative backgrounds were made during the interviews ($N = 25$).

> "This [the black background] is optimal considering potential distraction, but you have to meet people where they are and it would possibly be easier to get used to the system if you – I don't know – would have a typical Zoom background."

Our study leaves open how virtual backgrounds impact physical and social presence. Future research could examine whether a photorealistic scene with a unified background and elements like a table or a bonfire enhance presence and convey virtual togetherness.

6.7.5 Subjective Effect of Eye Contact

Although participants spent only *a little* more time gazing at one another in Gazing Heads, eye contact changed their perception substantially. Since gaze is hardly noticeable in Tiled View (see questionnaires) any "eye contact" was probably guided by auditory perception not visual cues. In interviews ($N = 7$) eye contact in Gazing Heads revealed a clear, positive effect on participants' experience:

> "For a discussion, the second system [Gazing Heads] is much more comfortable. You just have more of a feeling of being part of a group. For me, in seminars, [...] others were just sitting inside of their tiles simply looking [somewhere], not really taking part. Here you have the feeling of being integrated, even if you are not saying anything, especially when you are being looked at."

Interviews suggested, as also indicated by prior work [187], that eye contact in Gazing Heads facilitated a stronger feeling of engagement and social presence:

> "I think it's more personal because, when you are looking at me, I have the feeling you are actually looking in[to] my eyes, and then I want to [...] explain my point of view to you [...] and I know [...] if you are looking at my face you will see also my eyes."

Our observations from recordings suggest that synthesised head rotations make eye contact and attention selective and salient. They *amplify* associated positive effects such as increased social presence and higher engagement.

6.7.6 Comparison to Other Systems

So far we have seen that Gazing Heads conveys attention and improves engagement and social presence, unlike previous studies. What are the reasons for this difference? GazeChat did not improve users ability to perceive the attention of others, when compared to the Tiled View and no improvement in engagement, eye-contact or social presence could be measured [89]. This may partly be because GazeChat provides no live video. However, the mock-up study of He et al. [88] provides evidence that head rotations are superior to gaze correction for conveying attention. This is further backed up by Hydra's [201] user study, which used head rotations and found it improved users ability to perceive attention compared with TiledView — but no positive effects on social presence and engagement emerged here either. We argue that synthesised head rotations convey attention most clearly. Our interpretation is that they *amplify* the positive effects associated with gaze-awareness. A head rotation towards the new person in focus combined with instant direct eye contact creates

strong observable reactions evident in many of our recordings. An **implication** for future gaze-aware systems is that correcting a users frontal view without adding head rotation [71, 94, 254] is insufficient.

6.7.7 Inconclusive Results Regarding Turn-Taking

The speech and turn-taking analysis saw no significant difference between the two systems for 10 of the 11 measures suggested by Sellen [201], nor did the questionnaire. This may be explained by video recordings of sessions, in which participants seemed to rely more heavily on auditory cues for turn-taking. Participants gave partly contradictory answers about turn-taking: 51% found Gazing Heads easier with 19% for Tiled View, and 30% had no preference, the largest degree of ambiguity we observed in any direct comparison question. Participants made several comments in interviews about turn-taking being easier with Gazing Heads ($N = 28$):

> "Who is talking next or who is generally talking right now ... emerged organically." "Yeah right [...] when you started talking and you notice the others are looking at you or are turning towards you, then you knew ok I can talk right now."

and

> "You could directly address people. When I asked [other participant] his name, I just looked at him and that worked."

6.7.8 Camera Transitions Can Be Distracting

Using transitions between camera views instead of computer vision methods to synthesise head rotation was a limitation of our simulation, necessitated by technical limitations. Implementing the Gazing Heads concept this way came with limitations on its own, which may have negatively influenced the results of our experiment. In the previous chapter we saw Gazing Heads rated as less protected from distraction. Interviews indicated that this was due to the transitions between cameras ($N = 27$), for example:

> "You see this short fade between perspectives [...] and that instantly distracted me for half a second and I thought: Oh cool where is he looking right now? And then it was like ... ok what just happened [in the discussion]?"

Frequently, participants described the animated transitions not as head rotation but as fading, vanishing, switching or flickering ($N = 41$) , and others perceived transitions as too slow or lagging ($N = 8$).

> "I just don't like the way – I know it is still in development – but I just don't like the way we are turning into shadows when we are turning. It looks kind of weird (two group members agree)."

> "Suddenly they are looking at you and then suddenly [the] other person ... but you don't see the transition from one camera to another [all group members express their agreement]." "Exactly - you are lagging somehow."

Similar to the quote above some participants perceived transitions as too slow or lagging ($N = 8$). Some complained that not every gaze switch resulted in a transition, but that was a deliberate design choice made to reduce the number of transitions, as they are visually very salient.

> "[The transitions] are slower than my gaze. I frequently gaze from one head to another and the animation takes longer than my switching. Not sure if one can see that? [switches gaze to demonstrate the delay]."

With one group that suggested the dwell and animation time was too slow, we tried a Gazing Heads configuration where head rotations occurred more promptly and frequently. After four minutes of testing the group agreed it was "more natural" and preferable. This raises an open question about the optimal frequency for transmitting gaze among users. While transmitting every gaze would be overwhelming, as participants shift focus about every second, our chosen dwell and recovery times might have been too conservative.

6.7.9 Realism and Nonverbal Cues

Another prevalent interview topic was the design choice to show only heads, without clothing or shoulders ($N = 52$). Participants noticed that this strengthened the 3D illusion ($N = 13$). It also made the experience somewhat artificial (similar to a game) and gave the impression of a "work in progress" ($N = 25$). The most common concern was the absence of non-verbal communication from posture, shoulder shrugging and hand gestures ($N = 46$).

> "At some point during the game I shrugged my shoulders and I thought – ok – do they even see that? Then I thought ... oh no ... now I have to say something like "I don't care" or something like that."

At the same time it was noticed that reducing non-verbal communication to just the head increased the saliency of facial expressions ($N = 15$).

> "I think it's not that bad to have floating heads, because that way you focus on the facial expressions much more."

There was a variety of opinions about showing the upper body and clothing. Arguments against showing gestures or posture were rare, but not all participants wanted the actual appearance of their upper body to be visible since, amongst other things, it meant having to dress appropriately for a meeting:

> "Think about fat-shaming for example, or women who are often exposed to unpleasant gazes."

We tried 13 interview groups using the system without green turtlenecks, adding shoulders back into view, at the expense of less natural camera switches and exaggerated upper body movement. They were unsure whether this experience was more ($N = 27$) or less natural ($N = 30$) – some described this configuration as a "hybrid", that means, easier to adapt to, given stronger similarity to Tiled View, yet providing gaze awareness benefits ($N = 12$). An open question is the mapping of upper body movement which needs to be consistent with head rotation ("inverse kinematics") while still appearing natural.

6.7.10 Implementation on Commodity Hardware

Overall, our interviews confirmed that technical maturity and realism are the main obstacles to a full implementation of Gazing Heads. Even under sub-optimal lighting conditions, with users wearing glasses or having long hair, systems need to realistically convey users' gaze, unique facial expressions and facial features. Current solutions may degrade the experience to a point where standard single-view video conferencing is preferred over unrealistic gaze-aware solutions. There are several implications for future research. Computer vision methods which modify users gaze, *e.g.* by using Generative Adversarial Networks or creating avatars) need to tackle two issues: realism [18, 17] and latency [22, 69, 256]. For example, the head rotations shown by He et al. [89] appear fairly realistic but incidentally affect facial expression[1]. Several recent works [80, 180, 192] have claimed higher levels of realism since our study was performed in 2022, and it remains to be seen how effective they are in a system like Gazing Heads. There is a need of an objective evaluation of available methods under realistic conditions instead of clean benchmark data sets or cherry picked examples. Latency needs to be held down to an acceptable level (cf. Gazing Heads: 133.33 ± 33.33 ms) [69]. Gazing Heads also needs webcam eye-tracking with $4.86°$ of accuracy (Appendix E.6.) This is achievable with modern methods under ideal conditions [90] but accuracy under realistic system conditions remains to be confirmed.

[1] `https://youtu.be/dGY8NbG11Ng`

6.7.11 Finding More Sensitive Measures

We saw earlier that the nature of the task had a dominant influence on quantitative behavioural measures, consistent with prior work [225, 10, 117]. One implication is that quantitative results on gaze awareness may not generalise well across different studies with varying task scenarios and interface layouts. It also raises the question of exactly what alternative measures could be used. The measures of turn-taking [201] seem sufficient only to detect gross changes between tasks or comparing a physical to a virtual setting [201]. One possible interpretation is that artificially introduced gaze leaves video-mediated communication behaviour unchanged. Alternatively they may be the wrong measures [229]: "how many turns people take, how long those turns are, how many pauses people take [..] doesn't reflect people's real experiences of what those conversations are like." Compared to other studies, our wider range of measures did capture some differences in users' experiences. It remains for future work to find better quantitative measures to avoid relying on self-reported data. Since it is known that "People generally get along better and communicate more effectively when they look at each other" [117], one potential approach in the game setting may be to adapt measures of user cooperation and responsibility from social psychology and behavioural game theory.

6.7.12 Limitations

We are aware of four areas in which our experiments have limitations.

Realistic Setting The tasks (game and discussion) may not entirely represent typical video conferencing sessions in domestic or business settings. We chose unusually dynamic tasks. However this was done to help participants engage and get socially comfortable as quickly as possible, given that they generally did not know each other beforehand.

Questionnaires Neglected Differences Between Tasks Questionnaires asked about differences between the two systems, but did not explore differences between the game and the discussion. This was to avoid overburdening participants with questions and consuming more time, but admittedly may have caused some relevant and interesting effects to be missed. However, no differences in experience between tasks were mentioned during interviews.

Novelty Effect The novelty of the systems, especially Gazing Heads, may have influenced participants' behaviour and ratings. During interviews, many participants ($N = 50$)

emphasised that Gazing Heads was novel, though they clearly understood that Tiled View represented familiar video-conferencing applications.

Participant Demographics Most participants were young students (Section 6.5.1), accustomed to video conferencing, so results may not generalise well to older populations less familiar with technology. We conducted experiments in three languages for variety but the speech analysis results may not entirely generalise across languages. Lastly, gaze is believed to be influenced by socio-cultural norms [84, 249], and our population was predominantly Northern European and Western Caucasian, so that is another potential limitation to generality.

6.8 Conclusion

The Tiled View layout for video-conferencing is well-established, and our user study and prior work [201, 89] have all shown how challenging it is to improve upon it. There are two obstacles to realising that improvement: convincing realism [18] and low latency [22, 69, 30] — without which the user experience is severely compromised. In this study we have addressed the realism issue.

We have found that Gazing Heads represents a clear advance over present day video-conferencing in its capacity for conveying gaze and attention. In contrast to earlier studies by Sellen [201] and by He et al. [89], which also took Tiled View as baseline, Gazing Heads has been found to increase social presence, mutual eye contact, and user engagement. It unequivocally enhances the experience of users. We attribute these results to the amplifying effect of head rotations for conveying gaze. In its current design, Gazing Heads enhances highly interactive small group meetings. For other communication scenarios, like collaborative content editing or presentations with a large audience, alternative designs may be beneficial — an idea that is somewhat supported by our interviews. For meetings however, conveying attention to content, and to other participants, and reducing distractions for the presenter, seem to be particularly important.

Human communication in virtual space is a topic of considerable and growing significance because remote working has become a permanent and prominent feature of working life, but Zoom fatigue is a challenge. Any technical progress that may mitigate it could substantially impact the effectiveness, health and well-being of users. We believe that this study, and the Gazing Heads concept in particular, could represent an important step towards that goal.

Appendix A

Additional Publications During PhD Candidacy

During my time as a PhD candidate, before transitioning to the University of Heidelberg, I made several other contributions, which are not included in the main text. To fully explain my academic contributions as a PhD candidate, this appendix chapter briefly describes these works and includes a copy of each publication.

A.1 Minimalistic Explanations: Capturing the Essence of Decisions

When commencing research on the evaluation of explanation methods, I conducted a small study investigating whether the regions selected by a feature attribution explanation (LIME [184]) reduce the image to the regions users find to be semantically meaningful. This early work is not included in the main text due to its exploratory character. I include it here as it influenced my subsequent study designs and research questions. In particular, it led to the insight that highlighting pixels is very limiting for explaining a classification decision.

Publication This work was published as a Late-Breaking work in the Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems.

> Martin Schuessler and Philipp Weiß. 2019. Minimalistic Explanations: Capturing the Essence of Decisions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI). ACM, LBW2810:1–LBW2810:6. doi: 10.1145/3290607.3312823

Minimalistic Explanations: Capturing the Essence of Decisions

Martin Schuessler
Technische Universität Berlin
Weizenbaum Institute
Berlin, Germany
schuessler@tu-berlin.de

Philipp Weiß
Technische Universität Berlin
Weizenbaum Institute
Berlin, Germany
philippweiss@mailbox.tu-berlin.de

ABSTRACT

The use of complex machine learning models can make systems opaque to users. Machine learning research proposes the use of post-hoc explanations. However, it is unclear if they give users insights into otherwise uninterpretable models. One minimalistic way of explaining image classifications by a deep neural network is to show only the areas that were decisive for the assignment of a label. In a pilot study, 20 participants looked at 14 of such explanations generated either by a human or the LIME algorithm. For explanations of correct decisions, they identified the explained object with significantly higher accuracy (75.64 % vs. 18.52 %). We argue that this shows that explanations can be very minimalistic while retaining the essence of a decision, but the decision-making contexts that can be conveyed in this manner is limited. Finally, we found that explanations are unique to the explainer and human-generated explanations were assigned 79 % higher trust ratings. As a starting point for further studies, this work shares our first insights into quality criteria of post-hoc explanations.

KEYWORDS

explanations; interpretable machine learning; image classification; deep neural networks

CHI'19 Extended Abstracts, May 4–9, 2019, Glasgow, Scotland UK

ACM ISBN 978-1-4503-5971-9/19/05.
https://doi.org/10.1145/3290607.3312823

ACM Reference Format:
Martin Schuessler and Philipp Weiß. 2019. Minimalistic Explanations: Capturing the Essence of Decisions. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts (CHI'19 Extended Abstracts), May 4–9, 2019, Glasgow, Scotland UK*. ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3290607.3312823

INTRODUCTION

The impact of machine learning on our society is growing as it is becoming an integral part of many computer programs. Unfortunately, systems like deep neural networks that have significantly promoted the revival of machine learning research are inherently uninterpretable due to their sub-symbolic nature. Hence researchers are faced with a fundamental technical barrier to transparency as they have limited understanding of what these systems are learning and are unable to prove that they will work on unseen problems [8]. Nevertheless, transparency and explainability are an integral component of ethically aligned design [5, 14]. Consequently, interpretable machine learning research has seen a surge in interest and publications with two main streams of research: The first suggest new "simpler" models that are mathematically more interpretable yet exhibit comparable performance to uninterpretable models. The second seeks to explain black-box model predictions with post-hoc explanations without uncovering the mechanism behind them [8]. The running hypothesis that motivates such research is that displaying explanations can help novice and expert users to develop trust into a model [11]. However, there is minimal consensus on a definition for interpretability [6, 8] and scholars have argued that research in this field needs to build more strongly on research on explanation in philosophy, psychology and cognitive science [9]. Furthermore, human factors and real-world usability aspects are often neglected when new approaches are proposed, which may be because current interpretable machine learning research is relatively isolated from HCI research [2].

However, interaction with intelligent systems and agents is a traditional field of HCI. For example, Kulesza et al. [7] introduced *Explanatory Debugging Systems* that explain their decisions and incorporate user feedback, which was shown to lead to better predictions, sounder mental models and higher user satisfaction. Since their implementation has been limited to simple Naïve Bayes classifiers, these principles and findings may not translate to complex deep learning models. More recent work from our community includes work by Binns et al. [3] studying how different presentation styles of explanation influence justice perception or work by Rader et al. [10] studying how explanations of the Facebook news feed algorithm influence the beliefs and judgments.

In this work, we add to this body of research by investigating if minimalistic post-hoc explanations can capture the essence of a decision and if they align with human intuition.

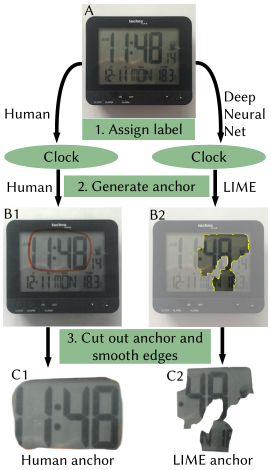


Figure 1: Anchor generation scheme: On the left branch a human assigns a label and highlights the anchor. On the right branch, a deep neural network assigns the label, and the LIME [11] algorithm creates the anchor. Both anchors are printed on paper and cut out by hand to smooth the edges.

METHOD

A "full" explanation of a complex model is often not feasible or even understandable for humans, which is why explanations need to be selective in the causes they present [9]. For the machine learning task of image classification where an image is assigned one of several possible labels, **anchors** are one possible way of providing such minimalistic explanations. An anchor is the reduction of the input image to the regions that supported the assignment of a label. In our pilot study, we compared algorithmically generated anchors to the gold standard of human explanations. For this purpose, we photographed several everyday objects and generated anchors for them algorithmically and manually.

Algorithmically Generated Anchors

To generate anchors algorithmically we used the Keras framework [4] with tensorflow [1]. We predicted a label for each photo using the *Inception v3 model* [13] trained with the 1000 class ImageNet training data (Figure 1 - Step 1). For the post-hoc explanation method, we restricted our experiment to *local interpretable model-agnostic explanations*, generated with the **LIME algorithm**. This algorithm was developed by Ribeiro et al. [11] in 2016. In a user study, they also demonstrated its ability to support users in identifying generalisation error and skewed datasets.

For a decision, LIME creates a sparse, linear model g with super-pixels as input. The resulting model is interpretable for two reasons: Firstly, the domain of g is a super-pixel representation of the image, which is meaningful for a human. Secondly, the sparsity constraint enforces that just a few of all super-pixels contribute to the classification by g, creating a very selective model. The anchor is obtained by reducing the input image to pixels that supported the decision (Figure 1 - Section B2). Anchors generated in this fashion can exhibit some rough edges which we smoothed manually. It is important to note here that different model architectures (e.g., vgg16) produce different anchors and how the architecture influences the anchors is an open research question.

Manually Generated Anchors

We showed photos of seven everyday objects to four volunteers recruited within our institute and asked them to assign a label to the image (Figure 1 - Step 1). Next, we instructed them to mark up regions of the image that they considered most relevant for their decision (Figure 1 - Step 2). If in doubt explainers were instructed to consider what regions they considered essential in such a way that their removal would make it much harder to identify the object. Finally, their selections were cut out from paper and glued back to paper smoothing the edges if necessary. Once we had created a couple of anchors in this fashion, they appeared to be considerably different from the algorithmically generated ones.

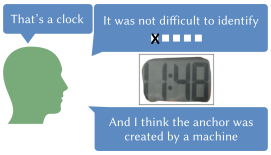


Figure 2: First stage of the experiment: For each image one of the two anchors are shown to subjects. They decide what the original label was, how difficult it is to recognise the label and finally how the anchor was created.

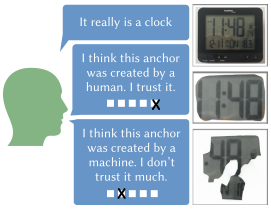


Figure 3: The second stage of the experiment: The subjects see both anchors and the original image. Again they decide which anchor was created by the algorithm. They also judge if they would trust the classifier given each explanation and assuming the machine created it.

Study Design

If anchors are selective in a human-understandable way, they should reduce an image to the essential parts. If this is the case, humans should be able to identify the object for which an anchor was generated if the anchor was generated for the correct object label. We hosted a pilot study with twenty participants, researchers from multiple disciplines, at the Weizenbaum Institute. In the first half participants were individually presented with seven anchors of the seven objects, randomly either algorithmically or manually created. In a questionnaire, they were asked to identify the object outlined by the anchor, give a difficulty rating for this task (five-point Likert scale) and select whether they think the anchor was generated by a human or by an algorithm (Figure 2). In the second part, we showed participants the original images of the object along with the anchors they had already seen and the ones they had not seen. Hence a manually and an algorithmically generated anchor were on display for each object. We also marked the anchors that explained a wrong label. In the questionnaire, we asked participants once again to determine for each anchor if a human or an algorithm generated it. Lastly, assuming the anchor had been generated by an algorithm they were asked to rate the likelihood that they trusted the underlying classifier to classify objects of the same type correctly in the future (Figure 3).

RESULTS

Fifteen out of twenty participants submitted their questionnaire which was optional. We analysed the data using two-way repeated measurement ANOVAs and report only significant results in this short work. As shown in Figure 4 the recognition rate was significantly lower for explanations that explained the wrong label (18.52% vs. 75.64%; $F_{(1,105)} = 40.14, p < 0.001$). Similarly, the difficulty rate was significantly higher ($M = 4.70, SD = 0.53$ vs. $M = 2.66, SD = 1.59$; $F_{(1,99)} = 43.0754, p < 0.001$). In the first part of the experiment participants were able to distinguish between algorithmically and manually generated anchors with an average accuracy of 57.45 % which increased to 82.52 % in the second part where anchors where displayed pairwise along with the original image. If an anchor explained an incorrect label, trust ratings were significantly lower as when it explained the correct label ($M = 2.17, SD = 1.05$ vs. $M = 3.89, SD = 1.09$; $F_{(1,205)} = 82.45, p < 0.001$) and participants trusted manually generated explanations significantly more than algorithmically generated ones ($M = 3.83, SD = 1.25$ vs. $M = 2.99, SD = 1.29$; $F_{(1,205)} = 6.90, p = 0.009$).

DISCUSSION

In our pilot study participants were able to identify the original object more accurately and with more ease when an anchor explained the right label. Hence, in most cases, **anchors seemed to reduce images to their essential parts for a given label while being very selective**. Nevertheless, an

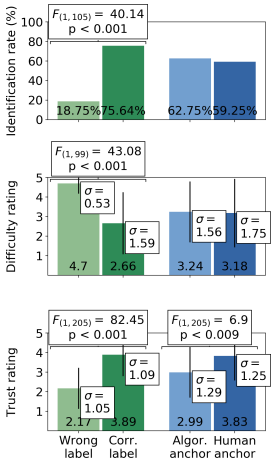


Figure 4: Study results. The three graphs compare different metrics for anchors of correctly and incorrectly labeled images (left), as well as anchors generated by humans and algorithms (right). Top: Identification rate of the correct object label. Middle: Difficulty rating of identifying the object. Bottom: Trust in the classifier's decision.

identification rate of 75.64 % is still leaving room for improvement. In future studies, we plan to allow participants to reveal additional regions interactively, which could identify important regions that had been left out by the explainer. Such feedback data could be used to improve or debug the classifier.

We also found that **explanations were unique to the explainer** (human subject or machine learning model respectively) and therefore considerably different from one another (i.e., anchors C1 and C2 in Figure 1). Hence it was easy for participants to distinguish between them once they were displayed side by side. Some participants mentioned that they saw a pattern in how they differed, stating that humans are more focused on the objects overall shape and the co-occurrence of region whereas the algorithm focussed on object-specific patterns in sub-regions. They also trusted the manually created anchors significantly more (3.89 vs. 2.17). Whether this is due to a general tendency to trust humans more is left to be investigated. Interestingly participants mentioned that they did not expect explanations to overlap or to be similar, but they expected them to align with their intuition. This shows that there can be more than one reasonable explanation for a given decision.

When creating anchors manually, participants often circled different regions that were overlapping or connected stating that the occurrence of both regions together or in a particular spatial arrangement is what made them assign a specific label (see Figure 5). However, mapping such an explanation to a set of sub-regions is not possible. Hence, **anchors can only communicate very few reasons for a given decision**. Future research could consult expertise from cognitive psychology and social science [9] about how humans generate and look at explanations. Such insights can be used to extend LIME or other post-hoc methods to convey more decision making context such as the relationships between regions. It is important to mention here that many interpretable models such as rule-based systems or classification trees provide explanations for the combination of features to a decision. Furthermore, explanations are not limited to the use of input features. Their expressiveness can be enhanced with the use of other media and modalities (see [8] for examples). Sevastjanova et al. [12] even outlined a very promising design space for the combination of *verbalisation* and *visualisation* to produce even richer explanations.

FUTURE WORK AND CONCLUSION

We aim to repeat this study with a more thorough design (no convenience sampling, better isolation of factors, improved shape of anchors, standardised questionnaires). In this experiment, we studied a very abstract notion of trust as the faith in a models performance. Following the argumentation of Doshi-Velez et al. [6] trust should instead be evaluated in respect to some real-world desiderata and more carefully operationalised. For example, one could base the reward for the experiment on the participant's ability to rely on the system appropriately. In such an experiment post-hoc explanations could be compared to real explanations, placebo explanation or simple model performance statistics. In future studies, we also seek to asses another quality indicator of explanations: their *decision-contrasting*

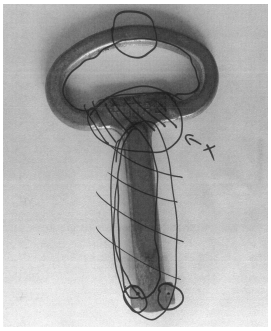


Figure 5: Participants highlights used to explain why s/he saw a key in this image of a bottle opener. Several circles cover almost the entire object because their arrangement as a whole was considered significant. The hatched area indicates that this region was of lesser importance.

Acknowledgements: Funded by the German Federal Ministry of Education and Research (BMBF) - NR 16DII113. During her fellowship at the Weizenbaum Institute, Stefania Druga provided helpful comments on this work. Berit Wiegmann helped to refine the infographics. We are also grateful to the anonymous reviewers for their valuable suggestions to mature the ideas presented in this paper.

capabilities [8, 9]. Since anchors only provide information about why a label was assigned, we plan to investigated if they can also provide useful information about why another label was not chosen.

In this work, we found that anchors are very minimalistic explanations that can be very selective. Even though they retain the essence of a decision, it is worth investigating how they could convey more decision-making contexts. We see this early work as a starting point for a series of human grounded evaluations [6] that asses the practical interpretability provided by post-hoc explanations and interpretable models.

REFERENCES

[1] Martín Abadi et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Article 582, 18 pages. https://doi.org/10.1145/3173574.3174156

[3] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Article 377, 14 pages. https://doi.org/10.1145/3173574.3173951

[4] François Chollet et al. 2015. Keras. https://keras.io.

[5] ACM U.S. Public Policy Council. 2017. Statement on Algorithmic Transparency and Accountability.

[6] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv:1702.08608 [cs, stat]* (Feb. 2017). arXiv:cs, stat/1702.08608 http://arxiv.org/abs/1702.08608

[7] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces (IUI '15)*. ACM, New York, NY, USA, 126–137. https://doi.org/10.1145/2678025.2701399

[8] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Commun. ACM* 61, 10 (Sept. 2018), 36–43. https://doi.org/10.1145/3233231

[9] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence* 267 (2019), 1–38. https://doi.org/10.1016/j.artint.2018.07.007

[10] Emilee Rader, Kelley Cotter, and Janghee Cho. 2018. Explanations As Mechanisms for Supporting Algorithmic Transparency. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. ACM, Article 103, 13 pages. https://doi.org/10.1145/3173574.3173677

[11] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '16)*. ACM, New York, NY, USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[12] Rita Sevastjanova, Fabian Beck, Basil Ell, Cagatay Turkay, Rafael Henkin, Miriam Butt, Daniel Keim, and Mennatallah El-Assady. 2018. Going beyond Visualization: Verbalization as Complementary Medium to Explain Machine Learning Models. In *Workshop on Visualization for AI Explainability*.

[13] Christian Szegedy et al. 2015. Rethinking the Inception Architecture for Computer Vision. *CoRR* abs/1512.00567 (2015). arXiv:1512.00567 http://arxiv.org/abs/1512.00567

[14] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-Being with Autonomous and Intelligent Systems. Version 2.

A.2 Power Dynamics in Data Annotation for Computer Vision

Along with the research presented in this work, the author contributed to several interdisciplinary research projects at the Weizenbaum Institute. Milagors Miceli, Tianling Yang, and I critically reflected on how ground-truth data is created for training computer vision models. This work has a strong connection to explaining image classifiers. Instead of focusing on black-box models, it focused on black-box datasets. Before image classification models can be trained, they require annotated data. When annotators assign meaning to raw data through the use of labels, they create an abstraction of reality that is later learned by computer vision models trained on this annotated data. Less attention is paid in the machine learning community to the fact that these models learn an arbitrary abstraction of reality that was created during the annotation process. We investigated image data annotation practices performed in industrial contexts to gain a deeper understanding of this sense-making process. Following a constructivist grounded theory approach, we conducted several weeks of fieldwork at two annotation companies. We analysed which structures, power relations, and naturalised impositions shape annotators' sense-making of data. We found that annotators are influenced by the interests, values, and priorities of actors above their station. Arbitrary classifications were vertically imposed on annotators and, through them, on data. This is in stark contrast to the prevailing notion that views data annotation as a neutral and objective practice that simply creates ground-truth labels. Instead, it is an exercise of power with multiple implications for individuals and society.

Publication This work was published in the Proceedings of the ACM on Human-Computer Interaction, won a Best Paper Award, and has been cited 134 times in the four years since its publication. It was presented at the 23rd ACM Conference on Computer-Supported Cooperative Work and Social Computing (CSCW):

> Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, 4, CSCW2. doi: 10.1145/3415186

Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision

MILAGROS MICELI, Technische Universität Berlin, Weizenbaum Institut, Germany
MARTIN SCHUESSLER, Technische Universität Berlin, Weizenbaum Institut, Germany
TIANLING YANG, Technische Universität Berlin, Weizenbaum Institut, Germany

The interpretation of data is fundamental to machine learning. This paper investigates practices of image data annotation as performed in industrial contexts. We define data annotation as a sense-making practice, where annotators assign meaning to data through the use of labels. Previous human-centered investigations have largely focused on annotators' subjectivity as a major cause of biased labels. We propose a wider view on this issue: guided by constructivist grounded theory, we conducted several weeks of fieldwork at two annotation companies. We analyzed which structures, power relations, and naturalized impositions shape the interpretation of data. Our results show that the work of annotators is profoundly informed by the interests, values, and priorities of other actors above their station. Arbitrary classifications are vertically imposed on annotators, and through them, on data. This imposition is largely naturalized. Assigning meaning to data is often presented as a technical matter. This paper shows it is, in fact, an exercise of power with multiple implications for individuals and society.

CCS Concepts: • **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Social and professional topics** → **Employment issues**; • **Computing methodologies** → **Supervised learning by classification**.

Additional Key Words and Phrases: Machine Learning, Computer Vision, Data Annotation, Image Data, Power, Social Inequity, Grounded Theory, Symbolic Power, Classification, Subjectivity, Data Creation, Work Place Ethnography, Training and Evaluation Data, Image Labeling

ACM Reference Format:
Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 1, 1, Article 115 (October 2020), 25 pages. https://doi.org/10.1145/3415186

1 INTRODUCTION

Power imbalances related to practices of classification have long been a topic of interest for the social sciences [9, 11, 13, 16, 32, 57]. What is (relatively) new is that arbitrary classifications are increasingly established and stabilized through automated algorithmic systems[57, 62]. With each system's outcome, meaning is imposed, and higher or lower social positions, chances, and disadvantages are assigned [6, 24, 37, 63]. These systems are often expected to minimize human intervention in decision-making and thus be neutral and value-free [23, 51, 73]. However, previous research has shown that they may contain biases that lead to discrimination and exclusion in several domains such as credit [37], the job market [70], facial recognition systems [19, 45, 71], algorithmic

Authors' addresses: Milagros Miceli, Technische Universität Berlin, Weizenbaum Institut, Berlin, Germany, m.miceli@tu-berlin.de; Martin Schuessler, Technische Universität Berlin, Weizenbaum Institut, Berlin, Germany, schuessler@tu-berlin.de; Tianling Yang, Technische Universität Berlin, Weizenbaum Institut, Berlin, Germany, tiangling.yang@tu-berlin.de.

2573-0142/2020/10-ART115
https://doi.org/10.1145/3415186

filtering [4, 62], and even advertisement [1]. Critical academic work has furthermore discussed the politics involved in data-driven systems [27, 30, 56] and highlighted the importance of investigating the capitalistic logics woven into them [20, 26, 81]. What the enthusiasm of technologists seems to render invisible is that algorithmic systems are crafted by humans and hence laden with subjective judgments, values, and interests [31, 44]. Moreover, before the smartest system is able to make predictions, humans first need to make sense of the data that feeds it [61, 66, 72]. Despite its highly interpretative character, data-related work is still often believed to be neutral, "comprising unambiguous data, and proceeding through regularized steps of analysis" [61].

The present paper investigates data annotation for computer vision based on three research questions: How do data annotators make sense of data? What conditions, structures, and standards shape that sense-making praxis? Who, and at what stages of the annotation process, decides which classifications best define each data point? We present a constructivist grounded theory [21, 59, 60] investigation comprising several weeks of fieldwork at two annotation companies and 24 interviews with annotators, management, and computer vision practitioners. We define data annotation as a sense-making [52] process where actors classify data by assigning meaning to its content through the use of labels. As we have observed, this process involves several actors and iterations and begins as clients transform their needs and expectations into annotation instructions. The sensemaking of data, so we argue, does not happen in a vacuum and cannot be analyzed independently from the context in which it is carried out.

We use Bourdieu's [13] concept of symbolic power, defined as the authority to impose meanings that will appear as legitimate and part of a natural order of things, as a lens to analyze the dynamics of imposition and naturalization inscribed in the classification, sorting, and labeling of data. Previous research in the field of data annotation has largely focused on workers' individual subjectivities as a major cause for biased labels [18, 40, 48, 77]. Conversely, our investigation introduces a power-oriented perspective and shows that hierarchical structures broadly inform the interpretation of data. Top-down meaning impositions that follow the demands of clients and the market shape data profoundly.

With this investigation, we seek to orient the discussion towards the interests and values embedded in the systems that potentially shape our individual life-chances [37]. Through the description of three observed annotation projects, we expose the deeply normative nature of such forms of data classification and discuss their effects on labels and datasets. Building on this perspective, we propose the incorporation of power-aware documentation in processes of data annotation as a method to restore context. We argue that reflexive practices can improve deliberative accountability, compliance to regulations, and the explication and preservation of effective data work knowledge. With this work, we also hope to inspire researchers to adopt a situated and power-aware perspective not only to investigate practices of data creation but also as a tool for reflecting power dynamics in their own research process.

2 RELATED WORK

2.1 Data Work as Human Activity

Previous work has argued that data-driven systems are often linked to "a technologically-inflected promise of mechanical neutrality" [41]. However, these systems require, in many cases, the intervention of human discretion in their deployment [2, 64], and even more frequently, in their creation [18, 22, 35, 36, 39, 48, 51, 61, 66, 67]. Moreover, critical research has argued that data-driven systems embody the personal and corporate values and interests of the people and organizations involved in their development [31, 51, 53, 57]. As Klinger and Svensson state, "arguments that technology had agency on its own hide the individuals, structures, and relations in power and

thus serve their interests, interests that become increasingly blurred" [53]. A view into the power dynamics encoded in data and systems is, as we will argue, of fundamental importance, especially considering that "the technical nature of the procedures tends to mask the presumptions that enter into the programming process, the choices that are made, and the conceivable alternatives that are ruled out" [57].

Besides technical exercise and operation, the development of data-driven products involves "mastering forms of discretion" [66] and is conditioned by the networked systems in which they are created, developed, and deployed [51, 73]. Kitchin [51] pinpoints various processes and factors that reveal extensive human interventions in data-driven systems, such as the translation of tasks into algorithmic models, available resources, the choice of training data, hardware and platform, the creative process of programming, and adaptation of systems to meet requirements of standards and regulations. He further argues that algorithmic systems are subject to the purposes of their creation: "to create value and capital; to nudge behaviour and structure preferences in a certain way; and to identify, sort and classify people" [51].

The examination of the provenance of data and the work practices involved in their creation is fundamental for the investigation of subjectivities and assumptions embedded in algorithmic systems. Passi and Jackson [66] propose the concept of *data vision* to describe the ability to successfully work with data through an effective combination of formal knowledge and tools, and situated decisions in the face of empirical contingency. Mastery of this interplay is essential to data analysts, which reveals "the breadth and depth of human work" inscribed in data [66].

Embedded in such processes are not only individual subjectivities, but also narratives, preferences, and values related to larger socio-economic contexts [8, 50, 67]: "Numbers not only signify model performance or validity, but also embody specific technical ideals and business values." [67]. Data practices such as the choice of training data, data capturing measurement interfaces [68], and the selection of data attributes [61] as well as the design of data in an algorithmically recognizable, tractable, and analyzable way [35, 61], all indicate that data is created through human intervention [61]. Feinberg points to the "interpretive flexibility" and situated nature of data and considers data as a product of "interlocking design decisions" made by data designers [35]. According to Muller et al. [61], the degree of human intervention will determine how deep and fundamental subjective interpretations are inscribed in data and its analysis.

The present paper unpacks data annotation practices with a human-centered perspective. The practices we have observed and analyzed are situated in outsourcing companies that provide annotation services for commissioning clients. As previous work has argued, service is situated in local, cultural, and social contexts [50] and is co-produced and co-created in the interactions between service providers and recipients [8]. This perspective sheds light on the situated [33, 66] and collaborative [35, 67] nature of data work, as clients and annotation teams both participate in the creation of datasets. Scrutinizing data annotation with a service perspective further requires taking into consideration institutional structures and organizational routines [3, 50].

Annotation tasks are, as we will argue, mainly about sensemaking [52], i.e. framing data to make it categorizable, sortable, and interpretable. Previous work in this space has largely focused on individual preconceptions, considering annotators' subjectivities to be a major source for labeling bias [18, 40, 48, 77]. Other researchers (we among them) explore factors beyond individual subjectivities that influence workers and labels, such as loosely-defined annotation guidelines and annotation context [36], the choice of annotation styles [22], and the interference between items in the same data batch [80]. In a thorough investigation into annotation practices in academic research, Geiger et al. [39] draw attention to the background of annotators, formal definitions and qualifications, training, pre-screening for crowdwork platforms, and inter-rater reliability processes.

The authors consider these factors to be likely to influence the annotations and advocate for their documentation.

With the present paper, we join the discussion around subjectivity in data annotation. By examining the processes and contexts that shape this line of work, we argue that subjectivity can also be shaped by power structures that enable the imposition of meanings and classifications.

2.2 Data, Classification, Power

Practices related to classifying and naming constitute the core of data annotation work. As Bowker and Star [16] have most prominently argued, classifications represent subjective social and technical choices that have significant yet usually hidden or blurry ethical and political implications [79]. Classification practices are constructed and, at the same time, construct the social reality we perceive and live in [11]. Therefore, they are also culturally and historically specific [46]. Adopting a critical position to examine these practices is essential because, as Durkheim and Mauss argue, "every classification implies a hierarchical order for which neither the tangible world nor our mind gives us the model. We therefore have reason to ask where it was found" [32].

Humans collect, label, and analyze data in the usually invisible context of a plan that determines what is considered data [17, 68] and how that data is to be classified [16]. "A dataset is a worldview", as Davis [29] wonderfully puts it. Accordingly, it can never be objective nor exhaustive because, "it encompasses the worldview of the labelers, whether they labeled the data manually, unknowingly, or through a third party service like Mechanical Turk, which comes with its own demographic biases. It encompasses the worldview of the built-in taxonomies created by the organizers, which in many cases are corporations whose motives are directly incompatible with a high quality of life." [29]. Furthermore, decisions about what information to collect and how to measure and interpret data define possibilities for action by making certain aspects of the social world visible – thus measurable – while excluding other aspects [30, 68]. Data-related decisions are infrastructural decisions [16, 68] as they "exercise covert political power by bringing certain things into spreadsheets and data infrastructures, and thus into management and policy" [68]. This way, datasets are powerful technologies [16] that bring into existence what they contain, and render invisible what they exclude. As Bowker argues, "the database itself will ultimately shape the world in its image: it will be performative." [15].

The performative character of datasets, that is, the power of creating reality through inclusion and exclusion, relates to Pierre Bourdieu's theorization of *symbolic power*. Symbolic power is the authority to sort social reality by separating groups, classifying, and naming them [10, 13]. Every act of classification is an attempt to impose a specific reading of the social world over other possible interpretations [57]. Thus, symbolic power is not merely a matter of naming or describing social reality but a way of "making the world through utterance" [13]. The *power* aspect here relates to the authority to lend legitimacy to certain definitions while delegitimizing others. This authority is unevenly distributed and correlates with the possession of economic, cultural, and social capital [11].

According to Bourdieu [13], dominant worldviews find their origin in arbitrary classifications that serve to legitimize and perpetuate power asymmetries, by making seem natural what is in fact political: "Every established order tends to produce (to very different degrees and with very different means) the naturalization of its own arbitrariness" [9]. The systems of meaning created through acts of symbolic power are arbitrary because they are not deducted from any natural principle but subject to the interests and values of those in a dominant position at a given place and time in history [12]. A combination of *recognition* and *misrecognition* is necessary to guarantee the efficacy of arbitrary classifications [9]: the authority to impose classifications must be *recognized as legitimate*, for the imposition to actually be *misrecognized in its arbitrariness* and be perceived as

natural. This process of *naturalization* allows arbitrary ways of sorting the social world to become so deeply ingrained that people come to accept them as natural and indisputable. As argued by D'Ignazio and Klein [30], "once a [classification] system is in place, it becomes naturalized as 'the way things are'". Thus, the worldviews imposed through symbolic power are rendered less and less visible in their arbitrariness, until disappearing into the realms of what is considered common sense. As we will argue, the interplay between recognition of authority and naturalization of arbitrary classifications decisively shapes annotations and data.

Previous investigations have related discriminatory or exclusionary outputs of data-driven systems to symbolic power: Mau argues that "advancing digitalization and the growing importance of Big Data have led to the rapid rise of algorithms as the primary instruments of nomination power" [57]. Here, nomination refers to the authority to name and classify. The author describes the ubiquity of an algorithmic authority embedded in a wide range of procedures and increasingly participating in the reinforcement of social classifications. Crawford and Paglen [27] discuss the politics involved in training sets for image classification. The authors expose the power dynamics implicit in the interpretation of images as it constitutes "a form of politics, filled with questions about who gets to decide what images mean and what kinds of social and political work those representations perform" [27]. Even if not directly referenced to Bourdieu, Crawford and Paglen's conclusion closely relates to what the French sociologist has described as the "social magic" [13] of creating reality through naming and classifying: "There's a kind of sorcery that goes into the creation of categories. To create a category or to name things is to divide an almost infinitely complex universe into separate phenomena. To impose order onto an undifferentiated mass, to ascribe phenomena to a category—that is, to name a thing—is in turn a means of reifying the existence of that category." [27]

Investigating data as a human-influenced entity [61] informed by power asymmetries [5] means understanding both data and power relationally. Data exists as such through human intervention [61] because, as we have seen, "raw data is an oxymoron" [42]. Similarly, Bourdieu [10] offers a relational view of power as enacted in the interaction among actors as well as between actors and field. In the discussion section, we will analyze the relation between annotators, data, and corporate structures. The symbolic power construct will then offer a valuable contribution to the discussion of assumptions encoded in datasets that reflect the naturalization of practices and meanings [9, 28].

3 METHOD

This investigation was guided by three research questions:

RQ1: How do data annotators make sense of data?

RQ2: What conditions, structures, and standards shape that sense-making praxis?

RQ3: Who, and at what stages of the annotation process, decides which classifications best define each data point?

We followed a constructivist variation of grounded theory methodology (GTM) [21, 59, 60]. The central premise of constructivist grounded theory is that neither data nor theories are discovered, but are formed by the researcher's interactions with the field and its participants [76]. This method provided tools to systematically reflect on our position, subjectivity, and interpretative work during fieldwork and at the coding stage.

Data was obtained through participatory observation (with varying degrees of involvement) and qualitative interviewing (in-depth and expert interviews). Fieldwork was approached exploratorily, guided by sensitizing concepts [21]. They helped to organize the complex stimuli in the field without acting as hypotheses or preconceptions. Phases of data collection and analysis were intertwined. Observations and interviews informed one another: while ideas emerging from the observations served to identify areas of inquiry for the interviews and even possible relevant

interview partners, statements from the interviews pointed many times at interesting actors, tasks, or processes needing to be more attentively observed. Through constant comparison [43], we were able to identify differences and similarities between procedures and sites.

3.1 Data Collection

3.1.1 Participatory Observations. Part of the value of open-ended observations guided by GTM is the opportunity to see the field inductively and allowing themes to emerge from the research process and the data collected. However, once in the field, researchers must somehow organize the complex stimuli experienced so that observing becomes and remains manageable because it is certainly not possible to observe all details of all situations. At this point, sensitizing concepts come into play to orient fieldwork [21]. Sensitizing concepts in this investigation include loosely defined notions such as "impact sourcing", "subjectivity", "quality assurance", "training", and "company's structure", which provided some initial direction to guide the observation during data gathering.

Fieldwork was conducted at two data annotation companies. At both locations, the level of involvement regarding observations varied from shadowing to active participant observations [58]. At both annotation companies, fieldwork was allowed to commence after a representative of the company and the researcher on the field signed non-disclosure agreements (NDA) and respectively consented for participating in the present study. Consequently, we are restrained from disclosing or using confidential information in this paper, particularly concerning the companies' clients.

3.1.2 Qualitative Interviews. Part of the fieldwork conducted consisted of intensively interviewing annotators and management. All interview partners were allowed to choose their code names or were anonymized post-hoc to preserve the identity of related informants.

Interviews with management in additional annotation companies were framed as expert interviews. While in-depth interviews aim at studying the informant's practices and perceptions, "the purpose of the expert interview is to obtain additional unknown or reliable information, authoritative opinions, serious and professional assessments on the research topic" [54]. The sampled interview partners were considered experts because they provided unique insights into the structures and processes within their companies and the overall market (see table 1 and section 3.2, Sample, for a detailed list of informants).

3.2 Sample

Four sources of information were exhaustively explored: we started with two impact sourcing companies dedicated to data annotation located in Buenos Aires, Argentina (S1) and Sofia, Bulgaria (S2). *Impact sourcing* refers to a branch of the outsourcing industry employing workers from poor and vulnerable populations to provide information-based services at very competitive prices. We chose annotation companies with rather traditional management structures over crowdsourcing platforms where hierarchies appear not as evident We assumed that clear hierarchical structures would make it easier to trace back labeling decisions and structures to real people. We also had the preconception that tensions related to exercising power would be more prominent with workers from vulnerable populations. Field access was another reason for our choice. Impact sourcing companies responded most openly to our proposed ethnographic research.

While conducting fieldwork in S2, we decided to look closer into the translation of clients' needs into annotation tasks and quality standards. Consequently, we also interviewed management employees in three similar yet larger annotation companies (S3) and engineers with a computer vision company using annotated training sets in Berlin, Germany (S4).

3.2.1 S1: The Annotation Company in Buenos Aires. At the time of this investigation in June 2019, S1 is a medium-sized enterprise centrally located in Buenos Aires and dedicated to data-related

Table 1. Overview of Informants and Fieldwork Sites

	Interview method	Medium and Language	Code name	Role
S1: FIELDWORK (Annotation company in Buenos Aires, Argentina)	Qual. in-depth interview	Face to face; Spanish (Native)	Sole	Team leader
			Elisabeth	Annotator, reviewer
			Noah	Annotator, tech leader
			Natalia	Project manager
			Paula	Founder
			Nati	QA analyst
S2: FIELDWORK (Annotation company in Sofia, Bulgaria)	Qual. expert interview	Skype, Face-to-face; English (Proficient)	Eva	Founder
	Qual. in-depth interview	Face-to-face; English (Proficient)		
	Qual. expert interview	Face to face; English (Proficient)	Anna	Intern in charge of impact assessment
		Face to face; English (Low-intermediate)	Ali	Project manager, reviewer
	Qual. in-depth interview	Face to face; English (Low-Intermediate)	Savel	Annotator
		Face to face; English (Upper-Intermediate)	Diana	Annotator
		Face to face; English (Low-Intermediate) with occasional translation by another informant	Hiva	Annotator
		Face to face; English (Intermediate)	Mahmud	Annotator
			Mariam	Annotator
			Martin	Annotator
		Face to face; English (Advanced)	Sarah	Annotator
		Face to face; another informant translated into English (Advanced)	Muzhgan	Annotator
S3: EXPERTS (Managers in large annotation companies)	Qual. expert interview	Zoom, English (Proficient)	Jeff	General manager in annotation company in Iraq
			Gina	Program manager in annotation company in Iraq
		Zoom, English (Native)	Adam	Country manager in annotation company in Kenya
		Zoom, English (Advanced)	Robert	Director in annotation company in India
S4: PRACTITIONERS (Computer vision company in Berlin, Germany)	Qual. in-depth interview	Face to face; English (Proficient)	Ines	Project manager, data protection officer
		Face to face; English (Advanced)	Dani	Product manager
		Face to face; English (Advanced) German (Native)	Michael	Computer vision engineer
		Face to face; English (Proficient)	Dean	Research scientist, lead engineer

microwork. The company has further branches in Uruguay and Colombia. The Buenos Aires office occupies a whole floor with large common work areas. This location employs around 200 data workers, mainly young people living in very poor neighborhoods or slums in and around Buenos Aires. The companyäs employment strategy is a conscious decision as part of its impact sourcing mission. At S1, workers are divided into four teams. Each team includes a project manager and several team leaders and tech leaders. Annotators perform their tasks in-house and assume mainly two roles: *creators*, doing the actual labeling work, or *reviewers*, who confirm or correct annotations. Besides annotations for visual data, the company also conducts content moderation and software testing projects. Most of the clients are large local or regional companies, including media, oil, and technology corporations. At the time of this investigation, S1 had just started to expand to Brazil and other international markets, which resulted in the need to train their workers in Portuguese and English.

One particularity of S1 is that they provide workers with a steady part- or full-time salary and benefits. This form of employment contrasts with the widespread contractor-based model in data annotation. Even so, annotators at S1 received USD1.70 per hour, the minimum legal wage in Argentina at the time of this investigation. These salaries left workers way below the poverty line in a country that accumulated around 53% annual inflation in 2019. Low salaries are not the only downside perceived by workers: informants also complained about the fixed work shifts and the

impossibility to work remotely, as the company does not allow its workers to take laptops or any other equipment home.

The interviews at the Argentine company were conducted in Spanish, the mother tongue of both interviewer and informants. Interview transcripts were coded and interpreted without translating them by the first and third authors, native and intermediate Spanish speakers, respectively. Coding without translation was done to preserve the original meaning of the statements. The quotations in this paper were translated upon completion of the analysis.

3.2.2 S2: The Annotation Company in Sofia. S2 is a small annotation company in the center of Sofia, Bulgaria. The company occupies a relatively small office. Work at this location can be quite chaotic, with workers coming and going to receive paychecks or instructions for new projects. The company focuses on the annotation of visual data, especially image segmentation and labeling. The visual data involves various types of images, including medical residue, food, and satellite imagery. The company's clients are mostly located in Europe and North America. At the time of this investigation in July 2019, ten active projects were handled by three employees in salaried positions and a pool of around 60 freelance contractors. As an impact sourcing company, S2 is committed to fair payment and works exclusively with refugees and migrants from the Middle East. The company also favors female workers among them. Contractors mostly work remotely with their own or company-provided laptops, with flexible hours. They are paid per picture and, sometimes, per annotation. Payment varies according to the project and the level of difficulty. Most informants were satisfied with the remuneration and flexible conditions. However, many of them expressed the desire to have more stability and continuity of work and income.

All interviews at this location were conducted in English. Most annotators had low to medium English skills, which represented a significant difficulty for the conduction of interviews. For example, some informants over-simplified their statements and were often not able to provide in-depth answers. The language barrier could not have been foreseen or mitigated, as the founder, whose English skills are impeccable, had assured us a selection of interview partners with similar language skills. The misunderstanding probably originated in the fact that all proposed informants were indeed able to understand English at a level that was sufficient to perform their work. It was, however, not enough for them to easily tell their stories. The language barrier required improvisation on researchers end, including the simplification of questions and the introduction of walk-through questions [58], allowing informants to show procedures directly while reducing language requirements (see table 1 for more details).

3.2.3 S3: The Experts. In grounded theory investigations, decisions regarding theoretical saturation often happen simultaneously with the gathering of data, forcing researchers to make quick decisions on whether the collection of further or different data is necessary. While conducting fieldwork in Bulgaria, the idea emerged that expert interviews with management in other, more prominent impact sourcing companies could provide further insights about the translation of clients' needs into actual annotation tasks, standards, and quality assurance (QA). Through this form of inquiry, we additionally sought to frame some of the fieldwork observations.

Three expert interviews were conducted: *Jeff and Gina* are, respectively, general and program manager with a microwork company based in Iraq. Jeff is also in charge of training future workers on data annotation. The company had initially been founded by a worldwide organization dedicated to humanitarian aid and quickly became a for-profit impact sourcing company. Jeff and Gina were interviewed simultaneously. *Adam* is the general manager at the Kenyan branch of an impact sourcing company with many hubs for data annotation throughout Asia and Africa. *Robert* is based in India and works as a director of machine learning with one of the oldest impact sourcing

companies dedicated to data annotation. The company has many branches in different Asian countries.

The informants are identified through code names. The names of their companies remain anonymous.

3.2.4 S4: The Practitioners. The demands, rules, and processes of clients represented a recurring topic of the interviews conducted within data labeling companies. It seemed that managerial roles within labeling companies implied the ability to mediate and translate the client's requirements into factual tasks for workers.

How do such requirements originate? On whose needs are they based? To start exploring these questions, we decided to briefly investigate companies ordering and deploying labeled datasets for their machine learning products. A visit to a computer vision company based in Berlin was then arranged and carried out. Four relevant actors were interviewed in-depth at this location: a project manager, the data protection officer, the lead engineer, and a data engineer.

While this company is not a direct client of S1, S2, or S3, it does commissions and utilizes labeled images for its main product.

3.3 Data Analysis

The resulting 24 interviews were transcribed. Transcriptions were integrated with several pages of field notes and various documents such as specific instructions provided by clients with labeling requirements, metrics for quality assurance, and impact assessments. We followed the grounded theory coding system [21] for the interpretation of data: Phases of *open, axial*, and *selective coding* were systematically applied.

By the end of the *open coding* phase, a set of 28 codes had emerged. The process of *axial coding* followed. We applied a set of premises [25] to make links between categories visible. The material was then meticulously coded using the renewed set of axial categories. As part of this process, we iteratively returned to the material to look for additional evidence and to test and revise the emergent understanding. This analysis led to a core set of seven axial codes (see Table 2). Finally, for the *selective coding*, we combined several axial codes to the core phenomenon *"imposition of meaning"*. Selective coding indicates deliberate interpretive choices by the researchers. Making such choices explicit during the analysis process is fundamental in constructivist grounded theory [21].

As a final step, we connected salient codes and categories to the core phenomenon as causal conditions, context, intervening conditions, action/interactional strategies, or consequences [25] (see Figure 2 "Paradigm Model").

4 FINDINGS

The annotation of visual data consists of a set of practices aiming at interpreting the content of images and assigning labels according to that interpretation. The observed work practices involve mainly two tasks: labeling and segmenting. Segmenting, formally called semantic segmentation, refers to the separation of objects within an image, thus classifying them as belonging to different kinds. Labeling is mainly about giving a name to each of the objects that were previously classified as different from each other. Sometimes, labeling also includes the assignment of keywords and attributes. Those attributes fill the ascribed classifications with meaning by putting in words what constitutes each class.

To illustrate our findings, we describe three of the observed annotation projects, that were particularly relevant to our research questions. Several of the practices and tensions described in these cases remained consistent across projects and even companies. Finally, we report four salient observations that emerged from the collected data as part of the coding process.

Table 2. Table of core phenomenon, axial categories, open codes, and explanatory memos.

	Axial Categories	Open Codes		Memos
IMPOSITION OF MEANING	CLASSIFICATION AS POWER EXERCISE	Briefing		Information of labeling instruction to labelers. Communication of client's wishes and expectations. Communication chain from client to labelers.
		Struggle over meaning		Struggle over the meaning of things. Power struggles to name things. Also moments of subversion from labelers.
		Imposition		One-way, top-down imposition of meaning during team meetings. Imposition of client's desires and/or views in view of discrepancies.
		Team Agreement		Democratic alignment of concepts and opinions within the team. Teamwork to reach an agreement on how to name things.
		Layering		Nomination instances within annotation companies. Actors deciding over the interpretation of data at different stages of the process.
	LABELING OF DATA	Tools		Different tools to perform tasks of data annotation, where they come from and how they may represent a constraint for the work.
		Agency		Room for agency while performing labeling tasks; agency here refers to the possession of resources to achieve desired results.
		Constraints		Things that could count as a constraint for subjectivity when performing tasks of data annotation.
		Standardization		How labeling is standardized. Efforts from company or client to standardize labeling tasks.
	REFLEXIVITY ON WORK IMPACT	Visions of future		How workers imagine the future in relation to the tasks they perform.
		Tech		Visions of impact of technology/AI on society. Impact of their work on society.
	IMPACT SOURCING	Training		Training received as part of the impact sourcing model. Training that could be helpful for future jobs (languages, software, etc).
		Chance		Chances to learn, to work in the desired field. Chances related to impact sourcing companies. Opportunities offered by companies to their employees.
		Impact on lives		Impact of job on worker's lives. What this job means for them and how their lives have changed with this job.
		Closeness to management		Indicators for flat hierarchies. Accessibility to management. Possibility to talk directly and honestly to management.
		Recruting		How the interviewee was recruited to work in the company. How she/he got to work there.
		Mobility chances		Chances to grow and/or be promoted within the company.
	BIAS	Misunderstanding		When the interviewer asks about biases and the interview partner offers an answer showing they have misunderstood the question.
		Unawareness		Not knowing what the concept of bias refers to. Not being aware of biases as a hazard related to their tasks.
		Not bias related		Claim that biases are not relevant for the type of projects they handle within the company
	CAPITALISTIC LOGISTICS	Company structure	Speed	Optimization of processes, so that they are faster and the client is satisfied.
			QA	Quality Assurance Processes. Especially QA as a selling argument for clients. Control as a selling point.
			Productivity	Processes related to increasing or controlling productivity, making workers produce more.
			Flexibility	Flexibility in working time, work place; not as a fixed/ regular employee; work with children etc.
			Roles	The division of roles and tasks in the work/ in the company.
		Market's logics		Things that are done in a certain way to go according to the demands of the market.
		Worker's struggle		Workers asking for better conditions/benefits. Expressions of disagreement with aspects of the working conditions.
		Control		Control mechanisms. Control of results. Control of employees.
		Clients		All things clients. Communication with clients, desires of the clients, relation to clients. Client as king.
	PERSONAL SITUATION	Plans		Plans for the future at a personal level. Hopes and dreams.
		Vulnerability		Related to the vulnerable background of workers. Personal struggle/difficulties.
		Previous work experience		What workers did before becoming labelers.
		Education		Related to workers' background. Achieved academic level. Plans for further education.

4.1 Project 1: Drawing Polygons

This project, conducted by S2 in Bulgaria, consisted of analyzing, marking, and labeling pictures of vehicles for a Spanish client. The client had provided several image collections, each containing photographs of damaged car exteriors. The source of the images and the exact purpose of the dataset were unclear for the Bulgarian team. Only Eva, the founder of the annotation company, was capable of sharing some vague information about the client and the planned product:

> "I think it's a company working for insurance companies. So, they are providing insurance companies with a tool or a service I believe that's going to be in the form of an app that their users, who are using the insurance or maybe car rental companies

> and so on, can use in order to report damages. And so, these damages can be processed very quickly and identify them automatically. I think this is the final goal. I believe they are in the very early stage still. They are still trying to gather enough photos and train enough, use enough data to train their models."

Eva was in charge of client communications and the final quality control for every project at S2. Ali, an annotator who generally acted as mediator between Eva and the team, worked on the project as well. Besides regular annotation tasks, Ali was in charge of selecting the annotators for this project, briefing them with the instructions, and answering questions. For this purpose, he maintained a project-specific Slack channel. Daily, he monitored the progress made by every labeler and reviewed the annotated pictures. Despite his prominent role, Ali had no information about the planned product or the purpose of the annotations. Lack of information and general unawareness of the machine learning pipeline was very common among annotators at S2 and, to a lesser extent, at S1 in Argentina. Eva agreed with this observation and added:

> "I think that in many cases it's too difficult for a lot people to imagine what's the data they're working on for."

Besides Eva, none of the annotators we interviewed in S1 could relate the terms "machine learning" or "artificial intelligence" to their work. Ali did not inquire about further details beyond the specific instructions for the "car accidents project" because the instruction sent by clients normally provided "all we need" to complete annotation tasks:

> I: "But why does the client need all these pictures annotated like this? Do you know?"
> B: "No. But I think ... I am not sure, because I don't ask about this."

In this case, the client had sent a PDF document containing step-by-step instructions and example pictures. Moreover, the client had provided the platform where the segmentation and annotation tasks were to be performed. The platform had been specially developed for this purpose and tailored to the client's needs.

The first task for the annotators was to select the part of the vehicle that appeared damaged from a sidebar containing different classes (e.g., door, tire, hood). After that, they drew a polygon around the damaged area. The drawing was very time-consuming, and Ali seemed to pay special attention to the correct demarcation of the damaged areas. After drawing the polygon, they would classify the type of damage and its severity. Unfortunately, the company commissioning these annotations requested that no further details about the specific commands and labels are shared in this investigation as the company considers them one of their strategic advantage.

Apart from Eva and Ali, five annotators working remotely completed the project team. For the general briefing and the project kick-off, they were summoned to the office. Eva explained the client's instructions in English and showed some examples of the pictures and the procedure. Ali translated into Arabic for annotators with low English skills. Afterward, each annotator sat at one of the work stations in the office and tested the task while Ali walked around observing how annotators performed, answering questions, and continuously commenting on how easy the work was. For the duration of the project, annotators working remotely would resolve questions with Ali via Slack. Occasionally, if Ali was not satisfied with the quality of the polygons, he would summon the annotators to the office and work with them for a few hours. The same procedure was followed in cases of visible labeling inconsistencies among workers. Eva highlighted the importance of these "alignment meetings" to ensure the uniformity of the labels through the standardization of workers' subjectivities:

> "Normally, issues in data labeling do not come so much from being lazy or not doing your work that well. They come from a lack of understanding of the specific requirements for the task or maybe different interpretations because a lot of the things ... Two people can interpret differently so it's very important to share consistency and like having everyone understand the images or the data in the same way [...]. But because a lot of these tasks are not that straightforward, it's just not ... It's not just choosing A or B. It's more like okay for example I have this car, where do I track the exact scratch or deformation? What kind of a level is it? Like, it's a little bit more complicated and that's why it's better to invest in the human capability and let's say the standardization of everyone's understanding."

4.2 Project 2: Building Categories

This project was conducted at S1, the Argentine annotation company. It constituted a test for the acquisition of an important client, namely a sizable local corporation. The potential client had simultaneously outsourced the project with different annotation companies, planning to sign a contract with the best performing team.

We find this project to be particularly interesting as it constitutes an exception to the usual procedure of labeling data according to categories instructed by clients. In this case, the annotators were in charge of developing a classification system for the annotations. Concretely, the task consisted of analyzing camera footage, counting, and classifying vehicles driving in a gas station. The annotators were in charge of coming up with logical, mutually exclusive categories for the labeling.

Three annotators, a reviewer, a team leader, and a quality assurance (QA) analyst sat together to analyze the first, 60-minutes-long video. They started by counting *all* vehicles driving in the gas station. After a few minutes, some analysts lost track and claimed they did not expect "just counting" to be so complicated. To simplify the task, the team leader suggested establishing categories first, so that each annotator could focus on counting only one category. They promptly agreed on five categories, namely cars, buses, trucks, motorcycles, and vans. While counting, new categories such as pick-ups, SUVs, and semi-trucks were suggested by annotators, approved by the team leader and the QA analyst, and finally added to the list. Also, several questions arose: Can SUVs be considered cars? Do ambulances and police cars constitute categories for themselves?

Several team members expressed being worried about not knowing the client's exact expectations. "We are not really used to this kind of ambiguity" reviewer Elisabeth said. She also shared an experience from a former project, where inconsistencies between the interpretations of client and annotators had arisen, even though the client had provided clear instructions for the annotations. On that occasion, Elisabeth had been entirely sure that her interpretation was right until the client corrected her work: "and you think you're doing everything right until the client comes and says, 'No, that's all wrong!'" The client's correction had led Elisabeth to the conclusion that "I had been wrong all along. It put us [the team] back on track."

As for the "gas station project", Nati, the QA analyst, announced to the team that, despite the freedom offered by the project, they would proceed "as usual" to resolve questions and, most importantly, to assess the correctness of allocated labels. Upon request of the interviewer, reviewer Elisabeth described the usual process in detail:

> "Whenever I cannot resolve the questions annotators bring to me, I ask the leader. If the leader cannot solve them either, we ask QA. Otherwise, they ask the contact person at the client's company."

Interviewer: "So, the client has the final say?"
Elisabeth: "Yes. And the client surely has their hierarchies to discuss a solution as well."

Despite the room offered to the team by the "gas station project" to shape data according to their own judgment, the client's figure seemed to be tacitly present at all times to orient annotators' subjectivities. QA analyst Nati summarized this observation most clearly:

> "We try to guess what the client would value the most, what will interest them, trying to put ourselves in their shoes, thinking, imagining [the client] wants this or that."

In her QA analyst role, Nati also paid special attention to optimizing the time needed to annotate each video. Having one annotator counting only one category significantly reduced task completion time but raised important questions about quality control and cost optimization, as Nati pointed out:

> "How are we going to check for accuracy if only one annotator is responsible for each class and we do not have enough reviewers?"

Nati additionally mentioned that the client would not accept the costs of cross-checking results.

For Nati and the QA department, this project involved two challenges: the first was guessing what the client was expecting from the annotations and which taxonomy would best serve that expectation. The second consisted in optimizing the performance of annotators to present a competitive offer to the potential client. Indeed, the Buenos Aires-based company seemed to put much effort into developing better ways of measuring performance and output quality. In this sense, Nati acknowledged the singularity of the "gas station project" as being uncommonly ambiguous compared to the rest of their projects which generally included clear guidelines for the labels. However, she still saw a good opportunity emerging from the open character of the project:

> "This is where the QA department makes its move and says, okay, we can measure all this. We try to offer value [...] going into details to see what we can measure and offer the client something they would value because then we also participate in the 'farming' process. If we offer clients valuable QA data, they will probably buy more hours from us."

4.3 Project 3: Classifying Faces

The third project brings us back to the Bulgarian company (S2). It dealt with collections of images depicting people. All images resembled those commonly found in a mobile phone's gallery: several selfies, group pictures of what seemed to be a family, a couple, a child holding a cat. Eva, the founder of S2, explained that the dataset was intended for a facial recognition model for mobile phones. The annotations had been commissioned by a local computer vision company.

The first task for the annotators consisted of classifying the faces in the images according to a very concise set of instruction sent via email by the commissioning client:

> (1) For each photo, draw a rectangular bounding box around each face in the photo.
> (2) Annotate each such face with the following labels: Sex: male or female. Age: baby (0-2 years old), boy or girl (2-16 years old), man or woman (16-65 years old), old man or old woman (65+ years old). Ethnicity: Caucasian, Chinese, Indian, Japanese, Korean, Latino, Afroamerican.

Additionally, five freely chosen keywords were to be attached to each image.

Founder Eva was in charge of the general quality control. Apart from her, three annotators completed the team. Ali, one of the annotators, also managed the project, mostly briefing annotators, tracking the completion of the task, and revising the bounding boxes. Despite the project's sensitive

character, Eva did not have further information about the images' provenance and whether the people depicted were aware their picture would be used in a computer vision product.

Because of the highly subjective character of this project and the specificity of the classes provided, we insistently asked annotators how they were able to differentiate and assign such labels that were, at least to the eye of the researcher on the field, not at all straightforward. Ali reacted very surprised to this kind of question, almost as if he would not understand our strong interest in this topic:

> "It's not difficult, it's easy! Because all information here [shows the email with the instructions]. You have information. The woman is between 15 to 65, I think. The old woman, 65 to more. Old woman and old man."
> Interviewer: "Yeah, but that's what I'm saying, I would have had difficulties telling whether the person in the picture is over 65."
> Ali: "No, no, because you see this picture, you make the zoom, and you see the face [he zooms in and points at the area around the eyes, probably trying to show wrinkles that are hard to recognize as such]. Everything is clear!"

Furthermore, Ali stated that this project was significantly easier to manage than others, given the fact that annotators had not raised any questions or difficulties: "I think this is a project nobody asked me about," he said. Ali's remarks coincide with the claims of the other annotators involved: the classification of the people shown in the images in terms of race, age, and sex seemed straightforward to them. The annotator in charge of keywording also claimed that this task was very easy because the attributes were, in most of the cases, "pretty obvious." When asked what would be the procedure if they were unsure about what labels to assign, Eva, the founder, answered that they would immediately seek the client's opinion:

> "In this case we usually obey everything that they say because you know their interpretations is usually the one that makes sense."

Later on, Eva referred to "the mobile libraries project" as one of the most "controversial" projects in her company's portfolio. While discussing bias-related issues and how these can affect labels, she also highlighted the importance of raising moral questions around this type of projects and working in solutions for undesirable biases. However, Eva argued that her clients would probably not be interested in investing time or money in these issues. Similarly, Anna, the intern in charge of conducting an impact assessment at S2, commented on clients' general attitude towards ethical issues related to the commissioned labels:

> "I think even if they knew they should be sensitive or should be a little conscious about these things I think it works in their favor to not be. It's totally about digital ethics but I feel like it maybe from a company perspective [...] that they would prefer an outsourcing company that doesn't ask too many questions."

Anna also allocated some responsibility with the annotation companies. She commented on the difficulty of explaining sensitive categories, such as race and gender, when workers and management have different mother tongues. In S2, around 98% of the workers are refugees from the Middle East:

> "Yes, I have observed the [mobile libraries] project ... I feel a lot of it is not that the company is not aware of these things, but I think it's maybe too complicated to explain to refugees. I think some of us are lacking the vocabulary that would translate all these nuances. [...] And I've never heard any of them... any of the refugees ask... I think that's also another factor. I think it's a combination of a lot of these: The difficulties to explain it and, maybe, the lack of curiosity or explicit curiosity on their end."

4.4 Salient Observations

4.4.1 Standardization. At both annotation companies and in all projects observed, data annotation was performed following the requirements and expectations of commissioning clients. Guidelines were generally tailored to meet the requirements of the product that would be trained on the annotated datasets, its desired outcome, and its revenue plan. Instructions and briefings, while providing orientation, aimed at shaping the interpretation of data and, as described by Eva in section 4.1, "standardizing everyone's understanding." As shown in Projects 1 and 2, quality assurance constituted another decisive instance towards standardization and compliance with clients' expectations. Encouraged to define what quality means in the context of their company, informants at both locations (S1 and S2) and among the experts (S3) gave more or less different versions of a similar answer: quality means doing what the client expects.

4.4.2 Layering. As shown in project 2, many roles and departments participate in annotation assignments. Annotators occupy the lower layer of the hierarchical structure where the actual labeling of data is carried out (see Fig. 1). In a more or less official way, every company has at least two more layers where control is exercised: reviewers and quality assurance analysts (QA). In between reviewers and QA, some companies also place team leaders, tech leaders, and project managers. Finding more layers is possible, depending on the project's and company's size. As described in Project 2, large corporations sometimes outsource the labeling of the same dataset with different annotation companies. The results will later be controlled and compared. Also, important clients often hire external consultants to evaluate the performance of annotation companies independently. Furthermore, some annotation companies outsource parts of large labeling projects, if they lack the human resources to complete the task. These practices add even more layers to the annotation process. According to the experts (S3) and practitioners (S4) we interviewed, the layered character of these procedures is not exclusive of S1 and S2 but can be generalized to other annotation companies.

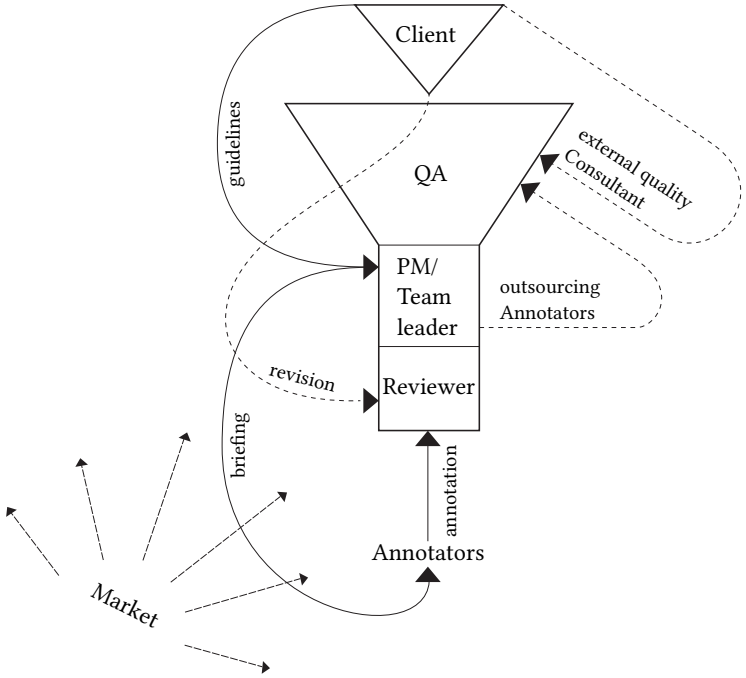


Fig. 1. Multiple actors on several layers of classification participate in processes of data annotation. The layers are hierarchical and involve different levels of payment, occupational status, and epistemic authority.

4.4.3 Naturalization. Our findings show that the top-down ascription of meanings to data through multi-layered structures were, for the most part, not perceived as an imposition by annotators. The interviews are abundant in statements such as "the labels are generally self-evident," and "the work is very straightforward."

The labels commissioned by clients and instructed by managers seemed to coincide in most cases with annotators' perceptions. In consequence, labels were hardly ever put under scrutiny or discussed. Moreover, annotators and managers generally perceived clients to be the ones to know exactly how data was supposed to be labeled since they held decisive information about the product they aimed at developing and the corresponding business plan. Additionally, in some cases, the image data to be labeled had been directly gathered by the commissioning company, which reinforced the idea that the client would know best how to interpret those images. This was reported by Eva (Founder of S2) in relation to a project involving satellite imagery. These perceptions contribute to the naturalization of the layers of classification depicted in Fig. 1. As illustrated by the projects described throughout this section, annotators broadly resolve doubts or ambiguities regarding the labels by asking their superiors. Both at S1 and S2, we found that the vertical resolution of questions prevailed over horizontal discussions and inter-rater agreement.

4.4.4 Profit-Orientation. Annotation companies mostly seek to optimize the speed and homogeneity of annotations to offer reasonable prices in the competitive market of outsourcing services. Several annotators (especially in S2) stated that project deadlines were often so short, that they were difficult to meet. Looking to cope with such a fast pace, workers relied even more on clear guidelines and efficient tools. Several informants at S1 and S2 stated that they found their work easier when clients provided clear instructions, a rather simple platform for the annotations, and a smaller number of classes to label. As shown by the "gas station project" (section 4.2), annotators tended to feel overwhelmed otherwise. In this sense, hierarchical structures did not solely aim at constraining workers' subjectivity but also provided orientation.

As expected from for-profit organizations, commissioning clients and annotation companies are primarily concerned with product and revenue plans. Moreover, as stated by Eva and Anna in section 4.3, some annotation companies may perceive a general disinterest of clients regarding the application of ethics-oriented approaches, i.e., transparent documentation and quality control for biased labels. A similar observation was reported by a QA analyst in S2 and confirmed by the four experts interviewed (S3). However, this does not mean that detrimental intentions guide clients. It merely states that ethical approaches involve monetary costs that clients cannot or will not bear. In short, several informants in S1, S2, S3, and S4 described an environment where market logics and profit-oriented priorities get inscribed in labels, even in projects involving sensitive classifications, as described in section 4.3.

5 DISCUSSION

Our observations show that annotators' subjectivities are, in most of the cases, subdued to interpretations that are hierarchically instructed to them and imposed on data. We relate this process to the concept of symbolic power, defined by Pierre Bourdieu [13] as the authority to impose arbitrary meanings that will appear as legitimate and part of a natural order of things. *Arbitrariness* is, in Bourdieu's conception [12, 14, 28], not a synonym of randomness. It refers to the discretionary character of imposed classifications and their subsumption to the interests of the powerful.

A *twofold naturalization* in the Bourdieusian sense [9] seems to facilitate the top-down imposition of meaning in data annotation: *First*, we found that classifications used to ascribe meaning to data are broadly naturalized. Annotators mostly perceive the labels instructed by clients and reassured by managers and QA as correct and self-evident. In a recent investigation, Scheuerman et al.

present a similar observation, describing how race and gender categories are generally presented as indisputable in image datasets [72]. In most of the cases observed by us, annotators, managers, and clients do not perceive assigned classifications as arbitrary or imposed. Hence, the labels are hardly ever questioned. *Second*, we have observed that the epistemic authority of managers and clients is also broadly naturalized by annotators. They are perceived to know better what labels correctly describe each data point. The higher the position occupied by an actor, the more accepted and respected their judgments. Even if annotators or management ever perceive principles of classifications as opposing personal or corporate values, the view persists that "the one who is paying" has the right to impose meaning. This way, clients have the faculty to impose their preferred classifications, just as they have the financial means to pay for labelers to execute that imposition. As illustrated by the "gas station project" in section 4.2, workers might even feel overwhelmed when clients do not overtly exercise their authority to instruct principles of classification. When annotators are challenged with making sense of the data themselves, the main rationale becomes "what would the client want?" in contrast to "what is contained in this data?". In this twofold naturalization lies, we argue, the efficacy of interpretations imposed on data: labels must be naturalized and thus perceived as self-evident if actors are to misrecognize the arbitrariness of their imposition [9].

As shown by our findings, the standardization of annotation practices and labels is assured throughout several layers of classification and control. The positions are depicted in Figure 1 as hierarchical layers positioned one above the other because they involve different levels of responsibility, payment, and occupational status. The number of layers, actors, and iterations involved hinders the identification of specific responsibilities. Moreover, no information regarding actors involved and criteria behind data-related decisions is registered. Annotation steps and iterations remain broadly undocumented. Accountability is diluted in these widespread practices. A problematic implication is that this multi-layered standardization process is hardly ever oriented towards social responsibility and usually responds to economic interests only [49]. There is no intention, however, to imply here that standardization is fundamentally harmful or that detrimental intentions lead the actors involved. We rather aim at showing how power structures can be stabilized through imposed standards [16] and argue that standardization can be dangerous if it is solely guided by profit maximization.

In this sense, we argue that the discussion on workers' subjectivity and personal values around data annotation should not let us researchers forget that datasets are generally created as part of large industrial structures, subject to market vicissitudes, and deeply intertwined with naturalized capitalistic interests. The challenge here is "to explicate the assumptions, concepts, values, and methods that today seem commonplace" [8] in this (and other) forms of service.

The main contribution of our investigation is the introduction of a power-oriented perspective to discuss the dynamics of imposition and naturalization inscribed in the classification, sorting, and labeling of data. Through this lens, we shed light on power imbalances informing annotation practices and shaping datasets at their origins. Our main argument is that power asymmetries inherent to capitalistic labor and service relationships have a fundamental effect on annotations. They are at the core of the interpretation of data and profoundly shape datasets and computer vision products.

There are at least two close-connected reasons why imposition and naturalization in the context of data creation are socially relevant and, in a way, different from power imbalances enacted through work practices in other settings: *First*, data practices involve particular ethical concerns because assumptions and values that inform data can potentially have devastating effects for individuals and communities [34, 63]. Algorithms trained on data that reproduces racists, sexist, or classist

classifications can reinforce discriminatory visions [62] "by suggesting that historically disadvantaged groups actually deserve less favorable treatment" [6]. Moreover, data about human behavior is increasingly sold for profit [81], which could result in surveillance [81] and exploitation [26]. *Second*, data-related decisions define possibilities for action, by making certain aspects of reality visible in datasets, while excluding others [15, 68]. This is relevant for state management and policy, i.e., to pinpoint places where intervention or allocation of resources is needed. However, the tendency of classification practices towards the erasure of residual categories [16] can cause tension and even be harmful for individuals who remain unseen or misclassified by data-driven systems [19, 71].

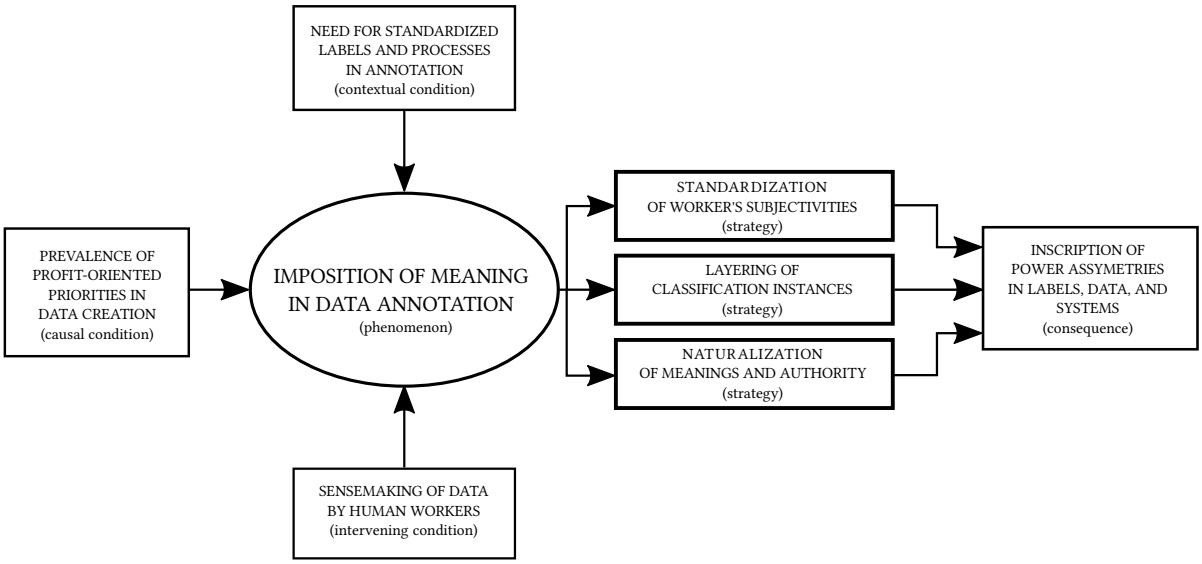


Fig. 2. Paradigm model resulting from the process of selective coding. It depicts the top-down allocation of meanings, its stabilization through annotation practices, and its effects on data (derived from the Grounded Theory Paradigm Model, by Corbin and Strauss). [25]

5.1 Implications for Practitioners

While annotation companies and their clients may or may not be aware that they are actively shaping data, the opacity surrounding embedded interests and preconceptions [72] is a significant threat to fairness, transparency, accountability, and explainability. Therefore, it is important that practitioners, i.e., corporations commissioning datasets and management at annotation companies, take steps to reflect, document, and communicate their subjective choices [38, 61, 65, 66, 72]. Promoting the intelligibility of datasets is fundamental because they play a key role in the training and evaluation of ML systems. Understanding datasets' origin, purpose, and characteristics can help better understand the behavior of models and uncover broad ethical issues [78].

Recent research work has highlighted the importance of structured disclosure documents that should accompany datasets [7, 38, 39, 47, 55, 78]. Fortunately, the machine learning research community has begun to promote similar reflexive practices: Following Pineau's suggestion [69], authors of NeurIPS and ICML conference are now requested to include a reproducibility checklist which encourages "a complete description of the data collection process, such as instructions to annotators and methods for quality control" if a new dataset is used in a paper. NeurIPS further requires authors to disclose funding and competing interests. They are also asked to discuss "the potential broader impact of their work, including its ethical aspects and future societal consequences." These

conferences are highly influential for ML practitioners and facilitate the adoption of the latest machine learning capabilities. It is certainly our hope that they will also inspire them to adopt such reasonable best practices and to engage in reflexive documentation.

In line with previous literature [7, 38, 49, 74, 78], we advocate for the documentation of purpose, composition, and intention of datasets. Moreover, the structures, decisions, actors, and frameworks which shape data annotation should be made explicit [39, 72]. We furthermore propose orienting documentation towards a reflexion of power dynamics. D'Ignazio and Klein [30] propose asking *who questions* to examine how power operates in data science. In this vein, we propose that disclosure documents include answers to questions such as: Whose product do the annotations serve, and how? Whose rationale is behind the taxonomies that were applied to data? Who resolved discrepancies in the annotation process? Who decided if labels were correctly allocated?

We argue that the annotation process already begins as clients transform their needs and expectations into annotation instructions. Therefore, the responsibility for documenting should not be solely placed with annotators but should be seen as a collaborative project involving annotation companies and commissioning clients. Given the hierarchical structures and power imbalances described in this paper, we find it extremely important that clients keep a record of the instructions that were given to annotators, the platforms on which annotations were performed, and the reasons for that platform choice, as well as the procedure employed for solving ambiguities, creating homogeneity, and establishing inter-annotator agreements. Extending dataset factsheets with a power-aware perspective could make power asymmetries visible and raise awareness about meaning impositions and naturalization. Yet, it is vital that documentation checklists are not prescriptive and produced exclusively in the vacuum of academia [38]. Instead, disclosure documents should be developed in an open and democratic exchange with annotation companies and their clients to accommodate real-world needs and scenarios [55].

Annotation companies and their clients might be reluctant to implement such a time-consuming documentation process. Moreover, they may regard some of the information as trade secrets, especially if it involves details about the intended product or if the structuring of the annotation process is considered a strategic advantage. We argue that allocating resources for documentation could nevertheless bring three pay-offs for organizations:

The first benefit is that proper documentation can foster deliberative accountability [67] and improve inter-organizational traceability, for instance, between annotation companies and clients. In addition, transparent documentation can help address the problematic dilution of accountability as a result of various actors and layers in the annotation process. In the context of this service relationship, accountability involves not only specific individuals but also organizations and includes factors such as organizational routines and processes of value co-creation [50]. Given the power imbalances that are inherent to this relationship [8], annotation companies could be motivated to keep track of decisions and procedures in the event of discrepancies with clients.

The second benefit is that documentation can facilitate compliance with regulations such as the GDPR and especially the "Right to Explanation" [67]. Serving as an external motivation, legal frameworks and regulations urge companies to put transparency as well as societal and ethical consequences of their products and services above the rationale of profit-maximization [49]. If there is no legal incentive and companies perceive transparency as coming at the cost of profit-oriented goals (as shown in our data), independently created transparency certifications and quality seals for datasets may provide an additional incentive given the momentum created around FATE AI.

The third benefit is that documentation may create a long term business asset because knowledge about practical data work is made explicit and persistent. Without documentation, such knowledge is often confined to workers with the "craftsmanship" to make situated and discretionary decisions [66], bearing the risk of knowledge loss due to worker flow or lack of traceability. At the same time,

documentation can have analytical value, improve communication in interdisciplinary teams, and ease comprehension "for people with diverse backgrounds and expertise" [61].

5.2 Implications for (CSCW) Researchers

Our research highlights the relation between human intervention and hierarchical structures in processes of data creation. It shows that power imbalances not only translate into asymmetrical labor conditions but also concretely shape labels and data. We firmly believe that researchers studying socio-technical systems in general, and data practices, in particular, could benefit from including a similar, power-aware perspective in their analysis. Such a perspective would primarily aim at making asymmetrical relations visible. Making power visible means exposing naturalized imbalances that get inscribed in datasets and systems [30].

We propose four (interconnected) reasons for integrating such a perspective into research:

First, this perspective could contribute to making work visible [30, 44, 75]. Especially in the case of machine learning systems where the enthusiasm of technologists tends to render human work invisible [44], research should emphasize the value of the human labor that makes automation possible. Furthermore, making "humans behind the machines" [53] visible could help contest any pretension of calculative neutrality attributed to automated systems.

Second, this paper argues that power relationships inscribed in datasets are as problematic as individual subjectivities. A power-oriented perspective allows researchers to "shift the gaze upwards" [5] and move beyond a simplistic view of individual behaviors and interpretations that, in many cases, could end up allocating responsibilities with workers exclusively. A view into coroporate structures and market demands can offer a broader perspective to this line of research.

Third, the investigation of organizational routines and hierarchies could help researchers approach the real-world practice of data work [67], develop context-situated recommendations, and assess their applicability in corporate scenarios. This could help establish open and democratic discussions between researchers and practitioners regarding the conception of solutions for undesired data-related issues [38, 55].

Finally, rigorous reflexion and documentation of power dynamics is not only advisable for practitioners working with data but is also fundamental for researchers investigating those work practices. Acknowledging that, just like data, theories are not discovered, but they are co-constructed by researchers and participants [76] is a significant step in this direction. Throughout this investigation, the constructivist variation of grounded theory [21] has constituted a fantastic tool to methodically reflect on the researchers' perspectives, interpretations, and position.

5.3 Limitations and Future Work

This paper has focused on the annotation of image data for machine learning as performed within impact sourcing companies. While our current results are bound to this context, the framework presented here could inspire further (comparative) research involving diverse actors in other annotation settings, such as crowdsourcing platforms.

6 CONCLUSION

This paper has presented a constructivist grounded theory investigation of the sensemaking of data as performed by data annotators. Based on several weeks of fieldwork at two companies and interviews with annotators, managers, and computer vision practitioners, we have described structures and standards that influence the classification and labeling of data. We aimed at contesting the supposed neutrality of data-driven systems by setting the spotlight on the power dynamics that inform data creation.

We found that workers' subjectivity is structurally constrained and profoundly shaped by classifications imposed by actors above annotators' station. Briefings, annotation guidelines, and quality control all aim at meeting the demands of clients and the market. We have argued that the creation of datasets follows the logics of cost effectiveness, optimization of workers' output, and standardization of labels, often at the expense of ethical considerations.

We have observed the presence of multiple instances of classification, with diverse actors among several hierarchical layers that are related to the possession of capital. We have argued that the many layers, actors, and iterations involved contribute to the imposition of meaning and, finally, to the dilution of responsibilities and accountability for the possible harms caused by arbitrary labels. Furthermore, our findings have shown that workers naturalize the imposed classifications as well as the epistemic authority of those actors higher in the hierarchy. Our observations indicate that power asymmetries, which are inherent to labor relations and to the service relationship between annotation companies and their clients, fundamentally shape labels, datasets, and systems.

We have furthermore discussed implications for practitioners and researchers and advocated for the adoption of a power-aware perspective to document actors and rationale behind the meanings assigned to data in annotation work. Finally, we have emphasized the importance of adopting a similar power-aware perspective in the CSCW research agenda, not only as a possible focus for future work but also as a tool for reflecting on researchers' own position and power.

7 ACKNOWLEDGEMENTS

Funded by the German Federal Ministry of Education and Research (BMBF) – Nr. 16DII113f. We would like to acknowledge the individuals and companies participating in this study: we dearly thank them for their openness! Special thanks to Philipp Weiß for his support whenever we struggled with formatting tables in Overleaf. We wish to thank our anonymous reviewers for their feedback, and Enrico Costanza, Walter S. Lasecki, Leon Sixt, Florian Butollo, Matti Nelimarkka, and Alex Hanna for valuable comments on earlier versions of this work. Special thanks to our research group leader Diana Serbanescu and PI Bettina Berendt for their continuous support for this project.

REFERENCES

[1] Muhammad Ali, Piotr Sapiezynski, Miranda Bogen, Aleksandra Korolova, Alan Mislove, and Aaron Rieke. 2019. Discrimination Through Optimization: How Facebook's Ad Delivery Can Lead to Biased Outcomes. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019), 199:1–199:30. https://doi.org/10.1145/3359301

[2] Ali Alkhatib and Michael Bernstein. 2019. Street-Level Algorithms: A Theory at the Gaps Between Policy and Decisions. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 530:1–530:13. https://doi.org/10.1145/3290605.3300760

[3] Luis Araujo and Martin Spring. 2006. Services, Products, and the Institutional Structure of Production. *Industrial Marketing Management* 35, 7 (Oct. 2006), 797–805. https://doi.org/10.1016/j.indmarman.2006.05.013

[4] Paul Baker and Amanda Potts. 2013. 'Why Do White People Have Thin Lips?' Google and the Perpetuation of Stereotypes via Auto-Complete Search Forms. *Critical Discourse Studies* 10, 2 (May 2013), 187–204. https://doi.org/10.1080/17405904.2012.744320

[5] Chelsea Barabas, Colin Doyle, JB Rubinovitz, and Karthik Dinakar. 2020. Studying up: Reorienting the Study of Algorithmic Fairness around Issues of Power. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, Barcelona, Spain, 167–176. https://doi.org/10.1145/3351095.3372859

[6] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. https://doi.org/10.15779/Z38BG31

[7] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[8] Jeanette Blomberg and Chuck Darrah. 2015. Toward an Anthropology of Services. *The Design Journal* 18, 2 (2015), 171–192. https://doi.org/10.2752/175630615X14212498964196

[9] Pierre Bourdieu. 1977. *Outline of a Theory of Practice.* Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511812507
[10] Pierre Bourdieu. 1985. The Social Space and the Genesis of Groups. *Theory and Society* 14, 6 (1985), 723–744. https://doi.org/10.1007/BF00174048
[11] Pierre Bourdieu. 1989. Social Space and Symbolic Power. *Sociological Theory* 7, 1 (1989), 14–25. https://doi.org/10.2307/202060
[12] Pierre Bourdieu. 1990. *The logic of practice* (reprinted ed.). Polity Press, Cambridge.
[13] Pierre Bourdieu. 1992. *Language and Symbolic Power* (new ed.). Blackwell Publishers, Cambridge.
[14] Pierre Bourdieu. 2000. *Pascalian Meditations.* Stanford University Press, Stanford, Calif.
[15] Geoffrey C. Bowker. 2000. Biodiversity Datadiversity. *Social Studies of Science* 30, 5 (Oct. 2000), 643–683. https://doi.org/10.1177/030631200030005001
[16] Geoffrey C. Bowker and Susan Leigh Star. 1999. *Sorting Things out: Classification and Its Consequences.* MIT Press, Cambridge, Mass.
[17] danah boyd and Kate Crawford. 2012. Critical Questions for Big Data: Provocations for a Cultural, Technological, and Scholarly Phenomenon. *Information, Communication & Society* 15, 5 (June 2012), 662–679. https://doi.org/10.1080/1369118X.2012.678878
[18] C. E. Brodley and M. A. Friedl. 1999. Identifying Mislabeled Training Data. *Journal of Artificial Intelligence Research* 11 (Aug. 1999), 131–167. https://doi.org/10.1613/jair.606
[19] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, Vol. 81. PMLR, 77–91.
[20] Ryan Burns. 2019. New Frontiers of Philanthro-capitalism: Digital Technologies and Humanitarianism. *Antipode* 51, 4 (April 2019), 1101–1122. https://doi.org/10.1111/anti.12534
[21] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis.* Sage Publications, London ; Thousand Oaks, Calif.
[22] Justin Cheng and Dan Cosley. 2013. How Annotation Styles Influence Content and Preferences. In *Proceedings of the 24th ACM Conference on Hypertext and Social Media - HT '13.* Association for Computing Machinery, Paris, France, 214–218. https://doi.org/10.1145/2481492.2481519
[23] Angèle Christin. 2016. From Daguerreotypes to Algorithms: Machines, Expertise, and Three Forms of Objectivity. *SIGCAS Computers and Society* 46, 1 (2016), 27–32. https://doi.org/10.1145/2908216.2908220
[24] Danielle Keats Citron and Frank Pasquale. 2014. The Scored Society: Due Process for Automated Predictions. *Washington Law Review* 89, 1 (2014).
[25] Juliet M. Corbin and Anselm L. Strauss. 2015. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory* (fourth edition ed.). SAGE, Los Angeles.
[26] Nick Couldry and Ulises A. Mejias. 2019. Data Colonialism: Rethinking Big Data's Relation to the Contemporary Subject. *Television & New Media* 20, 4 (May 2019), 336–349. https://doi.org/10.1177/1527476418796632
[27] Kate Crawford and Trevor Paglen. 2019. Excavating AI. https://www.excavating.ai.
[28] Ciaran Cronin. 1996. Bourdieu and Foucault on Power and Modernity. *Philosophy & Social Criticism* 22, 6 (Nov. 1996), 55–85. https://doi.org/10.1177/019145379602200603
[29] Hannah Davis. 2020. A Dataset Is a Worldview. https://towardsdatascience.com/a-dataset-is-a-worldview-5328216dd44d.
[30] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data Feminism.* The MIT Press, Cambridge, Massachusetts.
[31] Ravit Dotan and Smitha Milli. 2020. Value-Laden Disciplinary Shifts in Machine Learning. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20).* Association for Computing Machinery, Barcelona, Spain, 294. https://doi.org/10.1145/3351095.3373157
[32] Emile Durkheim and Marcel Mauss. 1963. *Primitive Classification.* University of Chicago Press.
[33] M. C. Elish and danah boyd. 2018. Situating Methods in the Magic of Big Data and AI. *Communication Monographs* 85, 1 (Jan. 2018), 57–80. https://doi.org/10.1080/03637751.2017.1375130
[34] Virginia Eubanks. 2018. *Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor.* St. Martin's Press, New York.
[35] Melanie Feinberg. 2017. A Design Perspective on Data. In *CHI '17: Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems (CHI '17).* Association for Computing Machinery, Denver, Colorado, USA, 2952–2963. https://doi.org/10.1145/3025453.3025837
[36] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. 2010. Annotating Named Entities in Twitter Data with Crowdsourcing. In *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk (CSLDAMT '10).* Association for Computational Linguistics, Los Angeles, California, 80–88. https://doi.org/10.5555/1866696.1866709

[37] Marion Fourcade and Kieran Healy. 2013. Classification Situations: Life-Chances in the Neoliberal Era. *Accounting, Organizations and Society* 38, 8 (Nov. 2013), 559–572. https://doi.org/10.1016/j.aos.2013.11.002

[38] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumeé III, and Kate Crawford. 2018. Datasheets for Datasets. *arXiv:1803.09010* (March 2018). arXiv:1803.09010

[39] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, Garbage out? Do Machine Learning Application Papers in Social Computing Report Where Human-Labeled Training Data Comes From?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, Barcelona, Spain, 325–336. https://doi.org/10.1145/3351095.3372862

[40] Bhavya Ghai, Q. Vera Liao, Yunfeng Zhang, and Klaus Mueller. 2020. Measuring Social Biases of Crowd Workers Using Counterfactual Queries. In *Workshop on Fair & Responsible AI at ACM CHI Conference on Human Factors in Computing Systems*. Honolulu, HI, USA.

[41] Tarleton Gillespie and Tarleton Gillespie. 2014. The Relevance of Algorithms. In *Media Technologies: Essays on Communication, Materiality, and Society*, Pablo J. Boczkowski and Kirsten A. Foot (Eds.). The MIT Press, 167–194. https://doi.org/10.7551/mitpress/9780262525374.003.0009

[42] Lisa Gitelman (Ed.). 2013. *"Raw Data" Is an Oxymoron*. The MIT Press, Cambridge, Massachusetts ; London, England.

[43] Barney G. Glaser and Anselm L. Strauss. 1998. *Grounded theory: Strategien qualitativer Forschung*. Huber, Bern.

[44] Mary L. Gray and Siddharth Suri. 2019. *Ghost Work: How to Stop Silicon Valley from Building a New Global Underclass*. Houghton Mifflin Harcourt, Boston.

[45] Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender Recognition or Gender Reductionism? The Social Implications of Embedded Gender Recognition Systems.. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18)*. Association for Computing Machinery, New York, 1–13. https://doi.org/10.1145/3173574.3173582

[46] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, Barcelona, Spain, 501–512. https://doi.org/10.1145/3351095.3372826

[47] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv:1805.03677* (2018).

[48] Christoph Hube, Besnik Fetahu, and Ujwal Gadiraju. 2019. Understanding and Mitigating Worker Biases in the Crowdsourced Collection of Subjective Judgments. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3290605.3300637

[49] Gunay Kazimzade and Milagros Miceli. 2020. Biased Priorities, Biased Outcomes: Three Recommendations for Ethics-Oriented Data Annotation Practices. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society. (AIES '20)*. Association for Computing Machinery, New York, NY, USA, 1–7. https://doi.org/10.1145/3375627.3375809

[50] Lucy Kimbell and Jeanette Blomberg. 2017. The Object of Service Design. In *Designing for Service: Key Issues and New Directions*. Bloomsbury Publishing, 81–94.

[51] Rob Kitchin. 2017. Thinking Critically about and Researching Algorithms. *Information, Communication & Society* 20, 1 (Jan. 2017), 14–29. https://doi.org/10.1080/1369118X.2016.1154087

[52] Gary Klein, Jennifer K. Phillips, Erica L. Rall, and Deborah A. Peluso. 2007. A Data-Frame Theory of Sensemaking. In *Expertise out of Context: Proceedings of the Sixth International Conference on Naturalistic Decision Making*. Lawrence Erlbaum Associates Publishers, Mahwah, NJ, US, 113–155.

[53] Ulrike Klinger and Jakob Svensson. 2018. The End of Media Logics? On Algorithms and Agency. *New Media & Society* 20, 12 (Dec. 2018), 4653–4670. https://doi.org/10.1177/1461444818779750

[54] Natalia M Libakova and Ekaterina A Sertakova. 2015. The Method of Expert Interview as an Effective Research Procedure of Studying the Indigenous Peoples of the North. *Journal of Siberian Federal University. Humanities & Social Sciences* 8, 1 (2015), 114–129. https://doi.org/10.17516/1997-1370-2015-8-1-114-129

[55] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–14. https://doi.org/10.1145/3313831.3376445

[56] Astrid Mager. 2012. Algorithmic Ideology: How Capitalist Society Shapes Search Engines. *Information, Communication & Society* 15, 5 (June 2012), 769–787. https://doi.org/10.1080/1369118X.2012.676056

[57] Steffen Mau. 2019. *The Metric Society: On the Quantification of the Social*. Polity, Cambridge ; Medford, MA.

[58] Frauke Mörike. 2019. Ethnography for Human Factors Researchers. Collecting and Interweaving Threads of HCI.

[59] Michael Muller. 2014. Curiosity, Creativity, and Surprise as Analytic Tools: Grounded Theory Method. In *Ways of Knowing in HCI*, Judith S. Olson and Wendy A. Kellogg (Eds.). Springer, New York, NY, 25–48. https://doi.org/10.1007/

978-1-4939-0378-8_2

[60] Michael Muller, Shion Guha, Eric P.S. Baumer, David Mimno, and N. Sadat Shami. 2016. Machine Learning and Grounded Theory Method: Convergence, Divergence, and Combination. In *Proceedings of the 19th International Conference on Supporting Group Work (GROUP '16)*. Association for Computing Machinery, Sanibel Island, Florida, USA, 3–8. https://doi.org/10.1145/2957276.2957280

[61] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300356

[62] Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism.* NYU Press, New York.

[63] Cathy O'Neil. 2017. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.* PENGUIN BOOKS, London.

[64] Juho Pääkkönen, Matti Nelimarkka, Jesse Haapoja, and Airi Lampinen. 2020. Bureaucracy as a Lens for Analyzing and Designing. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA., 1–14. https://doi.org/10.1145/3313831.3376780

[65] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, Atlanta, GA, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[66] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, Portland, Oregon, USA, 2436–2447. https://doi.org/10.1145/2998181.2998331

[67] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 1–28. https://doi.org/10.1145/3274405

[68] Kathleen H. Pine and Max Liboiron. 2015. The Politics of Measurement and Action. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. Association for Computing Machinery, New York, NY, USA, 3147–3156. https://doi.org/10.1145/2702123.2702298

[69] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché-Buc, Emily Fox, and Hugo Larochelle. 2020. Improving Reproducibility in Machine Learning Research (A Report from the NeurIPS 2019 Reproducibility Program). *arXiv:2003.12206* (April 2020). arXiv:2003.12206

[70] Alex Rosenblat, Tamara Kneese, and Danah Boyd. 2014. Networked Employment Discrimination. *SSRN Electronic Journal* (2014). https://doi.org/10.2139/ssrn.2543507

[71] Morgan Klaus Scheuerman, Jacob M. Paul, and Jed R. Brubaker. 2019. How Computers See Gender: An Evaluation of Gender Classification in Commercial Facial Analysis Services. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW (Nov. 2019). https://doi.org/10.1145/3359246

[72] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (2020). https://doi.org/10.1145/3392866

[73] Nick Seaver. 2019. Knowing Algorithms. In *digitalSTS: A Field Guide for Science & Technology Studies*. Princeton University Press, PRINCETON; OXFORD, 412–422.

[74] Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, and Gaëlle Loosli. 2018. Baselines and a Datasheet for the Cerema AWP Dataset. In *Conférence d'Apprentissage CAp (Conférence d'Apprentissage Francophone 2018)*. Rouen, France. https://doi.org/10.13140/RG.2.2.36360.93448

[75] Susan Leigh Star and Anselm Strauss. 1999. Layers of Silence, Arenas of Voice: The Ecology of Visible and Invisible Work. *Computer Supported Cooperative Work* 8, 1-2 (March 1999), 9–30. https://doi.org/10.1023/A:1008651105359

[76] Robert Thornberg. 2012. Informed Grounded Theory. *Scandinavian Journal of Educational Research* 56, 3 (June 2012), 243–259. https://doi.org/10.1080/00313831.2011.581686

[77] Fabian L. Wauthier and Michael I. Jordan. 2011. Bayesian Bias Mitigation for Crowdsourcing. In *Proceedings of the 24th International Conference on Neural Information Processing Systems (NIPS'11)*. Curran Associates Inc., Granada, Spain, 1800–1808.

[78] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A Human-Centered Agenda for Intelligible Machine Learning. In *Machines We Trust: Getting Along with Artificial Intelligence*.

[79] Eviatar Zerubavel. 1993. *The Fine Line: Making Distinctions in Everyday Life.* (2nd ed. ed.). University of Chicago Press.

[80] Honglei Zhuang and Joel Young. 2015. Leveraging In-Batch Annotation Bias for Crowdsourced Active Learning. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining (WSDM '15)*. Association for Computing Machinery, Shanghai, China, 243–252. https://doi.org/10.1145/2684822.2685301

[81] Shoshana Zuboff. 2019. *The Age of Surveillance Capitalism: The Fight for the Future at the New Frontier of Power.* Profile Books, London.

Received January 2020; revised June 2020; accepted July 2020

A.3 Documenting Computer Vision Datasets

The question remains as to how the data annotation process can be made more transparent. Scholars have suggested checklist approaches to document this process. However, based on the data we gathered and additional interviews, we found that this is insufficient to interrogate power differentials and naturalised preconceptions encoded in annotated data. We identified four key issues that hinder the documentation of image datasets and the effective retrieval of production contexts: the involvement of various actors, the diverse purposes and forms of documentation, the perception of documentation as a burden, and problems with the intelligibility of documentation. To address these issues, we suggest considering the social and intellectual factors that lead to the practice of data annotation to make the context of data production more explicit in the documentation. We believe that transparency and adherence to ethical standards can be improved by a collective consideration of the social and intellectual factors that shape data practices, which we refer to as reflexive documentation. We argue that it helps to expose the contexts, relations, routines, and power structures that shape data.

Publication This work has been published as a full paper and was presented at the 2021 ACM Conference on Fairness, Accountability, and Transparency. It has been cited 102 times since the three years of publication.

> Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (FAccT). ACM, 161–172. isbn: 978-1-4503-8309-7. doi: 10.1145/3442188.3445880

Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices

Milagros Miceli
Technische Universität Berlin
m.miceli@tu-berlin.de

Tianling Yang
Technische Universität Berlin
tianling.yang@tu-berlin.de

Laurens Naudts
Centre for IT & IP Law (CiTiP), KU Leuven
laurens.naudts@kuleuven.be

Martin Schuessler
Technische Universität Berlin
schuessler@tu-berlin.de

Diana Serbanescu
Technische Universität Berlin
diana-alina.serbanescu@tu-berlin.de

Alex Hanna
Google Research
alexhanna@google.com

ABSTRACT

In industrial computer vision, discretionary decisions surrounding the production of image training data remain widely undocumented. Recent research taking issue with such opacity has proposed standardized processes for dataset documentation. In this paper, we expand this space of inquiry through fieldwork at two data processing companies and thirty interviews with data workers and computer vision practitioners. We identify four key issues that hinder the documentation of image datasets and the effective retrieval of production contexts. Finally, we propose reflexivity, understood as a collective consideration of social and intellectual factors that lead to praxis, as a necessary precondition for documentation. Reflexive documentation can help to expose the contexts, relations, routines, and power structures that shape data.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; • **Social and professional topics** → **Quality assurance**; *Computing industry*; • **Computing methodologies** → Computer vision problems.

KEYWORDS

datasheets for datasets, dataset documentation, reflexivity, data annotation, training data, transparency, accountability, audits, machine learning

ACM Reference Format:
Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Conference on Fairness, Accountability, and Transparency (FAccT '21), March 3–10, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3442188.3445880

1 INTRODUCTION

Since the rise of deep learning and convolution neural nets, the field of computer vision has demonstrated some of the most impressive results in machine learning [41]. Reaching a new high in popularity, computer vision models are used in a broad range of applications, penetrating ever more aspects of daily life. Creating datasets for computer vision is not straightforward. Work practices involved in gathering, annotating, and cleaning image data comprise subjective choices and discretionary decision-making [35, 39, 40]. Such decisions range from the framing of real-world questions as computational problems [5, 38] to the establishment of taxonomies to label images [32]. Data is also "the product of unequal social relations" [19] that are present among data workers as well as in the relationship between those whose data is collected and those who make use of data for research and/or profit. The opacity of industrial practices regarding computer vision datasets is a significant threat to ethical data work and intelligible systems [49].

Recent research has proposed implementing structured disclosure documents to accompany machine learning datasets [4, 22, 23, 27]. Despite their good intentions, those efforts fail to effectively reflect power dynamics and their effects on data [19, 32]. For instance, Gebru et al. [22] propose that datasheets include the question "does the dataset identify any subpopulations?" [22] e.g. by race, age, or gender. This way of documenting dataset composition is helpful. However, we argue that disclosing if a dataset includes racial categories does not speak to the problem of such categories' reductiveness, nor makes the assumptions behind race classifications embedded in datasets explicit. In the same way, asking "who created this dataset?" [22] and "who was involved in the data collection process (...) and how were they compensated?" [22] remains insufficient to interrogate hierarchies in industrial settings and their effects on data [32] . Reflecting on interests, preconceptions, and power encoded in training data [16, 19, 46] is essential for addressing many of the ethical concerns surrounding computer vision products.

In this paper, we lay our focus at the intersection of manual data processing and computer vision engineering. We investigate how work practices involved in the production of computer vision datasets can be made explicit in documentation. Although data processing can cover a variety of activities, we refer to companies where human workers collect, segment, and label image training data. Data processing companies of this kind provide data services at the request of computer vision companies (hereinafter "requesters") that wish to outsource parts of dataset production. Work between service providers and requesters requires strong coordination efforts as it comprises many actors and iterations [32]. Collaboration

FAccT '21, March 3–10, 2021, Virtual Event, Canada
ACM ISBN 978-1-4503-8309-7/21/03.
https://doi.org/10.1145/3442188.3445880

is informed by negotiation over the meanings that are ascribed to images [16]. In this context, not all actors hold equal power to shape datasets: Data processing companies generally collect and interpret data according to categories instructed by requesters, and workers often trust the judgment of their managers in case of doubt or disagreement [32]. These dynamics have a crucial effect on the datasets that train commercial computer vision products. Making them explicit in documentation can help better understand models' behavior and uncover broader ethical issues.

We base our investigation on fieldwork at two data processing companies, and several interviews with data collectors, annotators, managers, and computer vision practitioners. We identify key aspects of the effective documentation of responsibilities, decision-making, and power asymmetries that decisively shape image datasets. Our investigation is framed by the following research questions: (RQ1) How can the specific contexts that inform the production of image datasets be made explicit in documentation? (RQ2) Which factors hinder documentation in this space? (RQ3) How can documentation be incentivized?

Given the complex interweaving of actors, iteration, and responsibilities involved, documenting the context of data transformations is crucial, yet hard to achieve. We propose *reflexivity*, understood as the consideration of social and intellectual factors that predetermine and shape praxis [7], as a crucial component for retrieving and documenting power dynamics in data creation. We borrow Bourdieu's "Invitation to Reflexive Sociology" [8] and translate it into an invitation to reflexive data practices. Our invitation regards reflexivity not as personal introspection but as a collective and collaborative endeavor [8].

We start by reviewing work that investigates the documentation of machine learning datasets and models. Then, we explore different conceptualizations of reflexivity. After offering an overview of research methods, informants, and fieldwork sites, we present our findings. These are organized around four salient documentation-related issues emerging from our analysis, namely *the variety of actors involved and the collaboration among them, the different purposes and forms of documentation, the perception of documentation as burden*, and *problems around the intelligibility of documentation.* Next, we discuss the implications of our findings and propose the implementation of reflexivity in disclosure documents. Finally, we introduce and discuss four motivations which could lead companies to implement reflexivity-driven documentation, namely, *preservation of knowledge, inter-organizational accountability, auditability*, and *regulatory intervention.*

2 RELATED WORK

2.1 Documentation of Datasets and Models

Previous work has pointed at the need for opening black-box algorithms by explicating their outcomes [37, 44] and documenting their modeling [26, 34]. A growing body of literature has investigated and developed structured disclosure documents or checklists for artificial intelligence models and services, which document their intended uses, testing methodologies and outcomes, actors involved, possible bias, and ethical problems [3, 15, 26, 34]. While these disclosure documents primarily focus on AI models and services, information relevant to training datasets is also required to be reported.

Recent research [4, 22, 23, 27] has called for applying similar structured procedures for documenting datasets specifically. This line of research advocates for and applies the systematic documentation of datasets' purpose, composition, collection process, preprocessing, uses, distribution [22, 23, 47], and maintenance [11, 22, 27, 47]. Several studies also draw special attention to the documentation of actors involved, including their characteristics and roles [4, 22, 23], the use of software and other tools [4, 22, 47], availability of training and additional resources for documentation [4, 23], and fair pay for workers [22, 23, 47]. Furthermore, ethical concerns have been raised in documentation regarding privacy [22, 27, 47] and potential harms of datasets [22, 47] (see Table 1).

Most prominently, Gebru et al [22] argue that documentation can improve transparency, accountability and reproducibility, and facilitate the communication between "dataset consumers and producers". They propose that every dataset be accompanied by a checklist which should be flexible enough to accommodate specific domains and "existing organizational infrastructure and workflows" [22]. Holland et al. [27] argue that documentation of datasets can enable consumers to select appropriate datasets better and, at the same time, improve data collection practices among dataset creators, as they would need to explain and justify their practices. They propose a dataset nutrition label that is composed of modules to be filled in through a combination of manual work and automated procedures. Geiger et al. [23] focus primarily on documentation of datasets in academic settings. They maintain that documentation not only contributes to increasing reproducibility and open science, but is also a matter of "research validity and integrity" [23].

Whereas current proposals and practices of documentation often prioritize reproducibility, power imbalances in contexts of data creation are not often accounted for. In their investigation of data annotation services, Miceli et al. [32] present evidence of how power asymmetries shape computer vision datasets. In particular, the authors show how the judgements of managers and, even more, of requesters remain unquestioned when it comes to interpreting and labeling data. In view of these dynamics, D'Ignazio and Klein [19] underline the importance of restoring the context where datasets are produced, be it "social, cultural, historical, institutional, (...) [or] material," and the identities of dataset creators. They explain that "one feminist strategy for considering context is to consider the cooking process that produces 'raw' "data" [19] and propose asking "who questions" to drive reflection and analysis on power and privilege. In line with this research, we highlight the importance of looking into processes of data creation and foster disclosure documents that go beyond datasets' technical features. We argue that the dimensions proposed or applied in structured dataset documentation formats (see Table 1) are necessary but insufficient to drive a much-needed reflection of industry practitioners' and researchers' position and influence on data. For such a reflection to be possible, datasets must be placed in the context of their production. This perspective would not only provide a better understanding of datasets' "functional limitations" but can also make power asymmetries in data settings [19] visible.

Table 1: Summary of descriptive dimensions in documentation frameworks proposed or applied in previous research. It should be noted that the dimensions are often interconnected and not mutually exclusive.

Descriptive Dimensions in Documentation	Authors / Proposed or Applied Documentation Form					
	Gebru et al. [22]: Datasheets	Geiger et al. [23]: manual and technology-assisted documentation	Bender and Friedman[4]: Data statements	Holland et al. [27]: Dataset Nutrition Label	Seck et al. [47]: Datasheets	Choi et al. [11]: Datasheets
Description of dataset's motivation: private or public? single use or open dataset?	✓		✓		✓	✓
Description of actors involved: e.g. funding providers, data workers, data subjects and so on	✓	✓	✓		✓	✓
Description of dataset's composition	✓		✓	✓	✓	✓
Description of dataset's collection process	✓		✓	✓	✓	✓
Account of data (pre-)processing steps (e.g., cleaning,labeling)	✓	✓	✓		✓	✓
Description of dataset's intended and recommended uses	✓			✓	✓	✓
Description of datasets' distribution	✓	✓		✓	✓	✓
Description of datasets' maintenance	✓			✓	✓	✓
Description of software and other tools used in data work	✓		✓		✓	
Reflection on potential impacts and ethical issues relevant to datasets	✓			✓	✓	✓
Description of training for data workers		✓	✓			
Formal definitions and instructions for annotation		✓	✓			
Payment for workers	✓	✓			✓	
Team composition and diversity	✓	✓	✓		✓	
Account for production settings and hierarchies		✓				
Procedures for solving discrepencies in data production		✓				
Rationale for data collection framing and labeling taxonomies			✓			

2.2 The Notion of Reflexivity

According to D'Ignazio & Klein [19], reflexivity is a precondition for restoring context in data creation. The authors define reflexivity as "the ability to reflect on and take responsibility for one's own position within the multiple, intersecting dimensions of the matrix of domination" [19]. The matrix of domination is a concept first termed by Patricia Hill Collins [13] to explain how systems of power are configured and experienced. Black feminist scholars and critical race theorists have given considerable attention to the importance of one's positionality with regard to race, gender, and class in scientific practice. The work of Dorothy Smith [48], Patricia Hill Collins [13], and Sandra Harding [25] in standpoint theory is an important strand in this space. Researchers in critical race theory further interrogate ideological positioning of privileged and dominant groups [2, 6, 18]. More broadly, scholars on positionality frame actors' positions in socio-political contexts and scrutinize researchers' personal identities and stances concerning the contexts of knowledge and study [9, 12, 31]. These positions shape researchers' view of the world and thereby the whole research process, i.e., how they perceive, construct and approach a research problem, how they report research findings, and the process of knowledge construction and production [9, 12].

Previous investigations in sociotechnical systems have introduced reflexivity by drawing experiences and methodologies from other disciplines to examine presumptions and taken-for-granted practices in machine learning and data science. Viewing machine learning via computational ethnography, Elish and boyd [20] underline the situated nature of knowledge work and argue in favor of methodological reflections and reflexive practices. Drawing on critical race methodologies and operationalization of race in other disciplines, Hanna and Denton et al. [24] argue that the widespread conception and operationalization of race in algorithmic systems as a fixed attribute is decontextualized and, therefore, problematic. Previous work has furthermore argued that machine learning systems have positionality. Among other factors, "they inherit positionality from data" [1]. Preconceptions and values get embedded in data, for instance, through collection and analysis methods and through the taxonomies used in data annotation. The sensemaking and classification of data through labels as performed by annotators [32] is "a judgement and as such informed by the knowledge, experiences, perspectives, and value commitments of annotators or labelers" [1].

As we will explain in Discussion, Pierre Bourdieu's conceptualization of reflexivity, understood as a relational construct and

an integral part of inquiry praxis, is at the core of the documentation framework we present in this paper. Bourdieu's writings on reflexivity offer a systematic investigation into social and intellectual factors that predetermine and shape researchers' practices in scientific work [7, 8, 21]. The Bourdieusian notion of reflexivity goes beyond personal experiences and regards researchers' position at the collective level, that is, in relation to other actors and the field of inquiry as a whole. Moreover, Bourdieu's reflexivity does not aim to undermine objectivity. Instead, it is presented as an analytical tool to sensitize researchers to "the social and intellectual unconscious" that condition their thoughts and practices in research, and is, therefore, an integral part of and a "necessary prerequisite" for scientific inquiry [8]. The French sociologist pinpoints three types of bias that may influence scientific research, which may be mitigated by introducing reflexivity.*The first bias* results from researchers' positions in the social structure, such as class, gender, and ethnicity. *The second bias* comes from researchers' position in academic disciplines, i.e., academic traditions, prevailing currents, and socio-organizational structures in specific disciplines that determine specific field epistemologies. *The third bias*, termed by Bourdieu as the intellectualist bias, is embedded in the scholarly gaze that places researchers outside or above the object of research and considers their engagement with problems as purely scientific and unconstrained from social positions and economic interests. In opposition to this idea, Bourdieu argues that researchers are participants rather than external observers and restores research practices as knowledge-producing activities rather than pure and disinterested investigations. In the Discussion section, we will come back to this notion of reflexivity. The three Bourdieusian levels of bias will be the base to discuss why reflexivity is fundamental for documenting data practices. Reflexivity to make individual and collective positions explicit and acknowledge their effects on data is not only crucial for conducting better science, as Bourdieu [8] argues. It could also help researchers and practitioners uncover broader ethical issues in computer vision systems.

3 METHOD

3.1 Data Collection

This investigation was organized around two phases, involving different (yet related) research foci and methods. Documentation practices are a critical aspect we investigated at both stages:

In the first phase, we focused on work practices in data processing companies, where human workers collect, segment, and label image training data. We conducted ethnographic fieldwork at two data processing companies of the "impact sourcing" sector located in Buenos Aires, Argentina, and Sofia, Bulgaria. Impact sourcing refers to a special type of business outsourcing processing company that intentionally employs workers from marginalized communities. As described on their websites and confirmed by our observations, the Argentine company employs young people living in slums, while the Bulgarian organization works with refugees from the Middle East.

The Buenos Aires-located company that we will call "Emérita" is a medium-sized organization. With branches in three Latin American countries, Emérita conducts projects in data annotation, content moderation, and software testing. Its clients are large regional corporations in diverse fields such as security, e-commerce, and energy. At the time of the observations, between May and June 2019, the Buenos Aires branch of Emérita had around 200 data-related employees who mostly worked 4 hours shifts, Mondays to Fridays, and were paid at the minimum wage.

"Action Data" is the code-name of the Bulgarian company. Action Data specializes in image data collection, segmentation, and labeling. Its clients are computer vision companies, mostly located in North America and western Europe. The company offers its workers contractor-based work and the possibility to complete their assignments remotely, with flexible hours. Contractors are paid per picture or annotation, and payment varies according to each project and its difficulty. At the time of the observations, in July 2019, the Bulgarian company was very small in size. Three employees in salaried positions and a pool of around 60 contractors handled operations.

At both sites, we conducted several weeks of observations, with different levels of interaction and involvement. All tasks observed were related to the production of datasets for computer vision and requested by computer vision companies. Moreover, we observed the on-boarding, briefing, and further training of workers as well as instances of communication between managers and teams, and managers and requesters. It is important to mention that the observations were primary conducted with a different research question in mind and focused on general work practices and not specifically on documentation. However, the exploratory character of the method and the rich interactions observed allowed us to extract useful insights for this investigation that were later corroborated by our interview partners.

In addition to the observations, fieldwork at both sites also consisted of intensively interviewing data collectors, annotators, and management. In total, we conducted sixteen in-depth interviews with an average length of 65 minutes, face-to-face, at both locations. Informants were aged 21 to 40. Eleven of them identified as female and four as male. None of them had received an education in tech-related fields or had technical knowledge prior to their current employment. At Emérita in Argentina, we conducted five in-depth interviews with data workers and employees in managerial positions. At Action Data, we conducted eleven in-depth interviews with workers and managers. Interview partners were asked to choose code names to preserve their identity and that of related informants. The interviews included accounts of specific work situations involving the interpretation of data, the communication with managers and clients, and the documentation of responsibilities and decisions. Moreover, the interviews covered task descriptions, general views on the company and the work, informant's professional and educational background, expectations for the future, and biographical details.

The second phase of this investigation dealt with the role of stakeholders at the opposite end of the service relationship, namely, the computer vision companies requesting data processing services. At fieldwork, we observed that requesters have a major influence on the documentation practice of data processing companies and decided to pursue this line of inquiry. Through expert interviews with computer vision engineers, data quality analysts, and managers, we investigated how task instructions are formulated and

communicated to data processing workers, and how this process is documented. The interviews revolved around the object, purpose, and responsibilities of documentation. Moreover, we discussed issues and possible solutions for implementing broader forms of documentation in industrial contexts at the intersection of data processing and computer vision.

We conducted a total of fourteen expert interviews. Four informants were managers with large data processing companies located in Kenya, India, and Iraq. In addition, six expert interviews were conducted with computer vision practitioners working on products including an aesthetics model that sorts and rates personal image libraries, a scanner that detects contamination on hands, and optical sorting equipment for the classification of waste. The computer vision practitioners work for companies located in Germany, Spain, and the United States. Finally, four of the interviews conducted at Emérita and Action Data revolved almost exclusively around the role of requesters in documentation and were framed as expert interviews.

While the goal of in-depth interviews is revealing practices and perceptions, the purpose of expert interviews is to obtain additional professional assessments on the research topic [29]. The sampled interview partners were considered experts because they were able to provide unique insights into widespread routines and practices in their and other companies. With an average length of 48 minutes and conducted face-to-face or remotely, the expert interviews allowed us to contextualize some of the practices observed at fieldwork and analize to what extent observations could be generalized to other settings.

3.2 Data Analysis

For the analysis, we integrated field notes with a total of thirty interview transcriptions and used constructivist grounded theory principles [10] to code and interpret the data. We conducted phases of open, axial, and selective coding and let the categories emerge from the data. We applied a set of premises [14] to make links between categories visible and make them explicit in our research documentation and in open discussions among three coders. We constantly compared the collected data to revise our emergent understanding or find additional evidence of observed phenomena. Four salient axial dimensions identified during the analysis process constitute the base for the findings we present in the following section.

4 FINDINGS

As stated in Introduction, this paper explores three research questions: (RQ1) How can the specific contexts that inform the production of image datasets be made explicit in documentation? (RQ2) Which factors hinder documentation in this space? (RQ3) How can documentation be incentivized? Our findings unpack documentation practices at the intersection of data collection, data annotation, and computer vision engineering. Through descriptions and interview excerpts, we describe salient dimensions emerging from our data: *actors and collaboration, documentation purpose, documentation as burden,* and *intelligibility of documentation.* These four dimensions reveal scenarios that should be taken into account for creating effective documentation procedures that are based on workers' needs and possibilities.

4.1 Actors and Collaboration

Our first research question inquires about ways of making the specific production contexts of image datasets explicit in documentation. In this section, we take a first step towards unpacking RQ1 by describing the characteristics of such production contexts.

The creation of computer vision datasets requires the collaboration of actors that often work in different organizations. At the intersection of data collection, data annotation, and computer vision engineering, not every actor has the same influence on data [32]. Power differentials become evident when deciding which data to collect, how to classify it, and how to label it. Many datasets are produced with a specific computer vision product in mind. Dataset design begins as the expected outcome of that product (in terms of computational output but also of revenue) is transformed into task instructions for data collectors and annotators. A typical assignments is illustrated by a data collection project of Active Data: the company received task instructions to collect images of diverse human faces from a Western European company, producing identification and verification systems. Eva, the founder of Active Data, offered more details:

> "They were interested in a diversity of five different ethnicities, so Caucasian, African, Middle Eastern, Latin American and Asian. Of course, very debatable whether these can be the five categories that can classify people around the world "

This type of assignment generally revolves around a client's envisioned computer vision product and underlying business idea. The technical assumptions of a classification system demand mutually exclusive categories, in this case even for a problematic concept such as race. Whether such categorisation captures the realities of data subjects or coincides with the values and believes of data workers is not negotiated. Written instructions formulated by the requester are passed along to project managers who brief workers. Workers then start collecting the images. For outsourcing companies, the rationale behind data-related decisions is "doing what the client ordered" and "offering value to the client." Conversely, the rationale shaping datasets in computer vision companies is "data needs to fit the model" and "data processing should be fast, cost-efficient, and high-quality."

Power differentials between service providers and requesters become even more evident given that the data processing companies participating in this investigation are located in developing countries, while their clients are in the Global North. In view of such asymmetries, decisions about what to document and the financial means to do so largely depend on the most powerful actors. Anna, an intern working at Action Data and in charge of auditing the company and conducting an impact assessment, concisely described these dynamics:

> Q: "What do you think are the potential drivers or reasons for the implementation of the more transparent approach to documenting systems and processes?"
> A: "If the customer demands it."

> Q: "Is this something you have heard before, customers demanding a more ..."
> A: "No."

Moreover, computer vision companies often regard some of the information that could or should be documented as confidential, especially if it involves details about the intended product or if some of the processes involved in producing the dataset are considered a strategic advantage. Given the collaborative nature of data creation, one stakeholder's opacity may affect others' inclination towards transparency. As Active Data's founder Eva (and several others of our informants) described, secrecy in computer vision hinders her company's attempts to document work processes:

> "It's also a small challenge of how to preserve some of the know-how throughout the different projects without of course revealing too much about the different processes that each client has, you know, the confidential information from each project."

In many cases, this issue leads to reluctance to share existing documentation with other stakeholders and the general public or to not document at all.

4.2 Documentation Purpose

The reasons for documenting the production of datasets and the forms of documentation vary with each organization. To start considering ways of incentivizing documentation (RQ3), we first must look into common needs and goals that different stakeholders may have in relation to disclosure documents. In this sense, we have identified four common documentation purposes: *preservation of knowledge, improvement of work practices, accountability,* and *disclosure of dataset's specifications.*

All data processing companies participating in this investigation carry out some form of project documentation. In a more or less structured way, companies document task instructions provided by clients. Instructions may change as projects develop, or workers might develop new practices according to clients' feedback. Soo is a project manager at the Kenyan branch of a large data processing company. During our interview, he explained how this form of documentation can help improve existing processes and practices:

> "We have a 'lessons learned'- folder where we put all these items. Like the client has said, 'You did not do well here.' We'll find in our process, there was this flaw. We will document that. And then what happens after we document is that information is stored to be used for that project and some future projects with the same kind of process work."

The *preservation of this form of praxis-based knowledge* is crucial because it helps organizations resolve doubts that might emerge, train future workers, and apply situated solutions to future projects. Similarly, documentation can also serve to *revise and improve work practices and flows*, as further described by Soo:

> "How can we improve this process? This did not go well. What was the issue? How did we solve it? How can we avoid this in future? And you will get information for a project that was done five years ago [...] The documentation helps us in making sure that we avoid repeating the same mistakes. And also, it helps us in looking for better ways of doing the work, how to measure where it is possible and also what other process we can improve, like in the process flow"

Given the differentials of power described in the previous section, documentation is many times perceived as useful for *accountability* between outsourcers and requeters. Several informants working at data processing companies highlight the importance of preserving task instructions and documenting changes instructed by clients. Keeping this type of record might serve as proof that tasks were carried out as instructed. In the next interview excerpt, the founder of Active Data describes how documentation might help resolve discrepancies if clients are not satisfied with the quality of the service provided or decide to demand more:

> "We also keep the client accountable so that they don't come up with a new requirement or something that we haven't mentioned before. So, SoWs [scope of work documents] are also for accountability of us towards the client as well so that the client can have a document where they can keep track of what the arrangement is and so on beyond our contract"

However, accountability within teams can become surveillance for workers: several informants account for the connection between project documentation and the measurement of workers' performance in data processing companies. The Argentine company, Emérita, directs great efforts to measure workers' performance and output quality and to transform those into numbers and charts. Nati, Emérita's continuous improvement analyst, described this process:

> "Within the project documentation, we have an external person who checks if the work the team did is right or wrong, then documents the percentage of right and wrong. [...] If something is wrong, we fix it before the client notices. But still, even when it is fixed, we record that there was something that was wrong and record who was responsible for the mistake."

Finally, in the case of datasets for public use or without a pre-established purpose, organizations might find it important to document and disclose *datasets' specifications.* This particular case was reported by our informants at Action Data, as the company had recently released two datasets for public use. During an interview, Eva contemplated the possibility of releasing a disclosure document along with the datasets:

> "It might be nice to implement some type of documentation at least for them [datasets for open use] because they're for external use and it might be good to know what the origin of the images are, what the process of annotation had been and so on."

It is worth mentioning that releasing datasets for public use is usually not within the scope of outsourcing companies. Investing resources to produce a pro-bono dataset represents a considerable effort for these companies. In the case of Active Data, the dataset was made publicly available as part of the company's marketing strategy.

4.3 Documentation as Burden

Relevant to start unpacking factors that hinder documentation (RQ2) is the fact that several informants see documentation as time-consuming, extra work that is likely to delay the completion of workers' "actual" tasks. This is a widespread view among the computer vision practitioners interviewed for this investigation and coincides with the observation that, among the different roles explored in this study, computer vision companies seem to be the least inclined to document work practices.

"Lack of time" is the most widespread answer when informants are asked why there are not more aspects of data creation reflected in reports. Documentation is broadly perceived as optional, a nice-to-have feature that is implemented only once all "important" issues are sorted. Andre, a US-based computer vision engineer with a start-up dedicated to producing scanners that detect contamination on hands, described his company's position on this issue:

> "[Documenting] is lower on our priority list than a bunch of other things that we need to do. It's just not the company's priority at this moment. There are other more valuable things to keep the company successful. As the engineering team grows, as we have more time to do those things and our work to meet the company's exact needs are less burdensome, then we'd go to more documentation."

Among our informants in computer vision companies, the view persists that documentation is an activity only large corporations can afford. As further reported by Andre, start-up teams are smaller, and workers are multitasking, which reinforces the view that there are more pressing issues than documentation:

> "That's one of the interesting things about start-ups. You don't have the time to document everything. [...] There is a lot of knowledge in every single person here that would take far too long to pull out of them and transfer to a new person and keep the company still running at the same time."

A similar observation was made by Eva, the founder of the Bulgarian data processing company, regarding her company's clients:

> "We've been working with quite a lot of new companies recently. Some of them are bigger corporations that have more let's say bureaucratic procedures and more detailed processes of description of everything that's happening around the project, while others are just start-ups that prefer very lightweight, minimum involvement and paperwork around their projects."

Lack of incentives, external or internal, is another reason why documentation might be perceived as a burden. For instance, some informants agreed that laws and regulations would be an excellent external incentive for technology companies to integrate documentation as a constitutive part of their work. In the absence of regulations, documentation is seen as optional extra work. As for internal incentives within organizations, several computer vision practitioners explained that documenting was not a part of their work routines and was therefore not encouraged by the company's structures. Emmanuel, a computer vision engineer based in Barcelona and working on optical sorting equipment for waste's classification, discussed the need for integrating documentation in existing workflows. He moreover imagines that extending projects' deadlines to prioritize documentation would not be seen as acceptable within his company's culture:

> "Time is a huge issue. I mean, I think planning is very important, get the time to do it [documenting] and that everybody knows this is supposed to be done. Because right now, documenting is not a task and I don't know that I would have a gap between projects so I could document. And this is never a priority for the company, they expect me to meet my deadlines, I can't just drop my deadlines to document. And this is a problem. If documenting was part of the deadline, companies wouldn't just leave it for another time"

Even in companies that integrate laborious documentation in their work processes, as is the case of Emérita, there are instances where documenting is just not profitable. Nati, one of our informants with the Argentine data processing company, describes one of those situations:

> "It happens sometimes that we do one-time projects that go only for one or two weeks. In those cases, documentation is a waste of time and money, because the client buys, let's say, eighty hours and you spend twenty documenting. It's just not profitable."

As expected, financial incentives, or the lack thereof, can also influence views on documentation.

4.4 Intelligibility of Documentation

To further investigate factors that hinder documentation (RQ2) it is necessary to explore issues around creating compelling, retrievable, and intelligible disclosure documents. To illustrate some relevant aspects related to structuring and providing access to documentation, we draw on the observations made during fieldwork at both data processing companies, Emérita and Active Data. Both companies have vast experience in the documentation of data collection and annotation projects.

In the case of the Argentine company, Emérita, due to the extension of documentation and the large number of projects conducted, navigating and maintaining disclosure documents has become difficult. Nati, a continuous improvement analyst, is in charge of addressing this issue:

> "What happened a lot was that information was repeated in many places. The objectives were written in three different documents. The people who were in the project were in two different systems [...] So, having that repeated was horrible, because every time people in the team changed, well, you needed to update many things and credentials"

Nati works on optimizing some of her company's internal processes, including documentation. For that purpose, she has surveyed project documents, observed how the company teams work, and discussed with them how documentation can be improved. Her main focus lies in producing documentation that can be easily retrieved and used, which can be very challenging:

> "For example, in the case of project guides, it was not clear what documentation had to be done, so everyone did what they wanted, or what they remembered, or what they knew, because someone told them, and when information was needed, they didn't know if it had been documented or not, or they didn't know where to find it. We lost a lot of information like this."

Further issues related to the intelligibility of documentation may arise depending on who is in charge of documenting and who are the users of documentation. In the case of Active Data, the Bulgarian company working with refugees from the Middle East, language and lack of technical knowledge is one of those issues:

> "Since we're working with people who very frequently do not have high levels of education or do not speak good English, I've heard a lot of complaints that people are not reading the training documents or they're not following them or they're asking questions that appear or are already answered in the training documents. So, it can be quite frustrating because people may not be used to following such documentation and they might need additional training just to know how to use this recommendation, how to read it and how to follow it"

Creating useful reports that can be easily retrieved and understood is challenging. How disclosure documents are created, indexed, and stored depends to a greater extent on the intended addressees of documentation. As illustrated by the previous interview excerpt, language is important if stakeholders with different levels of literacy will make use of documentation.

5 DISCUSSION

As described in Findings, work at the intersection of data collection, annotation, and computer vision engineering requires strong coordination efforts among actors that occupy different (social) positions. Documentation purpose, organizational priorities, and needs around documentation intelligibility vary across stakeholders. In such heterogeneous contexts, some actors hold more power than others and decisions made at the most powerful end will inevitably affect work practices and outputs at every level. These power differentials and their effects are broadly naturalized [17, 19, 32]. Despite their decisive effects on data, decisions and instructions that are rooted in such naturalized power imbalances are mostly perceived as self-evident and remain undocumented as a consequence.

Previous research has emphasized the importance of documenting machine learning datasets [22, 23, 27, 30, 49]. While we acknowledge that work for creating the foundations for our investigation, we also argue that the frameworks proposed are not sufficient to interrogate power differentials and naturalized preconceptions encoded in data. With our investigation, we move the focus away from documenting datasets' technical features and highlight the importance of accounting for production contexts. Our research questions address the challenge of documenting production processes that are characterized by the multiplicity of actors, needs, and decision-making power. In this and the following sections, we lay out implications of our observations and outline a documentation framework to address the contexts and issues described in Findings.

Given the collaborative nature of datasets production, we argue that documentation should not be carried out in the vacuum of each organization. The framework we propose regards dataset documentation as a collaborative project involving all actors participating in the production chain. This is not easy for sure. To address such challenge, we propose that reflexivity, understood as a collective endeavor [7], be an integral part of such collaborative documentation. As argued by Bourdieu [8], this form of collective reflexivity accounts for actors' social position and aims to interrogate praxis fields and the relations that constitute them. In a similar manner, reflexive documentation should help to make visible the interpersonal and inter-organizational relations that shape datasets. As described in the Related Work section, Bourdieu's notion of reflexivity covers three levels of hidden presupposition: the researcher's social position, the epistemology of each disciplinary field, and "the intellectualist bias", described as the scholarly gaze researchers use to analyze the social world as if they were not part of it [7, 8]. We take this perspective and transform Bourdieu's "Invitation to Reflexive Sociology" [8] into an invitation to reflexive data practices. What constitutes our invitation entails much more than observing how one actors' positionality affects data: If documentation is to be seen as a collaborative project, reflexivity of work practices should be understood as a collective endeavor, where widespread assumptions, field methodologies, and power relations are interrogated.

With this framework, we regard documentation in a two-fold manner: First, as an artifact (the resulting documentation) that enables permanent exchange among stakeholders participating in data creation. We envision disclosure documents that travel among actors and organizations, across cultural, social, and professional boundaries, and are able to ease communication and promote inter-organizational accountability. Second, we regard documentation as a set of reflexive practices (the act of documenting) intended to make naturalized preconceptions and routines explicit. Just as Bourdieu regards reflexivity as a "necessary prerequisite" for scientific inquiry [8], the reflexive practices involved in our documentation framework should be seen as a constitutive part of data work. If reflexivity is only regarded as a desirable goal related to AI ethics and not as actual *part of the job*, documentation will never be considered a priority and, as described in Findings, it will continue to be perceived as a burden.

5.1 Why Reflexivity?

Our research questions enquire about ways of making the contexts that inform the production of image datasets explict in documentation and about factors that hinder or incentivize the implementation of documentation in industry settings. In view of our findings, we argue that effective documentation should be able to reflect the dynamics of power and negotiation shaping datasets through work practices. However, making visible the hierarchies, worldviews, and interests driving decisions and instructions is extremely challenging. One major difficulty lies in their taken-for-grantedness: documenting naturalized power dynamics and decisions that are

largely perceived as self-evident [33] require intensive reflexive practice.

The three previously-mentioned levels of reflexivity proposed by Bourdieu (social position, field epistemology, and intellectualist gaze) can be useful to discuss why reflexivity should be at the core of documentation practices in data creation for computer vision. They provide an additional lens through which data practices can be approached, and as such, serve as a complement to on-going work and discussions regarding the documentation of datasets:

First, reflexive documentation should consider the social position of workers involved in dataset production, not just individually but in their relation to other stakeholders. Such consideration could help produce documentation that brings power imbalances into light and questions taken-for-granted instructions and hierarchies. This relational examination is especially important due to the widespread use of outsourced services for the collection and annotation of data: Workers at crowdsourcing platforms are subject to precarious employment conditions [28, 45]. In the impact sourcing companies presented in this paper, workers come from marginalized communities (refugees in Active Data, slum residents in Emérita). Most of them have no technical education. How does their social position affect these workers' ability and power to question the instructions commanded by computer vision engineers or data scientists in tech companies? This question becomes even more pressing if we examine the relationship that connects data processing services in developing countries with computer vision companies in the Global North. Documentation frameworks that are oblivious to the fact that production chains are shaped by asymmetrical relationships will never be effective in reflecting how those asymmetries affect data. In this sense, reflexive documentation should bring power differentials to light and, ideally, empower those in vulnerable positions to speak up and raise questions.

Second, reflexive documentation should serve to question field epistemologies. Examining the epistemology of computer vision might shed light on the assumptions, methods, and framings underlying the production of image datasets. As Crawford and Paglen [16] argue, computer vision is "built on a foundation of unsubstantiated and unstable epistemological and metaphysical assumptions about the nature of images, labels, categorization, and representation." Bringing these assumptions forward in documentation is important because socially-constructed categories, such as race and gender, are generally presented as indisputable in image datasets [46]. Furthermore, a fixed and universal nature is not only ascribed to the categories as such, but also to the correspondence that supposedly exists between images and categories, appearances and essences [16]. Reflexivity should help reveal the political work such assumptions perform behind their purely technical appearance.

Finally, reflexive documentation should help practitioners question the "intellectualist gaze" [7] in data work. This type of bias is the inclination to place ourselves outside the object of research. This form of examination would highlight the role of workers and organizations in creating data while questioning widespread notions such as "raw data" and "ground truth labels". Reflexivity should therefore help to adopt a relational view on data and data work, acknowledging data as a "human-influenced entity" [35] that is shaped by individual discretion, (inter-)organizational routines, and power dynamics.

5.2 Why Document?

Data processing services and computer vision companies might be reluctant to implement such an elaborate approach to documentation. Our third research question asks how can documentation be incentivized. In this section, we consider four ways in which the Bourdieusian framework previously outlined can constitute an asset for organizations, and thus serve as an incentive for the uptake of reflexive documentation.

5.2.1 Preservation of Knowledge. Reflexive documentation could make praxis-based and situated decision-making explicit and help preserve it in documentation. This knowledge can become long term business assets for companies. Moreover, reflexive documentation can preserve know-how relevant to data work [39] that may get lost due to workers flow. As the flow of workers brings about problems in task transfer and reinvestment in training new employees, documentation that preserves knowledge and methods for effective data work, be they project-specific or not, can ease the transition.

Furthermore, documentation can "have analytical value [and] improve communication in interdisciplinary teams" [32]. The framework offered in this paper highlights the collective nature of reflexivity. We argue that documentation that preserves praxis-based knowledge and best practices (as described in section 4.2) should be circulated among collaborating companies rather than be produced and retrieved in the vacuum of each organization. For one thing, sharing such documentation with other stakeholders may improve the quality of data work and of the datasets that are produced as a result. For another, documentation providing more details on discretionary decision-making and its contexts can enhance transparency and facilitate a better understanding of datasets before model development.

5.2.2 Inter-organizational accountability. Tracking decisions and responsibilities in environments and processes that involve multiple organizations can be challenging. As described in Findings, data processing companies use documentation to foster inter-organizational accountability and protect themselves in the face of disagreements with clients. At the same time, computer vision companies might consider documentation as a tool to keep track of the processing status of projects and audit requested tasks. Reflexive documentation could be especially useful to improve traceability, as the participation of many actors and iterations in data creation may lead to accountability dilution [32]. Moreover, documentation could provide "organizational infrastructure" that empowers individual advocates among workers to raise concerns and reduces the social costs for such actions [30]. An infrastructure based on the reflexivity framework outlined in this paper could facilitate the interrogation of intra- and inter-organizational relations, normative assumptions, and workflows shaping data at the three levels described in the previous section.

Conducting documentation at a collaborative level, which means to engage various actors and to accommodate documentation to their needs, can serve as a platform for permanent exchange among stakeholders. Enabling permanent exchange could help anticipate disagreements and misunderstandings, thus improving task quality and reducing completion time.

5.2.3 Auditability. Documentation based on reflexivity could constitute an asset for organizations to prevent issues before they are made public or weather the storm in the face of PR failures. Disclosure documents that are able to retrieve the context of dataset production could constitute a useful tool for auditability, for instance, when computer vision outputs are publicly questioned or for internal ethics teams who would like to perform an assessment for potential fairness concerns prior to the release of a model trained on such data [42, 43]. Such documents could help to identify problematic issues before they become public pushbacks. Moreover, in case of public failures, documentation could provide an audit trail that would allow organizations to address problems and offer solutions promptly. In this sense, public pressure could constitute an incentive for companies towards documentation.

In such cases, counting with reflexive documentation to audit datasets could help companies offer solutions that go beyond "throwing in more data" and are able to address issues at the three Bourdieusian levels previously described: identifying asymmetrical relationships that might have been encoded in datasets, interrogating widespread assumptions in computer vision, and questioning data, even "raw" data.

5.2.4 Regulatory Intervention. Organizations could also be pushed towards documentation through regulatory intervention. Yet, before any form of reflection, including the documentation thereof, can be imposed, a few observations can be made:

First, while documentation might be considered an important component or step of the reflexive process, it is neither constitutive to, nor sufficient for, reflection. Reflexivity represents a state of awareness, an encouragement for actors involved in data creation to more widely consider the impact of their practices. Reflexivity can already be valuable in itself. The policy end-goal is therefore to stimulate a reflexive mindset and to establish the right conditions for such a mindset to fully come to fruition. Conversely, if regulation only aims at pushing documentation, the danger exists that such regulatory requirements are approached as merely an administrative exercise towards compliance.

Second, if the encouragement of reflexivity through legal means would be desired, such mechanisms may already be (partially) present in existing initiatives. For instance, it could be argued that the EU General Data Protection Regulation's increased emphasis on accountability and risk-based responsibility stimulates some level of reflection where personal data are involved [36]. Reflexivity could moreover become an additional supportive tool for data workers as a means to detect and mitigate the impact data actions have on (fundamental) rights, and as such, contribute towards the compliance with existing legal frameworks.

Third, given the multiplicity of actors involved in data creation, regulatory initiatives should also carefully consider the actors they wish to target. Stakeholders should not only be targeted in isolation; instead, policy makers should understand the relationships these actors hold vis-a-vis one another, and the consequences that their relationships bear on the activities performed.

Finally, any regulatory response must adequately consider the power asymmetries described in this paper, including their manifestation within a globalized, international environment. Mechanisms of provenance, such as documentation, could help ensure and demonstrate that societal values and fundamental rights, as well as an appropriate level of reflexivity, have been maintained throughout the computer vision value chain, rather than purposefully avoided via outsourcing strategies and/or the exercise of power. Similarly, provenance may increase the accountability and responsibility of powerful entities in both their actions and their given instructions.

6 LIMITATIONS AND FUTURE WORK

This investigation was designed to be qualitative and exploratory. Our findings are bound to the specific contexts of the companies and individuals participating in our studies and cannot be generalized to all computer vision production settings. In the future, we seek to broaden this research by investigating ways of integrating the framework outlined in this paper in real-world production workflows and co-designing actionable guidelines for reflexive documentation together with industry practitioners.

7 CONCLUSION

Based on fieldwork at two data processing companies and interviews with data collectors, annotators, managers, quality assurance analysts, and computer vision practitioners, we described widespread documentation practices and presented observations related to the purpose, challenges, and intelligibility of documentation.

In view of these findings, we proposed a reflexivity-based approach for the documentation of datasets, with a special focus on the context of their production. We described documentation as a set of reflexive practices and an artifact that enables permanent exchange among actors and organizations. We argued that disclosure documents should travel across organizational boundaries, and be able to ease communication and foster inter-organizational accountability. We imagined documentation as a collaborative project and argued that reflexivity of work practices should therefore be understood as a collective endeavor, where not only personal positions but also praxis fields are interrogated.

Achieving a healthy balance between these elements and incentivizing practitioners and organizations to implement reflexive documentation is not easy. The challenge is nevertheless worth exploring if we aim at addressing some of the ethical issues related to the production of data for computer vision systems.

8 ACKNOWLEDGMENTS

Funded by the German Federal Ministry of Education and Research (BMBF) – Nr. 16DII113f. Laurens Naudts received support from the Weizenbaum Institute Research Fellowship programme. We dearly thank the individuals and organizations participating in this study. Thanks to Philipp Weiß for his help with Overleaf and to Leon Sixt, Matt Rafalow, Julian Posada, Gemma Newlands, and our anonymous reviewers for their valuable feedback. Special thanks to Prof. Bettina Berendt for her continuous support.

REFERENCES

[1] Yewande Alade, Christine Kaeser-Chen, Elizabeth Dubois, Chintan Parmar, and Friederike Schüür. 2019. Towards Better Classification. (2019), 4. https://drive.google.com/file/d/14uL1DQN8hRyDDDAm2WEleYbmxP7dqP72/view

[2] Michelle Alexander. 2012. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness* (revised edition ed.). New Press, New York.

[3] M. Arnold, D. Piorkowski, D. Reimer, J. Richards, J. Tsay, K.R. Varshney, R. K. E. Bellamy, M. Hind, S. Houde, S. Mehta, A. Mojsilovic, R. Nair, K. Natesan Ramamurthy, and A. Olteanu. 2019. FactSheets: Increasing trust in AI services through supplier's declarations of conformity. *IBM Journal of Research and Development* 63, 4/5 (2019), 6:1–6:13. https://doi.org/10.1147/JRD.2019.2942288

[4] Emily M. Bender and Batya Friedman. 2018. Data Statements for Natural Language Processing: Toward Mitigating System Bias and Enabling Better Science. *Transactions of the Association for Computational Linguistics* 6 (2018), 587–604. https://doi.org/10.1162/tacl_a_00041

[5] Bettina Berendt. 2019. AI for the Common Good?! Pitfalls, challenges, and ethics pen-testing. *Paladyn, Journal of Behavioral Robotics* 10, 1 (Jan. 2019), 44–65. https://doi.org/10.1515/pjbr-2019-0004

[6] Eduardo Bonilla-Silva. 2006. *Racism without Racists: Color-Blind Racism and the Persistence of Racial Inequality in the United States.* The Rowman & Littlefield Publishing Group, Inc., Lanham. OCLC: 781274997.

[7] Pierre Bourdieu. 2000. *Pascalian meditations.* Stanford University Press, Stanford, Calif. OCLC: 833852849.

[8] Pierre Bourdieu, Loïc J. D. Wacquant, and Loïc J. D. Wacquant. 1992. *An Invitation to Reflexive Sociology.* University of Chicago Press. tex.ids: bourdieu1992b googlebooksid: rs4fEHa0ijAC.

[9] B. Bourke. 2014. Positionality: Reflecting on the Research Process. *The Qualitative Report* 19, 33 (2014), 1–9. https://nsuworks.nova.edu/tqr/vol19/iss33/3

[10] Kathy Charmaz. 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis.* Sage Publications, London ; Thousand Oaks, Calif.

[11] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC : Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184. https://doi.org/10.18653/v1/D18-1241

[12] David Coghlan and Mary Brydon-Miller (Eds.). 2014. *The Sage encyclopedia of action research.* SAGE Publications, Inc, Thousand Oaks, California.

[13] Patricia Hill Collins. 1990. *Black feminist thought: knowledge, consciousness, and the politics of empowerment.* Number v. 2 in Perspectives on gender. Unwin Hyman, Boston.

[14] Juliet M. Corbin and Anselm L. Strauss. 2015. *Basics of qualitative research: techniques and procedures for developing grounded theory* (fourth edition ed.). SAGE, Los Angeles. https://us.sagepub.com/en-us/nam/basics-of-qualitative-research/book235578 tex.ids: dnsc2015.

[15] Henriette Cramer, Jean Garcia-Gathright, Sravana Reddy, Aaron Springer, and Romain Takeo Bouyer. 2019. Translation, Tracks & Data: An Algorithmic Bias Effort in Practice. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (CHI EA '19)*. ACM, New York, NY, USA, CS21:1–CS21:8. https://doi.org/10.1145/3290607.3299057 event-place: Glasgow, Scotland Uk.

[16] Kate Crawford and Trevor Paglen. 2019. Excavating AI: The Politics of Images in Machine Learning Training Sets. https://www.excavating.ai tex.ids: zotero-3263.

[17] Emily Denton, Alex Hanna, Razvan Amironesei, Andrew Smart, Hilary Nicole, and Morgan Klaus Scheuerman. 2020. Bringing the People Back In: Contesting Benchmark Machine Learning Datasets. *arXiv:2007.07399 [cs]* (July 2020). http://arxiv.org/abs/2007.07399 arXiv: 2007.07399.

[18] Robin J. DiAngelo. 2018. *White fragility: why it's so hard for white people to talk about racism.* Beacon Press, Boston.

[19] Catherine D'Ignazio and Lauren F. Klein. 2020. *Data feminism.* The MIT Press, Cambridge, Massachusetts. https://mitpress.mit.edu/books/data-feminism

[20] M. C. Elish and danah boyd. 2018. Situating methods in the magic of Big Data and AI. *Communication Monographs* 85, 1 (Jan. 2018), 57–80. https://doi.org/10.1080/03637751.2017.1375130

[21] Mustafa Emirbayer and Matthew Desmond. 2012. Race and reflexivity. *Ethnic and Racial Studies* 35, 4 (April 2012), 574–599. https://doi.org/10.1080/01419870.2011.606910

[22] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2020. Datasheets for Datasets. *arXiv:1803.09010 [cs]* (March 2020). http://arxiv.org/abs/1803.09010 arXiv: 1803.09010.

[23] R. Stuart Geiger, Kevin Yu, Yanlai Yang, Mindy Dai, Jie Qiu, Rebekah Tang, and Jenny Huang. 2020. Garbage in, garbage out? do machine learning application papers in social computing report where human-labeled training data comes from?. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, Barcelona, Spain, 325–336. https://doi.org/10.1145/3351095.3372862

[24] Alex Hanna, Emily Denton, Andrew Smart, and Jamila Smith-Loud. 2020. Towards a Critical Race Methodology in Algorithmic Fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAT* '20)*. Association for Computing Machinery, Barcelona, Spain, 501–512. https://doi.org/10.1145/3351095.3372826 tex.ids: hanna2020a.

[25] Sandra Harding. 1993. Rethinking Standpoint Epistemology: What is "Strong Objectivity"? In *Feminist Epistemologies.* Routledge, 49–82.

[26] Michael Hind, Stephanie Houde, Jacquelyn Martino, Aleksandra Mojsilovic, David Piorkowski, John Richards, and Kush R. Varshney. 2020. Experiences with Improving the Transparency of AI Models and Services. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems (CHI EA '20)*. Association for Computing Machinery, 1–8. https://doi.org/10.1145/3334480.3383051

[27] Sarah Holland, Ahmed Hosny, Sarah Newman, Joshua Joseph, and Kasia Chmielinski. 2018. The Dataset Nutrition Label: A Framework To Drive Higher Data Quality Standards. *arXiv:1805.03677* (2018). .http://arxiv.org/abs/1805.03677

[28] Lilly C. Irani and M. Six Silberman. 2013. Turkopticon: interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. Association for Computing Machinery, Paris, France, 611–620. https://doi.org/10.1145/2470654.2470742

[29] Natalia M Libakova and Ekaterina A Sertakova. 2015. The Method of Expert Interview as an Effective Research Procedure of Studying the Indigenous Peoples of the North. *Journal of Siberian Federal University. Humanities & Social Sciences* 8, 1 (2015), 114–129. https://doi.org/10.17516/1997-1370-2015-8-1-114-129

[30] Michael A. Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-Designing Checklists to Understand Organizational Challenges and Opportunities around Fairness in AI. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (CHI '20)*. Association for Computing Machinery, Honolulu, HI, USA, 1–14. https://doi.org/10.1145/3313831.3376445 tex.ids: madaio2020a.

[31] Frances A. Maher and Mary Kay Tetreault. 1993. Frames of Positionality: Constructing Meaningful Dialogues about Gender and Race. *Anthropological Quarterly* 66, 3 (1993), 118–126. https://doi.org/10.2307/3317515

[32] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proc. ACM Hum.-Comput. Interact.* 1, 1 (2020), 25. https://doi.org/10.1145/3415186

[33] Milagros Miceli, Martin Schüßler, and Tianling Yang. 2020. Between Subjectivity and Imposition: A Grounded Theory Investigation into Data Annotation. (2020), 19.

[34] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model Cards for Model Reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, 220–229. https://doi.org/10.1145/3287560.3287596

[35] Michael Muller, Ingrid Lange, Dakuo Wang, David Piorkowski, Jason Tsay, Q. Vera Liao, Casey Dugan, and Thomas Erickson. 2019. How Data Science Workers Work with Data: Discovery, Capture, Curation, Design, Creation. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems (CHI '19)*. Association for Computing Machinery, Glasgow, Scotland Uk, 1–15. https://doi.org/10.1145/3290605.3300356

[36] Laurens Naudts. 2019. How Machine Learning Generates Unfair Inequalities and How Data Protection Instruments May Help in Mitigating Them. In *Data Protection and Privacy : The Internet of Bodies* (first ed.), Ronald Leenes, Rosamunde van Brakel, Serge Gutwirth, and Paul De Hert (Eds.). Hart Publishing, Oxford, 71–92.

[37] High-Level Expert Group on Artificial Intelligence. 2019. *Ethics Guidelines for Trustworthy AI.* Technical Report. European Commission.

[38] Samir Passi and Solon Barocas. 2019. Problem Formulation and Fairness. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT* '19)*. Association for Computing Machinery, Atlanta, GA, USA, 39–48. https://doi.org/10.1145/3287560.3287567

[39] Samir Passi and Steven Jackson. 2017. Data Vision: Learning to See Through Algorithmic Abstraction. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (CSCW '17)*. Association for Computing Machinery, Portland, Oregon, USA, 2436–2447. https://doi.org/10.1145/2998181.2998331

[40] Samir Passi and Steven J. Jackson. 2018. Trust in Data Science: Collaboration, Translation, and Accountability in Corporate Data Science Projects. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW (Nov. 2018), 1–28. https://doi.org/10.1145/3274405

[41] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. 2018. A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Comput. Surv.* 51, 5 (Sept. 2018), 92:1–92:36. https://doi.org/10.1145/3234150

[42] Inioluwa Deborah Raji and Joy Buolamwini. 2019. Actionable Auditing: Investigating the Impact of Publicly Naming Biased Performance Results of Commercial AI Products. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society (AIES '19)*. Association for Computing Machinery, 429–435. https://doi.org/10.1145/3306618.3314244

[43] Inioluwa Deborah Raji, Andrew Smart, Rebecca N. White, Margaret Mitchell, Timnit Gebru, Ben Hutchinson, Jamila Smith-Loud, Daniel Theron, and Parker Barnes. 2020. Closing the AI Accountability Gap: Defining an End-to-End Framework for Internal Algorithmic Auditing. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. ACM, Barcelona Spain, 33–44. https://doi.org/10.1145/3351095.3372873

[44] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco California USA, 1135–1144. https://doi.org/10.1145/2939672.2939778

[45] Niloufar Salehi, Lilly C. Irani, Michael S. Bernstein, Ali Alkhatib, Eva Ogbe, Kristy Milland, and Clickhappier. 2015. We Are Dynamo: Overcoming Stalling and Friction in Collective Action for Crowd Workers. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems - CHI '15*. ACM Press, Seoul, Republic of Korea, 1621–1630. https://doi.org/10.1145/2702123.2702508

[46] Morgan Klaus Scheuerman, Kandrea Wade, Caitlin Lustig, and Jed R Brubaker. 2020. How We've Taught Algorithms to See Identity: Constructing Race and Gender in Image Databases for Facial Analysis. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW1 (2020). https://doi.org/10.1145/3392866 Article 058.

[47] Ismaïla Seck, Khouloud Dahmane, Pierre Duthon, and Gaëlle Loosli. 2018. Baselines and a datasheet for the Cerema AWP dataset. In *Conférence d'Apprentissage CAp (Conférence d'Apprentissage Francophone 2018)*. Rouen, France. https://doi.org/10.13140/RG.2.2.36360.93448

[48] Dorothy E. Smith. 1990. *The conceptual practices of power: a feminist sociology of knowledge*. Northeastern University Press, Boston.

[49] Jennifer Wortman Vaughan and Hanna Wallach. 2020. A Human-Centered Agenda for Intelligible Machine Learning. In *Machines We Trust: Getting Along with Artificial Intelligence*. http://www.jennwv.com/papers/intel-chapter.pdf

A.4 Legal Framework for Regulating Algorithmic Decision-Making

My critical perspective on the obscurity of many machine learning models led to collaborations with legal scholar Ferdinand Müller and computer scientist Elsa Kirchner. In our joint work, we join the debate on whether Europe needs a regulatory body akin to a "KI-TÜV" to ensure the safe and responsible use of algorithmic decision systems. We suggest an EU-wide regulatory approach to manage the risks of these systems effectively to ensure safety and transparency while supporting technological progress and innovation. Concretely, we recommend jointly considering systemic and application-specific risks. To illustrate this, we present three examples of systemic risk from our research: biased models, black box predictors, and self-adopting systems. Based on this, we suggest regulating AI use on five ordinal risk levels.

Publication This work, written in German[1], has been published and presented at the 6th GRUR Conference. Three years after this publication, a risk-based approach of less granularity was adopted in Article 9 of the European Union's Artificial Intelligence Act [237].

> Ferdinand Müller, Elsa Kirchner, and Martin Schüßler. 2021. Ein „KI-TÜV" für Europa? Eckpunkte einer horizontalen Regulierung algorithmischer Entscheidungssysteme. In *GRUR Junge Wissenschaft Intelligente Systeme – Intelligentes Recht*. Vol. 2020/21. Nomos Verlagsgesellschaft, 85–106. isbn: 978-3-8487-8142-3. `https://www.nomos-shop.de/nomos/titel/intelligente-systeme-intelligentes-recht-id-99401/`

[1] An English summary is available at `https://www.weizenbaum-institut.de/en/news/detail/ein-ki-tuev-fuer-europa0/`.

Ein „KI-TÜV“ für Europa? Eckpunkte einer horizontalen Regulierung Algorithmischer Entscheidungssysteme

Ferdinand Müller, Dr. Elsa Kirchner, Martin Schüßler[*]

Die Diskussion um einen Rechtsrahmen für KI-Anwendungen läuft auf Hochtouren. Allerorts werden Strategien, Leitlinien und Empfehlungen veröffentlicht. Doch was macht die Technik in ihrem Wesenskern aus? Was ist ihr spezielles Risiko? In diesem Beitrag soll erörtert werden, wie eine innovationsoffene, aber dennoch Sicherheit schaffende Bestimmung gefunden werden kann.

A. KI-Technik auf dem Prüfstand

I. Technologie des täglichen Lebens

Automatisierte und teilautomatisierte *Algorithmische Entscheidungssysteme* (AES) werden mit steigender Tendenz zu einer Technik, die uns ständig umgibt und mit der wir auch im Alltag immer häufiger interagieren. Neben dem Einsatz der Systeme durch Unternehmen kommen diese zunehmend auch bei Hoheitsträgern zur Anwendung.[1] Dabei können sie in bestimmten Lebensbereichen enorme Auswirkungen haben.

Ob dabei auch neue Risiken für immaterielle oder materielle Rechtsgüter entstehen, die eine Regulierung erforderlich machen, ist Gegenstand zahlreicher aktueller Debatten.[2] Eine repräsentative Umfrage des TÜV-Ver-

---

* Ferdinand Müller arbeitet am interdisziplinären *Weizenbaum-Institut für die vernetzte Gesellschaft* und beschäftigt sich mit den Grundlagen künstlicher Intelligenz im Rechtsverkehr. Martin Schüßler, ebenfalls vom *Weizenbaum*, ist Informatiker und Mensch-Maschine-Interaktions-Forscher im Bereich *explainable AI*. Dr. Elsa Kirchner vom *Deutschen Forschungszentrum für Künstliche Intelligenz* (DFKI) ist Biologin und hat in der Informatik im Bereich der Mensch-Maschine-Interaktion promoviert. Sie arbeitet an Methoden des *Interactive Machine Learning*.

1 Überblick über laufende Projekte von AlgorithmWatch: “Automating Society”, abrufbar unter: https://kurzelinks.de/jl15.

2 *L. Käde/S. von Maltzan*, Die Erklärbarkeit von Künstlicher Intelligenz (KI), CR 2020, S. 66 ff.; *A. Allar*, Rechtliche Herausforderungen Künstlicher Intelligenz, ZUM 2020, S. 325 ff.; *H. Zech*, Risiken Digitaler Systeme: Robotik, Lernfähigkeit

bandes in Deutschland kam zu dem Schluss, dass eine Mehrheit der Befragten die Einrichtung einer besonderen Prüfstelle und eine damit einhergehende staatliche Regulierung befürwortet.[3] Brauchen wir einen KI-TÜV? Was könnten die Eckpunkte einer auf AES bezogenen "Hauptuntersuchung" sein? Dieser Beitrag versucht die Besonderheiten dieser Systeme herauszuarbeiten, ihre Risiken zu bestimmen und schließlich die Eckpunkte einer geeigneten Regulierung vorzustellen.

II. Eine Definition

Über die technischen Eigenschaften und die daraus erwachsenden rechtlichen Risiken von AES treten immer wieder Unklarheiten auf. Dies resultiert auch aus der übermäßigen Verwendung des Schlagwortes "Künstliche Intelligenz" und den mit diesem Begriff verbundenen, teilweise stark überzogenen Erwartungen.[4] Die Bezeichnung entstand in einer Zeit, in der man von den Hürden der Erschaffung eines "maschinellen Denkens" noch keine konkrete Vorstellung hatte und man davon ausging, dass letztlich jede Tätigkeit von einem automatisierten Computer lösbar sein müsste.[5] Dass dies zu bewerkstelligen ungleich schwieriger ist, als ursprünglich angenommen, zeigt sich nicht zuletzt an der ambivalenten Entwicklung, wel-

---

und Vernetzung als aktuelle Herausforderungen für das Recht, Weizenbaum Insights 2020, abrufbar unter: https://kurzelinks.de/i2sq; *B. Jakl*, Das Recht der Künstlichen Intelligenz – Möglichkeiten und Grenzen zivilrechtlicher Regulierung, MMR 2019, S. 711 ff.; *S. Meyer*, Künstliche Intelligenz und die Rolle des Rechts für Innovation, ZRP 2018, S. 233 ff.; *G. Borges*, Rechtliche Rahmenbedingungen für autonome Systeme, NJW 2018, S. 977 ff.; *T. Burri*, Künstliche Intelligenz und internationales Recht, DuD 2018, S. 603 ff.; *M. Herberger*, "Künstliche Intelligenz" und Recht, NJW 2018, S. 2825 ff.; *M. Martini*, Algorithmen als Herausforderung für die Rechtsordnung, JZ 2017, S. 1017 ff.; zur Diskussion in USA siehe *A. Tutt*, An FDA For Algorithms, Administrative Law Review 69 (2017), S. 83 ff.

3 "Sicherheit und Künstliche Intelligenz", Studie des VdTÜV vom Januar 2020, abrufbar unter: https://kurzelinks.de/fm13.

4 Zu den schwierigen begrifflichen Fahrwassern siehe *Herberger*, Künstliche Intelligenz (Fn. 2), S. 2825 (2826 f.) oder auch *N. Braun Binder*, Künstliche Intelligenz und automatisierte Entscheidungen in der öffentlichen Verwaltung, SJZ 2019, S. 467 (469); vgl. dazu auch *G. Marcus/E. Davis*, Rebooting AI: Building Artificial Intelligence We Can Trust, New York: Pantheon 2019, S. 4 f.

5 *J. McCarthy/M. L. Minsky/N. Rochester/E. Shannon*, A Proposal For The Darthmouth Summer Research Project On Artificial Intelligence, 31.08.1955, abrufbar unter: https://kurzelinks.de/zfnp.

che von teilweise jahrelangen Stagnationen gekennzeichnet ist.[6] Mit der Zeit gab man (vorerst) die Utopie einer "starken" künstlichen Intelligenz, die umfassend auf ihre Umgebung einwirken kann, zugunsten der Entwicklung anwendungsorientierter "schwacher" KI-Technologien auf.[7] Hier zeichnen sich seit einigen Jahren deutliche Fortschritte ab, die neben den Steigerungen in der verwendeten Computertechnik vor allem auch auf die massive Nutzung personenbezogener Daten rückführbar sind.[8]

Algorithmische Entscheidungssysteme sind dabei als eine Ausprägung schwacher KI-Technologie einzuordnen.[9] Die zugrundeliegende Idee ist hier, dass durch das System eine Entscheidung oder eine Entscheidungsempfehlung erstellt wird, welche durch die maschinelle bzw. algorithmische Auswertung einer bestimmten Datenmenge schnellere und fundiertere Annahmen ermöglichen soll als eine allein händisch getroffene Entscheidung. AES basieren grundsätzlich auf denselben technischen Prämissen, die für die gesamte elektronische Datenverarbeitung gelten: digitale Algorithmen (1) verarbeiten digital und/oder analog (bspw. durch Sensoren) aufgenommene (2) maschinell lesbare Informationen (Daten[10]) (3), was zu einem sich digital oder analog (bspw. Bewegung mittels Aktuatoren) manifestierenden Ergebnis führt (4). Mehrere Besonderheiten unterscheiden AES jedoch von herkömmlicher Computertechnik: Die bei ihnen verwendeten Modelle setzen sich, je nach Komplexität des Anwendungsfalls, aus einer umfangreichen Datengrundlage (vgl. "Big Data") zusammen (5). Ihre Modellierung erfolgt durch algorithmische Optimierungsverfahren und unter Einbringung spezifischen Domänenwissens (6). Des Weiteren agieren diese Systeme bei der Informationsgewinnung, -verarbei-

6 Vgl. die Situation im sog. "KI-Winter", Begriff geprägt von *D. Crevier*, AI: The Tumultuous History Of The Search For Artificial Intelligence, New York, Basic Books 1993.

7 Vgl. Darstellung in der KI-Strategie der Bundesregierung vom November 2018, abrufbar unter: https://kurzelinks.de/h00f, S. 4 f.

8 *M. Martini*, Blackbox Algorithmus – Grundfragen einer Regulierung künstlicher Intelligenz, Berlin/Heidelberg 2019, S. 14.

9 Nach dem englischen Begriff *algorithmic decision making systems*, bspw. verwendet bei Unterstanding algorithmic decision-making: Opportunities and challenges, Gutachten des Wissenschaftlichen Dienstes des Europäischen Parlaments, vom März 2019, abrufbar unter: https://kurzelinks.de/z43t; im Deutschen verwendet bei *K. A. Zweig*, Algorithmische Entscheidungen: Transparenz und Kontrolle, Gutachten der KAS von 01/19, abrufbar unter: https://kurzelinks.de/l42v.

10 Zu den verschiedenen Ebenen des Datenbegriffs siehe *H. Zech*, "Industrie 4.0" – Rechtsrahmen für eine Datenwirtschaft im digitalen Binnenmarkt, GRUR 2015, S. 1151 (1153 f.).

tung und -ausgabe in einem höheren Maße selbstständig (7), so dass durch die Entkoppelung von menschlicher Mithilfe auch die Bearbeitung äußerst großer Datenmengen in relativ kurzer Zeit möglich wird. Weiterhin werden bei bestimmten AES zur Informationsverarbeitung sehr komplexe Modelle eingesetzt, wie bspw. künstliche neuronale Netze (8), für welche biologische neuronale Netze (wie bspw. im Gehirn) als Inspiration dienten.[11] Die Informationsverarbeitung erfolgt hier mittels gewichteter und häufig nicht-linearer Signalweitergabe entlang zahlreicher Entscheidungsschichten. Dies ermöglicht die Modellierung komplexerer Sachverhalte, stellt aber auch eine große Hürde für die Nachvollziehbarkeit dieser Modelle dar. Manche AES sind zudem nicht statisch, sondern passen die eigenen Parameter im Laufe des Einsatzes und/oder im Vorfeld an (9).

In den meisten Fällen ist ein Mensch als Supervisor und Letztentscheider Bestandteil des Systems, der die Entscheidungsempfehlung formuliert und/oder umsetzt, so dass von *teilautomatisierten* AES gesprochen werden sollte (10). Systeme, die in hohem Maße selbständig agieren und dabei gänzlich ohne menschlichen Operator für ihren Kernbetriebsablauf auskommen, kann man hingegen als *(voll-)automatisierte* AES beschreiben (11). Letztere werden bspw. im börslichen Hochfrequenzhandel eingesetzt, wo sie zur selbstständigen Konzeption, Formulierung und Ausführung von tausenden Orders pro Sekunde fähig sind.[12]

Die technischen Spezifikationen ermöglichen einerseits, dass der Einsatz von AES bei dafür geeigneten Tätigkeiten zu einer enormen Effizienzsteigerung führen kann. Dies sind vor allem Bereiche, bei denen die Entscheidungsgrundlage aus einer erheblichen Datenmenge besteht oder wo Entscheidungen in Sekundenbruchteilen notwendig sind.[13] Andererseits führen diese Eigenschaften auch zu neuartigen Risiken, bei denen fraglich ist, ob sie durch die bestehende Rechtsmaterie abgedeckt werden. Diese Risiken können sich aus den Eigenschaften des Systems selbst ergeben, aber auch aus ihrer (sektor-)spezifischen Anwendung.[14] Aus der Strukturierung der Risiken (Teil B.) und aus ihrer gegenseitigen Koppelung (Teil C.) ergeben sich letztlich die für AES entscheidenden regulatorischen Implikationen, welche am Ende des Beitrages behandelt werden sollen.

---

11 *Zech*, Risiken (Fn. 2), S. 14 ff.

12 Zu den Grundlagen des algorithmenbasierten Wertpapierhandels siehe *H.-P. Kollmann*, Autonome und intelligente Wertpapierhandelssysteme, Tübingen 2019, S. 52 ff.

13 *Martini*, Algorithmus (Fn. 8), S. 13.

14 Vgl. *Martini*, Algorithmus (Fn. 8), S. 115, "doppelte Ungewissheit" beim Einsatz von AES.

III. Suche nach einem Prüfprogramm

Fast jedes große Industrieland hat mittlerweile eine eigene Strategie zum Umgang mit der neuartigen Technik vorgelegt, wo oft betont wird, einen geeigneten Ansatz für die Regulierung finden zu wollen.[15] Während viele dieser Papiere lediglich Analysen eines Regulierungsbedarfes angekündigt haben, hat man bspw. in den USA auf lokaler Ebene schon zu entsprechenden Maßnahmen gegriffen.[16]

In Europa ist insbesondere die EU der Treiber einer Regulierung. Ein gesamteuropäischer Ansatz erscheint am sinnvollsten, da mögliche Regulierungsvorhaben auf nationalstaatlicher Ebene in den Binnenmarkt eingreifen könnten.[17] Die Kommission kündigte im April 2018 an, über eine Erweiterung des bisher geltenden Rechtsrahmens nachzudenken.[18] In den Mitteilungen vom Dezember 2018[19] und April 2019[20] stellte sie ihr Konzept vor, die Entwicklung und Anwendung von AES innerhalb des europäischen Wertekanons stattfinden zu lassen ("menschenzentrierte KI"). Im Rahmen ihrer Initiative hatte die Kommission auch die sog. *High Level Expert Group* eingerichtet, die aus Sachverständigen der Wirtschaft und Wis-

---

15 Bspw. USA, vom 11.02.19, abrufbar unter: https://kurzelinks.de/9n5q; China, vom 08.07.2017, abrufbar unter https://kurzelinks.de/of2z, Frankreich, vom 29.03.2018, abrufbar unter https://kurzelinks.de/992h und Deutschland, vom Dezember 2018, abrufbar unter: https://kurzelinks.de/mxy4; für einen Überblick siehe *K. Walch*, "AI Laws Are Coming", Forbes vom 20.02.20, abrufbar unter: https://kurzelinks.de/xjx0.

16 Zum Gesetzesvorhaben der Stadt New York siehe *Martini*, Algorithmus (Fn. 8), S. 87; für die Ergebnisse der Arbeitsgruppe siehe https://kurzelinks.de/le1t; die Ergebnisse wurden jedoch gespalten aufgenommen, siehe *C. Lecher*, "NYC's algorithm task force was 'a waste'", The Verge vom 20.11.2019, abrufbar unter: https://kurzelinks.de/ogry.

17 So auch Gutachten der Datenethikkommission, vom 23.10.19, die den Erlass einer "EU-Verordnung für Algorithmische Systeme" empfiehlt, abrufbar unter: https://kurzelinks.de/bmbt, S. 181 ff.; wohl auch Auffassung der Kommission; vgl. "Weißbuch zur Künstlichen Intelligenz", vom 19.02.2020, abrufbar unter: https://kurzelinks.de/e5uu, S. 17; ebenfalls befürwortend *R. H. Weber/S. Henseler*, Regulierung von Algorithmen in der EU und in der Schweiz, EuZ 2020, S. 28 (42).

18 Europäische Kommission, Mitteilung vom 25.04.2018, abrufbar unter: https://kurzelinks.de/lwm6, S. 17 f.

19 Europäische Kommission, Mitteilung vom 07.12.2018, abrufbar unter: https://kurzelinks.de/dlo3.

20 Europäische Kommission, Mitteilung vom 08.04.2019, abrufbar unter: https://kurzelinks.de/li5d.

senschaft besteht.[21] Die Gruppe hat bis dato zwei Dokumente vorgelegt: zum einen die "Ethik-Leitlinien für eine vertrauenswürdige KI", in welchen die sich aus der Grundrechtecharta ergebenden Rahmenbedingungen einer ethischen KI-Technologie dargelegt wurden.[22] Konkrete rechtliche Vorschläge enthalten die "Leitlinien" jedoch nicht; auch wurde kritisiert, dass die Aufteilung in sich teilweise wiederholende und unstrukturierte Einzelanforderungen verwirrend sei.[23] Zum anderen wurden im Juni 2019 die "Policy and Investment Recommendations" veröffentlicht, in der auch auf eine Überprüfung des "legal framework" hingewiesen wird.[24] Im Mitte Februar veröffentlichten "KI-Weißbuch" hat die Kommission die Ergebnisse der bisherigen Bestrebungen zusammengefasst und zu Anmerkungen für mögliche Veränderungen des Rechtsrahmens zu einer bis Ende Mai laufenden öffentlichen Konsultation aufgerufen.[25] Kritisiert wurde, dass kaum handfeste Maßnahmen im Weißbuch enthalten sind und sich die aufgestellte Risikobewertung als vage und konturlos erweist.[26]

Was wohl alle bisher vorgestellten Gutachten, Mitteilungen und Strategien vereint, ist das Streben nach einer strukturierten Risikobewertung für AES, bei der die Kritikalität von damit operierenden Anwendungen anhand eines übergreifenden Modells bestimmt werden kann.[27] Ein mögliches Beispiel eines solchen Modells soll am Ende dieses Beitrages erörtert werden (Teil C.). Zu erwähnen für Deutschland ist noch, dass auch die Entwicklung von industriellen Standards[28] und Zertifizierungen[29] in Angriff genommen wird.

---

21 Zur Agenda und Zusammensetzung siehe Website der Europäische Kommission, abrufbar unter: https://kurzelinks.de/50ro.

22 Dokument abrufbar unter: https://kurzelinks.de/l5me.

23 *H.-U. Dettling/S. Krüger*, Erste Schritte im Recht der Künstlichen Intelligenz: Entwurf der "Ethik-Leitlinien für eine vertrauenswürdige KI", MMR 2019, S. 211 (213 f.).

24 Dokument abrufbar unter: https://kurzelinks.de/aiwo.

25 Dokument abrufbar unter: https://kurzelinks.de/e5uu, die Ergebnisse der Konsultation lagen zum Zeitpunkt der Erstellung des Beitrags noch nicht vor.

26 *Y. Borutta/M. Haag/H. Hoffmann/J. Kevekordes/V. Vogt*, "Fundamentalkritik" des White Papers, abrufbar unter: https://kurzelinks.de/nhi7, S. 3 f.

27 Gutachten der Datenethikkommission (Fn. 17), S. 177; Weißbuch (Fn. 17), S. 19 f.; Policy and Investment Recommendations (Fn. 24), S. 37 f.

28 Projektübersicht des DIN, abrufbar unter: https://kurzelinks.de/sw2m.

29 *J. Heesen/J. Müller-Quade/S. Wrobel*, "Zertifizierung von KI-Systemen – Impulspapier", Plattform Lernende Systeme, abrufbar unter: https://kurzelinks.de/w2fm, S. 7 ff. mit Aufzählung weiterer Projekte.

IV. DSGVO als Lösung?

Von einigen Beteiligten des Diskurses wird vorgetragen, eine entsprechende Anwendung der DSGVO würde für die Handhabung zumindest mancher der mit AES verbundenen Risiken ausreichen. Es geht dabei unter anderem um die Frage, ob sich aus der DSGVO bzw. aus dem Auskunftsrecht in Art. 15 Abs. 1 lit. h, Art. 22 DSGVO Anforderungen an die Transparenz von AES herleiten lassen.[30] Aus dem "Recht auf Erklärbarkeit" ("right to explanation") datenverarbeitender Prozesse soll die Erklärbarkeit des ganzen Systems folgen.

Gegen diesen "Umweg" der Regulierung von AES können mehrere Einwände erhoben werden: Zum einen richtet sich die DSGVO an alle möglichen Formen der Datenverarbeitung – sei es nun händisch mit "Zettel und Stift" oder komplett digital mittels eines AES. Die DSGVO will damit kein bestimmtes Technologierisiko regulieren, sondern konzentriert sich technologieneutral nur auf den Teil des Prozesses, in dem personenbezogene Daten verarbeitet werden.[31] Nach ihrem Regelungskonzept erfasst die DSGVO also nur die Risiken, die für die persönlichkeitskeitsrechtliche Ebene der Datenverarbeitung und die damit verbundenen Rechte bestehen und nicht auch alle anderen Rechtsgüter, die durch den Einsatz von AES gefährdet sein können.[32] So kann es möglich sein, dass ein AES datenschutzrechtlich auf keine Bedenken stößt, aber andere Rechtsgüter in unzulässiger Weise beeinträchtigt. Nimmt man etwa den aktuellen Anwendungsbereich von Art. 22 DSGVO ernst, kann die Einhaltung der Vorschrift sogar zur Gefährdung von Rechtsgütern aktiv beitragen: Droht etwa ein autonomes Fahrzeug mit einer Person zu kollidieren, dürfte es nicht die personenbezogenen (Standort-)Daten desjenigen verarbeiten, dem es ausweichen möchte.[33]

---

30 Siehe *S. Wachter/B. Mittelstadt/C. Russell*, Counterfactual Explanations Without Opening The Black Box: Automated Decisions And The GDPR, Harvard Journal of Law & Technology 31 (2018), S. 841 (861 ff.).; oder auch *L. Franck*, in: P. Gola (Hrsg.), Datenschutz-Grundverordnung, 2. Auflage, München 2018, Art. 15 Rn. 19; *L. K. Kumkar/D. Roth-Isigkeit*, Erklärungspflichten bei automatisierten Datenverarbeitungen nach der DSGVO, JZ 2020, S. 277 ff. jeweils m.w.N.

31 Bspw. Erwägungsgrund 15 DSGVO.

32 So auch *Martini*, Algorithmus (Fn. 8), S. 80 f.; zu den im Zusammenhang mit dem Schutz personenbezogener Daten betroffenen Grundrechte s. *S. Ernst*, in: B. P. Paal/D. A. Pauly (Hrsg.), DSGVO/BDSG, 2. Auflage, München 2018, Art. 1 Rn. 11.

33 *M. Kroker*, Art. 22 DSGVO – ein Schuss in den Ofen?, PinG 2020, S. 255 ff. (257).

Zum anderen bezieht sich Art. 22 DSGVO wenn dann nur auf die Datenverarbeitung *vollautomatisierter* AES, wobei *teilautomatisierte* AES von der Anwendung der Vorschrift ausgenommen sein sollen.[34] Wie weiter oben schon dargestellt, fallen deshalb die Mehrzahl an AES, welche auf einer Kombination von maschineller Entscheidungsempfehlung und menschlicher Umsetzung beruhen, aus dem Anwendungsbereich der Vorschrift heraus. Dieser Umstand und generell der starre Verbotscharakter der Norm lassen auch für die DSGVO selbst an dieser Stelle gesetzgeberischen Anpassungsbedarf erkennen.[35] Die Reichweite des oben angesprochenen Auskunftsrechts – also welche Inhalte der involvierten Verarbeitungslogiken dem "right to explanation" unterliegen sollen – ist auch nach der "Schufa"-Rechtsprechung des BGH nur mit Unsicherheiten bestimmbar.[36] § 31 BDSG hilft hier für die Regulierung von AES-spezifischen Risiken letztlich genauso wenig weiter; dem Gesetzgeber ging es wohl nur darum, Scoring von Wirtschaftskarteien trotz Datenschutzes zu ermöglichen.[37] Auch sollte man angesichts einer zurzeit eher europarechtskritischen Rechtsprechung von einer extensiven Auslegung europäischer Normen absehen.[38]

Gelegentlich wird auch hervorgebracht, dass bereits die Datenschutzgrundsätze in Art. 5DSGVO einen operationalisierbaren Anforderungskatalog an die technischen Gegebenheiten von AES bereitstellen würden.[39] Ein genauer Vollzug der Grundsätze oder auch eine Sanktionierung bei deren Nichteinhaltung gestaltet sich aber aufgrund des bestimmungs- und

34 Siehe Erwägungsgrund 71 DSGVO; siehe *S. Schulz*, in: P. Gola (Hrsg.), DSGVO, 2. Auflage, München 2018, Art. 22 Rn. 1; *R. B. Abel*, Automatisierte Entscheidungen im Einzelfall, ZD 2018, S. 304 (305).

35 *M. Martini,* in: B. P. Paal/D. A. Pauly (Hrsg.), DSGVO/BDSG, 2. Auflage, München 2018, Art. 22 Rn. 46; auch kann daran gezweifelt werden, ob eine aus den 1970er Jahren stammende Vorschrift zur Regelung moderner KI-Technik noch passend ist, vgl. *S. Golla*, In Würde vor Ampel und Algorithmus, DÖV, S. 673 ff. (679 f.).

36 *F. Schmidt-Wudy*, in: H. A. Wolff/S. Brink (Hrsg.), BeckOK: Datenschutzrecht, 31. Edition, München 2020, Stand 01.05.20, Art. 15 DSGVO Rn. 78.3; vgl. *W. Krämer*, Die Rechtmäßigkeit der Nutzung von Scorewerten, NJW 2020, S. 497 ff.; siehe auch "OpenSCHUFA"-Kampagne der Open Knowledge Foundation Deutschland und AlgorithmWatch, abrufbar unter: https://openschufa.de/.

37 *E. M. Frenzel*, in: B. P. Paal/D. A. Pauly (Hrsg.), DSGVO/BDSG, 2. Auflage, München 2018, § 31 BDSG Rn. 10, der am Überblick des Gesetzgebers zweifelt.

38 BVerfG NJW 2020, S. 1647 ff.

39 Vgl. *M. Rost,* Künstliche Intelligenz – Normative und operative Anforderungen des Datenschutzes, DuD 2018, S. 558 ff. (559).

abwägungsbedürftigen Inhalts dieser Norm eher als schwierig.[40] Zusammenfassend lässt sich damit sagen, dass eine ausdrückliche Neuregelung für das spezielle AES-Risiko auch trotz Geltung der DSGVO erforderlich sein wird.[41]

B. Risikobereiche von AES

I. Systembezogene Risiken

Nachfolgend werden systembezogene Risiken aufgeführt, die anwendungsunabhängig auftreten können. Diese betreffen die Komponenten des AES – bspw. das Modell, den verwendeten Algorithmus, die (Trainings-)Daten, die operationale Datenbasis oder den menschlichen Umgang mit dem System. Die folgende Darstellung versteht sich als ein möglicher Annäherungsversuch an die Risiken dieser Systeme. Es existieren bereits andere Modelle, die die Risikofelder verschieden interpretieren und abgrenzen.[42] Dieser Beitrag plädiert für die generelle Berücksichtigung von drei systembezogenen Risikobereichen anwendungsunabhängig bei AES zu berücksichtigen sind: die (unbewussten) Verzerrungen von Parametern ("Biased AI"), die Intransparenz der Systemarchitektur ("Black Box AI") und die nachträgliche Veränderung von AES durch maschinelles Lernen.

1. Unbewusste Verzerrungen – "Biased AI"

Als Begründung für die Einführung von AES wird oftmals deren Objektivität und Neutralität hervorgehoben. Die Entscheidungen der Maschine seien unbeeinflussbar und daher nicht korrumpierbar, die "nackten Zahlen scheinen frei von jedem ideologischen Makel".[43] Die Gefährlichkeit solcher Annahmen beschrieb als einer der Ersten *Joseph Weizenbaum*, ein Pionier der kritischen Informatik. Gleich dem alten Witz über den Betrun-

40 *A. Roßnagel*, in: Simitis/Hornung/Spiecker gen. Döhmann (Hrsg.), Datenschutzrecht, 1. Auflage, Baden-Baden 2019, Art. 5 Rn.: 22.

41 Auch *Martini*, Algorithmus (Fn. 8), S. 339.

42 Vgl. etwa *Zech*, Risiken (Fn. 2), ("Robotik", "Lernfähigkeit" und "Vernetzung") oder *G. Teubner*, Digitale Rechtssubjekte, AcP 2018, S. 155 ff. ("Autonomierisiko", "Verbundrisiko", "Vernetzungsrisiko").

43 *D. L. Burk*, Algorithmic Legal Metrics, Notre Dame Law Review, Forthcoming, abrufbar unter: https://kurzelinks.de/ywz8, S. 14.

kenen, der nur im Lichtkegel der Laterne seine verlorenen Schlüssel sucht, weil es "dort heller ist", beschränke die gewählte Herangehensweise an ein Problem immer die dafür entwickelte Lösung. Die "Macht der Computer" sei ein gutes Beispiel für das Risiko, welches "allen […] sich selbst bestätigenden Denksystemen innewohnt".[44]

Diese Ambivalenz zeigt bspw. die Kontroverse um die von der US-Justiz eingesetzten AES auf. In vielen Bundesstaaten werden mittlerweile in bestimmten Phasen des Strafverfahrens AES benutzt, um die Rückfallwahrscheinlichkeit mutmaßlicher Straftäter*Innen für die Festlegung der Bewährung oder Kaution zu berechnen. Bereits 2014 hatte der Generalbundesanwalt der USA die Vermutung geäußert, dass der Einsatz solcher Systeme aufgrund ihrer Berechnungsgrundlage, welche u.a. auf verschiedenen statischen Faktoren über den Bildungsstand, sozioökonomischen Hintergrund und die Nachbarschaft des Angeklagten beruht, eher zur Verschärfung von Ungleichbehandlungen führen könnte.[45] Tatsächlich zeigte eine Untersuchung in 2016, dass die Wahrscheinlichkeit eines "false positives" – also eine fälschlicherweise zu hoch angenommene Rückfallwahrscheinlichkeit – für People of Colour signifikant höher sein kann.[46] Der Einsatz solcher Systeme – gerade für hoheitliche Tätigkeiten – würde damit zu nicht hinnehmbaren Diskriminierungen führen. Im Fall der auch zu diesen Zwecken eingesetzten Software *COMPAS* konnten Wissenschaftler*Innen 2018 zeigen, dass die Bewertungen des Systems kaum besser waren, als die von Testpersonen ohne juristische Vorkenntnisse.[47] Des Weiteren gelang es ihnen, ein vollständig interpretierbares Modell unter Nutzung von nur 7 statt 137 (!) Parametern zu gestalten, welches sogar eine leicht verbesserte Genauigkeit erzielte.

Zwar entsteht das eigentliche Problem – die vorurteilsbehaftete Einstufung eines Menschen aufgrund zugeschriebener Eigenschaften – nicht erst durch den Einsatz von AES. Es setzt sich aber in der Technik fort und kann sich dabei verstärken.[48] Die sich aufdrängenden Fragen, die sich

---

44 *J. Weizenbaum*, Die Macht der Computer und die Ohnmacht der Vernunft, 14. Auflage, Berlin 1978, S. 178 ff.

45 *E. Holder*, Rede am 01.08.2014, abrufbar unter: https://kurzelinks.de/2xme.

46 *J. Angwin/J. Larson/S. Mattu/L. Kirchner*, "Machine Bias", ProPublica vom 23.06.2016, abrufbar unter: https://kurzelinks.de/kw4p; siehe auch "Der Algorithmus ist Rassist", Beitrag von Spiegel.de vom 09.09.2016, abrufbar unter: https://kurzelinks.de/8zz0.

47 *J. Dressel/H. Farid*, The accuracy, fairness and limits of predicting recidivism, Science Advances 4 (2018), abrufbar unter: https://kurzelinks.de/e0ap.

48 Über die Probleme der Quantifizierung sozialer Faktoren siehe *Burk,* Metrics (Fn. 43), S. 6 f.

beim nachträglichen Bekanntwerden solcher "Biases" stellen, sind in der Praxis aktuell schwierig bis teilweise gar nicht zu beantworten: Wo befindet sich die Verzerrung genau? Ist bereits die Abbildung des zu lösenden realen Problems auf ein informationstechnisches Modell verzerrt? Oder entsteht die Verzerrung erst durch das Training bzw. die verwendeten Trainingsdaten oder dann später im Einsatz?

Die Wissenschaft steht hier bei vielem noch am Anfang.[49] Das "AI Now Institute", welches sich der Bekämpfung des Phänomens verschrieben hat, empfiehlt u. a. die Herstellung der Transparenz der entsprechenden Systeme, rigorose Testläufe während des gesamten Einsatzes von AES und fortwährende Beobachtung auf diskriminierende Faktoren, die interdisziplinäre Einbeziehung sozialwissenschaftlicher Erkenntnisse über den Kontext des Einsatzes und eine ausführliche Risikobewertung im Vorfeld.[50]

2. Intransparenz maschineller Prozesse – "Blackbox AI"

Mit dem Begriff "Blackbox AI" bezeichnet man ein AES, dessen Ergebnisausgabe sich nicht durch einen Menschen nachvollziehen lässt. Zu unterscheiden ist hierbei zwischen zwei weit verbreiteten Ursachen.

a. Intransparenz aufgrund technischer Ursachen

Die erste mögliche Ursache ist technischer Natur. Es gibt AES, die Modelle verwenden, welche sich aufgrund ihrer internen Komplexität der Interpretierbarkeit durch den Menschen entziehen. Bspw. können im Falle von künstlichen neuronalen Netzen Parameter in unüberschaubarer, nicht linearer Weise zusammenwirken, was ein großes Hindernis für die Verständlichkeit darstellen kann. Des Weiteren kann schon alleine die pure Anzahl der in einem Modell verwendeten Parametern so groß sein, dass die Vorhersagbarkeit von Ergebnissen nicht mehr gegeben ist. Als aktuelles Beispiel sei hier das Turing-NLG Sprachmodell von Microsoft genannt,

---

49 Für eine Übersicht siehe *A. Olteanu/C. Castillo/F. Diaz/E. Kıcıman*, "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries", Front. Big Data, Beitrag vom 11.07.2019, abrufbar unter: https://kurzelinks.de/axeg.

50 *S. M. West/ M. Whittaker/K. Crawford*, Discriminating Systems – Gender, Race, and Power in AI, AI Now Institute, April 2019, abrufbar unter: https://kurzelinks.de/u5ur.

welches aus 17 Milliarden Parametern besteht.[51] Fortschritte bei der informationstechnischen Lösung von besonders komplexen Problemstellungen (wie der menschlichen Sprache oder dem maschinellen Sehen) gehen aktuell fast immer mit einer enormen Steigerung technischer Komplexität einher, wenngleich auch Ansätze existieren, die versuchen, zumindest die Anzahl der Parameter so gering wie möglich zu halten.[52]

Zumindest für manche Anwendungsszenarien von AES besteht die Möglichkeit, interpretierbare Modelle einzusetzen. Deren Einsatz ist aber zum Teil mit erheblichem Mehraufwand verbunden. Für zahlreiche andere Szenarien gibt es nach aktuellem Stand kaum interpretierbare Modelle, deren Einsatz nicht einen erheblichen Genauigkeits- und somit auch Effizienzverlust mit sich bringen würde. Aus diesem Grund ist der Einsatz von hochkomplexen und nicht nachvollziehbaren Modellen weit verbreitet.

Trotz umfangreicher Forschung im Bereich "Interpretable Machine Learning" und "Explainable Artifical Intelligence" (XAI) gibt es zum Zeitpunkt der Veröffentlichung dieses Beitrages kaum Ansätze, die die Verständlichkeit solcher Modelle im Einsatz zuverlässig und in empirisch nachweisbarer Form steigern würden. Hier steckt die Forschung noch in den Anfängen. Erschwerend kommt hinzu, dass viele der vorgestellten Ansätze keine modellgetreuen und meist auch unvollständige Erklärungen generieren.[53] Folglich besteht keine Sicherheit, dass die Erklärungen ein Modell wirklich wahrheitsgetreu beschreiben.

b. Intransparenz aufgrund wirtschaftlich-organisatorischer Ursachen

Die zweite Ursache für Intransparenz ist wirtschaftlicher Natur. AES-Modelle enthalten nicht selten Informationen, die Gegenstand von Urheber- oder Schutzrechten nach der Geschäftsgeheimnisrichtlinie sind.[54] Auch im

---

51 *C. Rosset*, "Turing-NLG: A 17-billion-parameter language model by Microsoft", Microsoft Research Blog vom 13.02.2020, abrufbar unter: https://kurzelinks.de/1cdq.

52 Vgl. *J. Ba/R. Caruana*, Do Deep Nets Really Need to Be Deep?, Advances in Neural Information Processing Systems 27 (2014), abrufbar unter: https://kurzelinks.de/ex85.

53 *C. Rudin*, Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead, Nature Machine Intelligence 1 (2019), S. 206, abrufbar unter: https://doi.org/10.1038/s42256-019-0048-x.

54 Vgl. *S. Hetmank/A. Lauber-Rönsberg*, Künstliche Intelligenz – Herausforderungen für das Immaterialgüterrecht, GRUR 2018, S. 574 (575); *A. Rosenkötter/S. Seeger*, Das neue Geschäftsgeheimnisgesetz, NZBau 2019, S. 619 ff.

Schufa-Urteil wurde argumentiert, dass eine Offenlegung der Berechnungsformel schützenswerte Interessen der Auskunftei verletzen würde.[55] Das Wissen über die Modellierung und Konstruktion eines AES ist von enormem wirtschaftlichem Wert. Ein erklärbares Modell kann potentiell von der Konkurrenz nachgebaut werden. Auch aus diesem Grund werden zahlreiche AES (bspw. von Google, Amazon, Microsoft und IBM) ausschließlich als Dienstleistung über eine Schnittstelle angeboten, die keine weiteren Informationen über die dahinterliegenden Algorithmen und Modelle preisgeben. Für Unternehmen kann es daher lohnend sein, sich hinter dem Argument der technisch geschuldeten Intransparenz zu verstecken, um größeren wirtschaftlichen Erfolg zu erzielen. Es bedarf deshalb einer genauen Prüfung, ob für einen vorgesehenen Anwendungsfall wirklich ein Modell hoher Komplexität und darauffolgender technisch begründeter Intransparenz zum Einsatz kommen muss, oder ob nicht ein transparenteres Modell geschaffen werden kann.[56]

3. Veränderungsfähigkeit

Das dritte systembezogene Risiko stellt die Veränderungsfähigkeit von AES dar. Durch einen sich permanent wiederholenden Prozess kann erreicht werden, dass eine Abbildung, also ein Modell eines Prozesses, einer Datenlage oder auch kognitiver Phänomene auf Maschinen sich Schritt für Schritt verbessert.[57] Dieser iterative Prozess wird – als ein Teilbereich von KI – mit dem Begriff des *maschinellen Lernens* bezeichnet. AES können sich dieses Prozesses bedienen, um *vor* ihrem Einsatz ein Modell zu erlernen, auf dessen Basis sie sich verhalten. Maschinelle Lernverfahren können jedoch auch eingesetzt werden, um *während* des Einsatzes eines AES dieses bzw. dessen genutztes Modell zu verändern. Veränderungsfähigkeit meint hier die (nachträgliche) Anpassung des (Ausgangs-)Modells des AES während der Nutzung durch neu aufgezeichnete bzw. gewonnene Daten.

Für viele Anwendungsbereiche kann es starke Vorteile mit sich bringen, wenn das fertig trainierte Modell noch während des Einsatzes angepasst werden kann. So kann es bspw. sinnvoll sein, ein medizinisches Diagnosesystem nachzutrainieren, wenn sich der Datenpool der Patientengruppe er-

55 BGHZ 200, 38 ff.
56 So auch *Rudin*, Interpretable Models (Fn. 53).
57 Vgl. *A. M. Turing*, Computing Machinery and Intelligence, Mind 59 (1950), S. 433 ff.

weitert hat. Hierzu werden die neuen Daten genutzt, um das Modell so anzupassen, dass es sowohl die alten, als auch die neuen Daten optimal abbildet. In anderen Fällen kann ein Modell, welches für die Anwendung auf eine bestimmte Gruppe von Menschen ausgerichtet ist, durch Neugewichtung auf ein bestimmtes Individuum angepasst werden, so dass es die besonderen Daten der Person gut beschreibt und optimal erklärt.

Die Veränderungsfähigkeit birgt jedoch auch Risiken. Haben wir es mit einem nachlernenden AES zu tun, kann im Prinzip nicht mehr gewährleistet werden, dass die Funktion und Korrektheit, die zu Beginn des Einsatzes attestiert wurde, während der weiteren Verwendung gegeben bleibt. Trotzdem sind nach unserer Auffassung nachlernende AES die relevanteste Form zukünftiger AES, da sie in der Lage sind, sich weiter zu trainieren und zu verbessern (1), sich individuellen Anforderungen anzupassen (2) und sogar ungewünschte Änderungen (wie bspw. Abnutzung) kompensieren können (3). Tatsächlich liegt also eine besondere Herausforderung darin, Regulierungs- und Zertifizierungsmöglichkeiten zu schaffen, die dieser Art von AES und deren Veränderlichkeit gerecht werden. Die Integration von Lernverfahren kann für die Gesamtfunktionalität und Sicherheitskonzepte von AES eine erhebliche Relevanz aufweisen.

II. Anwendungsbezogene Risiken – Beispiel Exoskelette

Die zweite Risikoebene eines Einsatzes von AES bezieht sich auf deren Anwendungsbereich. Die Bandbreite der durch den Einsatz von AES betroffenen Rechtsgüter ist aufgrund der Vielzahl von Anwendungsmöglichkeiten groß. Die Risiken können dabei teilweise bereits durch bestehende Normen abgedeckt sein. In den meisten Fällen ergeben sich jedoch aus der Koppelung mit den systembezogenen Risikobereichen neue Risiken, die regulatorischen Bedarf verdeutlichen. Die Einwirkungsmöglichkeiten von AES auf materielle wie immaterielle Rechtsgüter können dabei immens sein – sei es analog durch Robotik oder auch rein digital wie etwa durch die Einstufung der Rückfallwahrscheinlichkeit eines Straftäters. Im Folgenden soll anhand der Entwicklung eines medizinischen Exoskeletts erläutert werden, welche Relevanz die Lernfähigkeit für die Therapie hat, welche Risiken dadurch auftreten können und wie diesen entgegengewirkt werden kann.

Exoskelette[58] sind robotische AES, die direkt am Körper eingesetzt werden. Neben passiven Exoskeletten, die Kräfte umleiten und so das Heben und Tragen vereinfachen sollen, werden etwa zu medizinischen Zwecken auch aktive Exoskelette entwickelt, die über Motoren Kräfte in den Körper des Trägers leiten. So möchte man etwa Gelähmten die Möglichkeit körperlicher Fortbewegung erleichtern oder sogar zurückgeben. Im Folgenden soll anhand der Entwicklung eines solchen medizinischen Exoskeletts erläutert werden, welche Relevanz die Lernfähigkeit für die Therapie hat, welche Risiken dadurch auftreten können und wie diesen entgegengewirkt werden kann.

Der Patient soll bspw. einen Arm mit Hilfe des AES wieder bewegen können. Das System muss also "wissen", wann und wohin sich der Arm bewegen soll. Dazu lernt es, Muster aus biologischen Signalen, wie etwa dem Elektroenzephalogramm (EEG), zu erkennen, die eine Bewegungsintention ableiten lassen,[59] um im Bedarfsfall durch Krafteinleitung über Motoren gezielte Bewegungen zu ermöglichen. Da sich die EEG-Signale in ihrer Stärke und Struktur je nach individueller Schädigung und Tagesform der Patienten stark unterscheiden, ist es ungleich schwieriger, einen Therapieerfolg mit einem "klassischen" System ohne Lernfähigkeit zu erzielen. Durch kontinuierliche Anpassung des Systems im Vorfeld und während des Einsatzes – auch als "Assist-as-needed"-Methode bezeichnet – kann so eine schrittweise Rehabilitation des Patienten ermöglicht werden.[60]

Durch die direkte Anwendung des Systems am Menschen entstehen erhebliche Risiken. Einerseits besteht die Gefahr, dass die durch ein technisches System typischerweise auftretenden Emissionen (Strahlung, Wärme, Vibrationen, Austritt von Betriebsmitteln wie Schmierstoffe, etc.) aufgrund der engen Verbindung intensiver auf den Menschen einwirken können. Andererseits kann auch das Gerät selbst potentiell Körperteile im direkten Kontakt verletzen oder quetschen; deshalb muss die Konstruktion

58 Altgriechisch etwa für "Außenskelett".

59 Bspw. zur Erkennung von EEG-Signalen siehe *B. Libet/C. A. Gleason/E. W. Wright/D. K. Pearl*, Time Of Conscious Intention To Act In Relation To Onset Of Cerebral Activity (Readiness-Potential): The Unconscious Initiation Of A Freely Voluntary Act, Brain 106 (1983), S. 623 ff.

60 *N. Will/E. A. Kirchner/F. Kirchner*, Künstliche Intelligenz und robotergestützte Rehabilitation, in: H. Hanika (Hrsg.), Künstliche Intelligenz, Robotik und autonome Systeme in der Gesundheitsversorgung, Sternenfels 2019, S. 101 ff.; *E. A. Kirchner/N. Will/M. Simnofske/P. Kampmann/L. M. Vaca Benitez/J. de Gea Fernández/F. Kirchner*, Exoskelette und künstliche Intelligenz in der klinischen Rehabilitation, in: M. Pfannstiel/P. Da-Cruz/H. Mehlich (Hrsg.), Digitale Transformation von Dienstleistungen im Gesundheitswesen V, Wiesbaden 2019, S. 413 ff.

hohen ergonomischen Anforderungen genügen. Für viele dieser Risiken bestehen bereits Anforderungen auf Grundlage der DIN EN ISO 13482.

Jedoch ergeben sich aufgrund der Lernfähigkeit und der autonomen Steuerung des Systems neue Eigenschaften, die nicht durch die Norm abgedeckt sind. Auch ohne die durch die Lernfähigkeit indizierte ständige Veränderung des Systems gibt es spezielle Risiken, die nicht völlig beherrschbar sind. Bspw. sind die verwendeten Biosignale – wie das EEG – aufgrund ihres Rauschens nicht vollständig interpretierbar. Auch können nur schwer berechenbare Restkräfte des Patienten bestehen.

Zur Abwendung dieser Risiken können direkt bei der Konstruktion und beim Betrieb des Systems hard- und softwareseitige Maßnahmen getroffen werden. So können Endanschläge eingebaut werden, die die Einwirkungsmöglichkeiten der Antriebe begrenzen. Auch sorgen Stromlimits dafür, dass die verwendeten Motoren nicht übersteuern. Gleichermaßen können Limits in der Software dafür sorgen, dass der Arbeitsbereich der Aktuatoren eingeschränkt wird und beim Versagen der Sensorik ein Abschalten des Roboters sichergestellt ist. Diese Maßnahmen können u.a. dafür sorgen, dass eine Fehlfunktion auf der Softwareseite, die autonome Funktionen umsetzt oder Lernen ermöglicht, nicht zu einer Gefährdung des Nutzers führen kann. Durch die Integration des Embedded Brain Reading Ansatzes, der eine sichere Nutzung von unsicheren Ergebnissen aus der Klassifikation der EEG-Signale ermöglicht, kann das Zusammenspiel zwischen lernender Komponente und robotischer Assistenz weiter verbessert werden.[61]

C. Fazit – Eckpunkte einer Regulierung

I. Grundlagen

Nach der Betrachtung möglicher systemischer und anwendungsspezifischer Risiken, die von der Technologie ausgehen können, kehren wir zur Ausgangsfrage des Beitrages zurück: Was könnten die Eckpunkte einer möglichen Regulierung sein?

Wie bei A.III dargestellt, versuchen die beteiligten Akteure einen "risikoadaptierten Regelungsansatz" zu finden. Für die Ermittlung und rechtliche Bewertung eines "Risikos" sind als Maßstäbe weniger feststehende

61 *E. A. Kirchner/R. Drechsler*, A Formal Model for Embedded Brain Reading, Industrial Robot 40 (2013), S. 530 ff.

Schadenspotenziale oder Eintrittswahrscheinlichkeiten von Schäden relevant, vielmehr geht es um die Eingrenzung des Ausmaßes einer möglichen Fehleinschätzung.[62] Das Vorhandensein von "Risiken" lässt sich im Gegensatz zu dem Bestehen von "Gefahren" auch nicht völlig verhindern, ergeben sie sich doch fast zwangsläufig aus der Nutzung von Hochtechnologie, will man damit den gesellschaftlichen Reichtum vergrößern.[63] Das Versprechen der Wohlstandsmehrung wird auch beim Einsatz von AES abgegeben.[64] Es wird jedoch nur einzulösen sein, wenn die dargestellten Risiken sich nicht übermäßig zum Nachteil der Betroffenen realisieren und ein möglichst breiter Teil der Gesellschaft in den Genuss der Vorteile der Technik kommt.

Ziel der Regulierung muss es also sein, die Risiken von AES adäquat zu bewerten und durch entsprechende Verpflichtungen dafür Sorge zu tragen, dass Schädigungen möglichst verhindert werden und kontrollierbar bleiben.[65] Jedoch kann auch eine allzu risikoaverse Regulierung zu einem gesamtgesellschaftlichen Schaden führen – und zwar dann, wenn dadurch Chancen zur Förderung von Innovation und Zukunftsfähigkeit ungenutzt bleiben.[66] Auch der individuelle Schaden kann immens sein, wenn der Einsatz von AES zu stark sanktioniert wird. Betrachten wir nur das Beispiel des Gelähmten, der – wie in Teil B.II dargestellt – durch die Nutzung eines medizinischen Exoskeletts wieder die Möglichkeit auf eine selbstbestimmte Fortbewegung hätte. Das Spannungsverhältnis bei der Regulierung bewegt sich zwischen der staatlichen Schutzpflicht einerseits und andererseits der verfassungsrechtlich garantierten Handlungs- und Berufsfreiheit derjenigen Personen und Unternehmen, die AES entwickeln, einsetzen und vertreiben.[67]

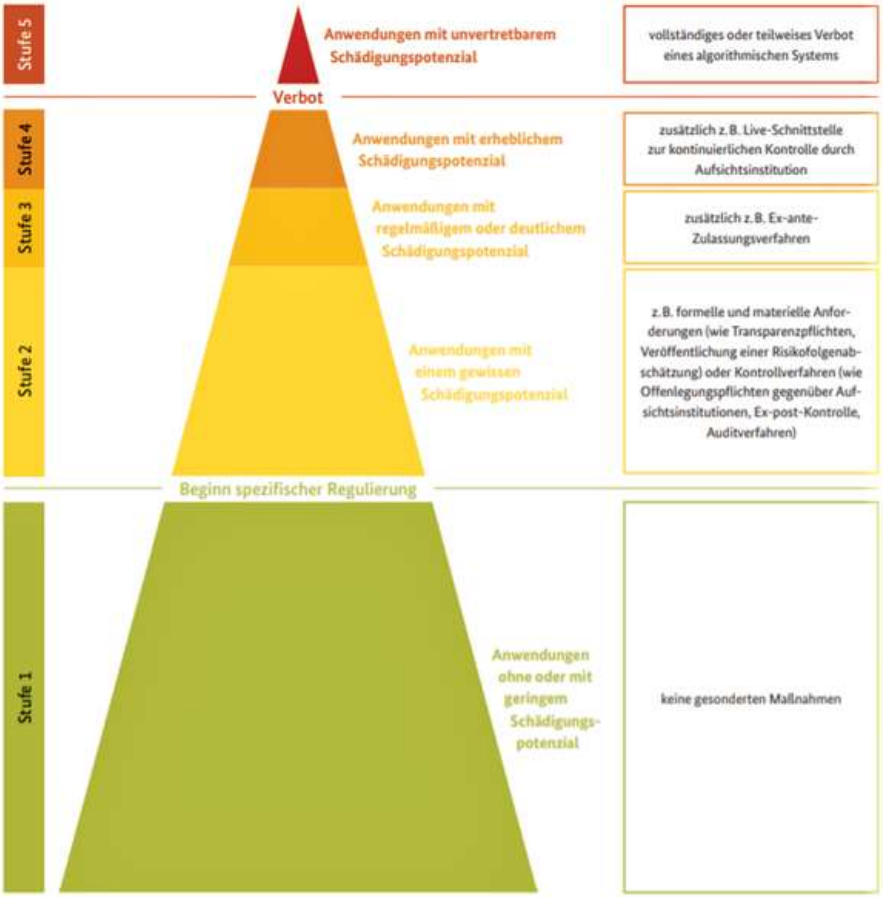
II. Dynamische Koppelung der Risikobereiche

Wie bewertet man das spezifische Risiko eines AES? Eine bereits bei A.III in Aussicht gestellte horizontale Regulierung läuft durch ihren universel-

62 *A. Scherzberg*, Risiko als Rechtsproblem, VerwArch 1993, S. 484 (497 f.).
63 *U. Beck*, Risikogesellschaft – Auf dem Weg in die andere Moderne, 23. Auflage, Berlin 2016, S. 25 ff.
64 Bspw. Bundesregierung, KI-Strategie (Fn. 7), S. 25.
65 Vgl. *U. Di Fabio*, Risikoentscheidungen im Rechtsstaat, Tübingen 2019, S. 41 ff.
66 *A. Scherzberg*, Risikosteuerung durch Verwaltungsrecht? Ermöglichung oder Begrenzung von Innovation?, VVDStRL 2004, S. 214 (233 f.).
67 *Martini*, Algorithmus (Fn. 8), S. 109.

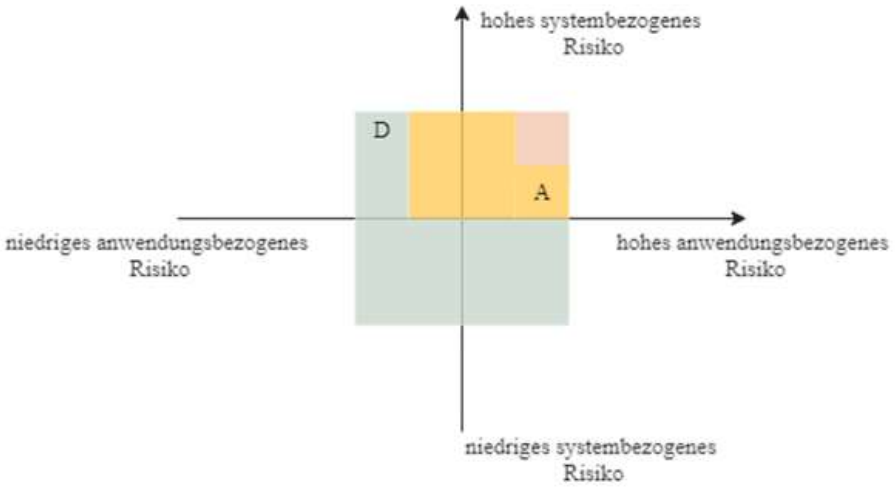
len Anspruch latent Gefahr, den Nutzen des Einsatzes von AES durch starre Anforderungen zu stark einzuschränken. Ähnlich wurde und wird bei der DSGVO argumentiert.[68] Einige Elemente der bestehenden Technikregulierung eignen sich auch für den Einsatz bei AES.[69]

Dieser Beitrag möchte im Gegenzug dazu einen ganzheitlichen, aber dennoch dynamischen Ansatz vorstellen: die Koppelung der Risikobereiche von AES. Dieses Modell soll auf der "Kritikalitätspyramide" der *Datenethikkommission* (s.u.) aufbauen.[70]



68 Zum Prozess der Evaluation der DSGVO nach 2 Jahren siehe *C. Geminn/C. Leontopoulos*, Stellungnahmen zur DSGVO, ZD-Aktuell 2020, 07024.
69 *Martini*, Algorithmus (Fn. 8), S. 113 ff.
70 Gutachten der Datenethikkommission (Fn. 17), S. 177.

Durch eine Kopplung der anwendungsbezogenen Risiken mit den systembezogenen Risiken soll die "Pyramide" aber zu einer Matrix werden, innerhalb derer Regulierungsmaßnahmen verortet werden können. Dies folgt aus der Überlegung heraus, dass aufgrund der Vielzahl von Anwendungsmöglichkeiten von AES bei gleichzeitig bestehenden anwendungsunabhängigen systembezogenen Risiken eine eindimensionale Bewertung ("niedriges Risiko – hohes Risiko") zu kurz gegriffen wäre. Das anwendungsbezogene Risiko als solches kann ja auch bereits durch andere Techniknormen reguliert sein. Zu überprüfen ist daher, ob sich gerade aus der Kombination mit den systembezogenen Risiken ein – möglicherweise höheres – AES-spezifisches Risiko ergibt, welches besondere Regulierungsmaßnahmen erforderlich machen könnte. Wenn anwendungsbezogen mit gewichtigen Rechtsgütern operiert wird, sollten auch erhöhte Anforderungen an die Bewältigung der systembezogenen Risiken gestellt werden. Das Gleiche gilt aber auch umgekehrt – falls mit geringerwertigen Rechtsgütern operiert wird, dann sollte es zu einer Verringerung der Anforderungen kommen, damit Regulierungsmaßnahmen verhältnismäßig sind.

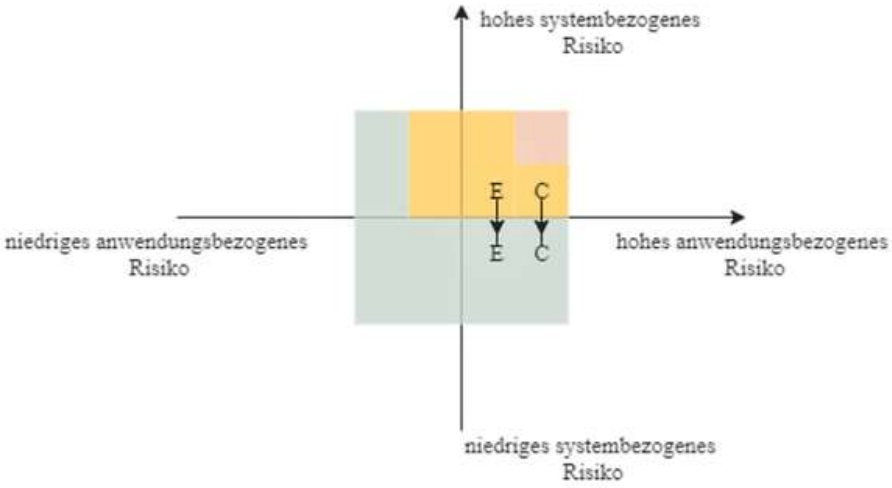


Grün bedeutet ein nicht vorhandenes bis niedriges Erfordernis an regulatorischen Maßnahmen, Gelb eine bestehende Erforderlichkeit und Rot steht für insgesamt hohe Anforderungen an den legalen Betrieb bis hin zu einem Verbot.

Eine Dating-App ("D") operiert mit einem AES. Dieses ist aufgrund der Modellgestaltung intransparent. Dadurch, dass das AES mit weniger gewichtigen Rechtsgütern wie den persönlichen Daten und vielleicht dem Vermögen des Benutzers auf vertraglicher Basis interagiert, besteht ein ge-

ringes anwendungsbezogenes Risiko, weshalb das hohe systembezogene Risiko hier eher nicht zu einem regulierungsbedürftigen AES-spezifischen Risiko führt.

Ein AES, welches in einem autonomen Fahrzeug eingesetzt wird ("A"), operiert regulär mit hohem anwendungsbezogenem Risiko. Höherwertige Rechtsgüter wie Leben, Gesundheit und Eigentum des Nutzers wie auch unbeteiligter Dritter wären potenziell betroffen. Deshalb kann nur ein niedriges systembezogenes Risiko toleriert werden, sonst droht eine Verschiebung in den roten Bereich.



Auch an den Einsatz des medizinischen Exoskeletts aus dem Beispiel oben sind erhöhte Anforderungen zu stellen. Einerseits bestehen anwendungsbezogene Risiken – etwa für die Rechtsgüter körperliche Integrität und Gesundheit, aber auch für die sensiblen Gesundheitsdaten etwaiger Patient*Innen. Andererseits bestehen bspw. durch die Nutzung der schwer interpretierbareren Bio-Signale und die Ermöglichung der Lernfähigkeit auch systembezogene Risiken. Die Herausforderung liegt hier darin sicherzustellen, dass etwa eine "angelernte" Fehlinterpretation des Systems – also die Realisierung der systembezogenen Risiken – nicht auch eine Realisierung anwendungsbezogener Risiken zur Folge hat. Dies könnte bspw. durch die Verwendung bestimmter Hard- oder Softwarekomponenten geschehen, die etwa innerhalb des Systems das durch die Biosignalverarbeitung erhöhte systembezogene Risiko kompensieren können.[71] Dadurch

71 s. B. II. am Ende.

lässt sich insgesamt auch eine Reduzierung des AES-spezifischen Risikos erreichen ("E").

Auch am Beispiel der COMPAS-Software (Teil B.I.1) war zu sehen, dass das AES-spezifische Risiko mithilfe der Wissenschaftler*Innen verringert werden konnte, indem das systembezogene Risiko verringert wurde (durch die Reduktion der Komplexität des Systems). Auch hier hat sich die Position in der Matrix verändert ("C").

Mit dem Ergebnis dieser Risikobewertung können dann regulatorische Maßnahmen verbunden werden. Beispielhaft wäre die Einführung einer verschuldensunabhängigen Gefährdungshaftung für risikointensivere AES-Anwendungen verbunden mit einer Pflichtversicherung.[72] Vorstellbar ist auch die Implementierung eines sektorspezifischen Registers und die Einrichtung einer Aufsichtsbehörde, wie von der *Datenethikkommission* vorgeschlagen. Auch denkbar sind Auflagen, die Betreiber von bestimmten AES verpflichten könnten, ihre systembezogenen Risiken in einer besonderen Weise zu evaluieren – wie das bspw. die AI NOW Gründer*Innen vorsehen (Teil B.I.1). Als schwierig erweist sich auch die regulatorische Einhegung des Risikos der Veränderungsfähigkeit – wenn man bedenkt, dass das Ordnungs- und Technikrecht sonst eher statische Zustände von Objekten beurteilt, die sich im Laufe des Einsatzes nicht (zumindest nicht beabsichtigt) verändern. Eine Lösung könnten hier Live-Schnittstellen sein, die aktuelle Zustände übermitteln und die bei Erreichen kritischer Limits einen Alarm senden. Eine andere Herangehensweise wäre etwa auch die Einführung einer Pflicht zum Nachweis der Absicherung veränderbarer Komponenten durch übergreifende Sicherheitsmaßnahmen, welche ein Gesamtversagen des AES verhindern. Wichtig ist auch die Unterscheidung zwischen *vollautomatisierten* und *teilautomatisierten* AES. Dieser Unterschied könnte Auswirkungen auf das Zertifizierungserfordernis, die Wartungsintervalle, eine veränderte Aufsichts- oder Genehmigungspflicht haben. Möglicherweise lassen sich all diese Maßnahmen für AES-Anwendungen auch in einer einzelnen, europaweit geltenden horizontalen Regulierung kombinieren, an deren Anfang dann tatsächlich ein "KI-TÜV" steht, der das spezifische AES-Risiko ermittelt und entsprechende Verpflichtungen der Betreiber nach sich zieht. Dadurch lässt sich auch Rechtssicherheit für den Einsatz und die weitere Entwicklung von AES gewinnen, so dass ein gesellschaftlicher Mehrwert durch die Technologie entstehen kann.

Die Definition von AES, die Betrachtung und Abgrenzung ihrer Risiken und die Einbettung dieser in die Matrix sind lediglich als die Eckpunk-

72 *H. Zech*, Künstliche Intelligenz und Haftungsfragen, ZfPW 2019, S. 198 (215).

te einer möglichen Regulierung zu verstehen. Offen bleibt bspw. eine exakte Verhältnismäßigkeitsprüfung von regulatorischen Maßnahmen zur Risikoabwägung einer einzelnen AES-Anwendung, die Gegenstand eines Folgebeitrags sein könnte.

Appendix B

XAI User Study Literature Review

In this part of the appendix, we provide additional details on our structured literature review of study designs for evaluating explanations, as presented in Section 2.5. We report the search query and selection criteria used with the empirical-to-conceptual method [166]. Finally, we list all the publications that were considered.

B.1 Search Query

Our search queries were composed of *groups* and *terms*. Groups refer to a specific aspect of the research question and limit the search scope. Terms have a similar semantic meaning within the group domain or are often used interchangeably. We were interested in the intersection of 3 groups that can be phrased using different terms. Table B.1 shows our used groups and terms.

Group	**Terms**
1 - Explainable	*explainability, explainable, explanation, explanatory, interpretability, interpretable, intelligibility, intelligible, scrutability, scrutable, justification*
2 - AI	*XAI, AI, artificial intelligence, machine learning, black-box, recommender system, intelligent system, expert system, intelligent agent, decision support system*
3 - Human Subject Evaluation	*user study, lab study, empirical study, online experiment, human experiment, human evaluation, user evaluation, participant, within-subject, between-subject, probe, crowdsourcing, Mechanical Turk*

Table B.1. Search query: The query was composed of groups and terms.

B.2 Selection Criteria

The search, executed on Scopus in September 2019, returned a total of 653 publications. We filtered the search results by six exclusion criteria (EC) and one inclusion criterion (IC). We were interested in primary studies that report the setup and result of human subject evaluations in the XAI context (IC-1). We limited the survey to publications written in English (EC-1), which address the *black-box explanation problem*, according to Guidotti et al. [82] (EC-2). We exclude publications that do not report *human-grounded* or *application-grounded* evaluations according to Doshi-Velez and Kim [52] (EC-3). Naturally, we also excluded publications where the full text could not be retrieved (EC-4). We further limited the review to scientific papers (EC-5). Finally, we removed duplicates, like copies from arXiv (EC-6). We applied the exclusion criteria in cascading order, which means if we excluded publications due to one EC, we did not assess any of the following criteria.

B.3 Reviewed Publications

After screening, 133 publications remained for analysis, of which we reviewed 35 for the initial taxonomy which was published as a workshop paper (1-35 in the list below.) Subsequent studies drew resources away from the planned more extensive review. However, we keep reviewing studies as they were published (36-52 in the list). Based on this, the author refined the taxonomy. In its current version, the types of tasks are more detailed than in the initial version.

1. Bussone et al., *"The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems"* [32]
2. Schaffer et al., *"I Can Do Better than Your AI: Expertise and Explanations"* [196]
3. Kulesza et al., *"Principles of Explanatory Debugging to Personalize Interactive Machine Learning"* [121]
4. Lim et al., *"Why and Why Not Explanations Improve the Intelligibility of Context-Aware Intelligent Systems"* [132]
5. Weitz et al., *""Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design"* [250]
6. Teso and Kersting, *"Explanatory Interactive Machine Learning"* [230]
7. El Bekri et al., *"A Study on Trust in Black Box Models and Post-hoc Explanations"* [59]
8. Zhou et al., *"Interpretable Basis Decomposition for Visual Explanation"* [264]
9. ElShawi et al., *"Interpretability in healthcare: A comparative study of local machine learning interpretability techniques"* [60]

10. Hutton et al., *"Crowdsourcing Evaluations of Classifier Interpretability"* [95]
11. Ehsan et al., *"Automated Rationale Generation: A Technique for Explainable AI and its Effects on Human Perceptions"* [57]
12. Kouki et al., *"Personalized Explanations for Hybrid Recommender Systems"* [120]
13. Lakkaraju et al., *"Faithful and Customizable Explanations of Black Box Models"* [126]
14. Nunes et al., *"Pattern-based Explanation for Automated Decisions"* [171]
15. Harbers et al., *"Design and Evaluation of Explainable BDI Agents"* [85]
16. Sato et al., *"Explaining Recommendations Using Contexts"* [194]
17. Ming et al., *"Interpretable and Steerable Sequence Learning via Prototypes"* [148]
18. Tsai and Brusilovsky, *"Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance"* [234]
19. Grover et al., *"BEEF: Balanced English Explanations of Forecasts"* [81]
20. Tintarev and Masthoff, *"Evaluating the effectiveness of explanations for recommender systems"* [233]
21. Razak et al., *"Interpretability and Complexity of Design in the Creation of Fuzzy Logic Systems — A User Study"* [182]
22. Lamy et al., *"Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach"* [127]
23. Lakkaraju et al., *"Interpretable Decision Sets: A Joint Framework for Description and Prediction"* [125]
24. Hohman et al., *"Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models"* [93]
25. Ribeiro et al., *"Anchors: High-Precision Model-Agnostic Explanations"* [185]
26. Dodge et al., *"Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment"* [50]
27. Millecamp et al., *"To Explain or not to Explain: the Effects of Personal Characteristics when Explaining Music Recommendations"* [145]
28. Gedikli et al., *"How should I explain? A comparison of different explanation types for recommender systems"* [72]
29. Nugent et al., *"The Best Way to Instil Confidence Is by Being Right"* [169]
30. Subramanian et al., *"SPINE: SParse Interpretable Neural Embeddings"* [224]
31. Nguyen et al., *"An Interpretable Joint Graphical Model for Fact-Checking From Crowds"* [164]

32. Muhammad et al., *"A Live-User Study of Opinionated Explanations for Recommender Systems"* [157]
33. Eisenstadt et al., *"Explainable Distributed Case-Based Support Systems: Patterns for Enhancement and Validation of Design Recommendations"* [58]
34. Zanker, *"The Influence of Knowledgeable Explanations on Users' Perception of a Recommender System"* [261]
35. Lully et al., *"Enhancing explanations in recommender systems with knowledge graphs"* [137]
36. Adebayo et al., *"Debugging Tests for Model Explanations"* [5]
37. Borowski et al., *"Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization"* [25]
38. Chu et al., *"Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction"* [41]
39. Kaur et al., *"Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning"* [103]
40. Mertes et al., *"GANterfactual — Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning"* [142]
41. Kim et al., *"Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)"* [108]
42. Cai et al., *"The Effects of Example-Based Explanations in a Machine Learning Interface"* [33]
43. Shen and Huang, *"How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels"* [203]
44. Ribeiro et al., *""Why Should I Trust You?"* [184]
45. Kulesza et al., *"Fixing the Program My Computer Learned"* [123]
46. Springer and Whittaker, *"Progressive Disclosure"* [220]
47. Springer et al., *"Dice in the Black Box: User Experiences with an Inscrutable Algorithm"* [219]
48. Feng and Boyd-Graber, *"What Can AI Do for Me?"* [66]
49. Lai and Tan, *"On Human Predictions with Explanations and Predictions of Machine Learning Models"* [124]
50. Buçinca et al., *"Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems"* [31]
51. Kim et al., *"HIVE: Evaluating the Human Interpretability of Visual Explanations"* [111]
52. Ramaswamy et al., *"Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability"* [181]

Appendix C

XAI Study I: Instructions and Collected Replies

C.1 Collected Replies

Participants' answers were anonymised and made available as a public dataset:

`https://doi.org/10.5522/04/11638275.v2`

C.2 Participant Instructions

Study participants had to complete a tutorial spanning 8 pages.

Page 1

Hello and thank you for participating in this study. Please read the following instruction carefully. It contains valuable information which will allow you to **earn additional rewards** during this study.

One of the successful applications of machine learning (ML) is image recognition. It can be used to assign "labels" of recognized objects to photos. For this, the ML system has to be "trained" on a large number of photos, which were manually labeled. The set of photos used for training is called the **"training set."**

For this study, we pre-trained a system to recognize 20 different labels. The 20 labels are: aeroplane, bicycle, bird, boat, bottle, bus, car, cat, chair, cow, dining table, dog, horse, motorbike, person, potted plant, sheep, sofa, train, tv monitor.

So if any of these appear in a photo, the system should recognize them and assign the corresponding label.

Page 2

So when is the system correct and when is it wrong? For each label, the system calculates a score (from 0 to 1). The score will be higher if the system is very sure about assigning a label and it will be lower if it is unsure. Each photo can contain multiple objects. Therefore it could need to be assigned multiple labels. The system will assign all labels which scores are higher than a predefined threshold. Please note that ML systems are generally not 100% accurate. They may work well on some photos (hopefully most of them), but make mistakes on other photos (hopefully just a few of them). It is useful to consider the following 4 outcomes where the system makes mistakes or is correct.

1. The image contains object X (e.g. a cat) and the system correctly recognizes it (hurrah!) – this is a **"true positive" (TP)**.
2. The image does NOT contain object X (e.g. a cat) and the system correctly recognizes that there is no such object (hurrah!) – this is a **"true negative" (TN)**.
3. The image contains object X (e.g. a cat), but the system does not recognize it (oops!) – this is a **"false negative"** (FN).
4. The image does NOT contain object X (e.g. a cat), but the system falsely recognized such an object in the image (oops!) – this is a **"false positive"** (FP).

Looking at some examples for each of the outcomes can reveal which images the system recognizes well and with which images it is struggling. In this study, we ask you to study such examples and estimate how the system will perform.

Page 3

Your main task is to estimate whether the system can successfully assign a label to several photos. To help you with this, we will show you 12 example photos from the training set that are visually similar to the photo you are currently working on. Concretely you will be shown:

- 6 photos that are True Positive (TP) examples
- 3 photos that are False Negative (FN) examples
- 3 photos that are False Positive (FP) examples

Page 4

Because each image can contain multiple objects, the system accepts the predicted labels which satisfy the following criterion: **The score of the predicted label has to be higher than a pre-defined threshold.** Please note that each "category" has its own threshold, which is shown as small red lines. See this images as an example:

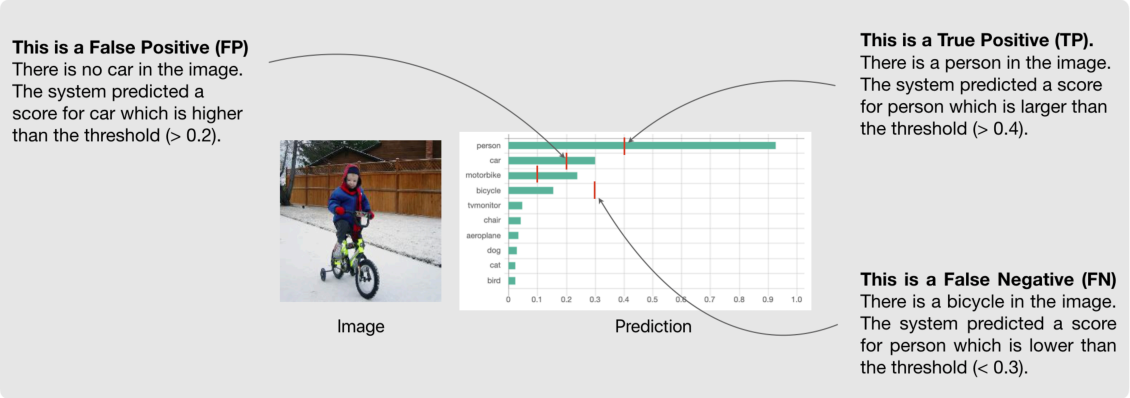


Figure C.1. Screenshot of participants' instructions on how to read classification results, scores, and score thresholds.

Page 5 (Participants were shown Figure C.2)

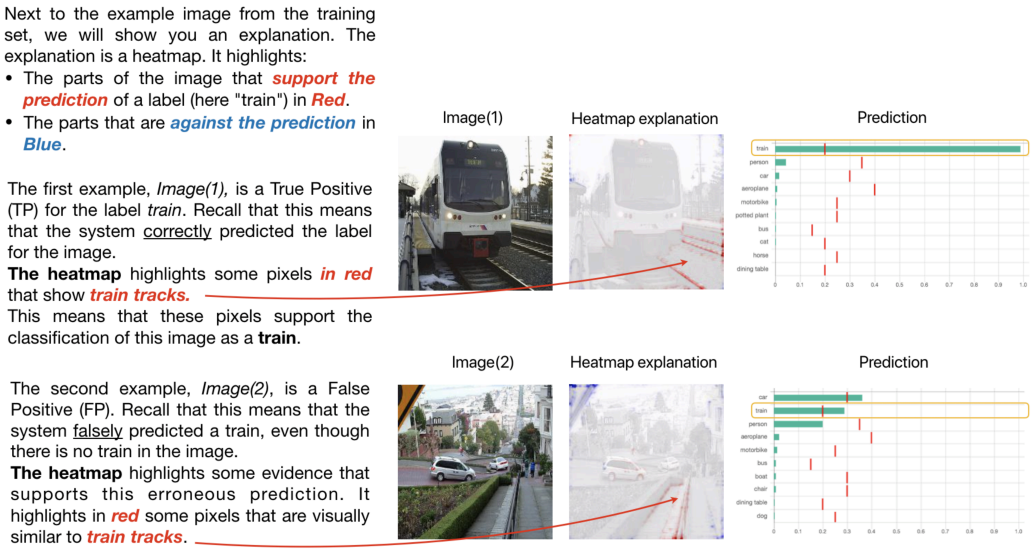


Figure C.2. Screenshot of participants' instructions for saliency maps.

Page 6 (Participants were shown Figure C.3)

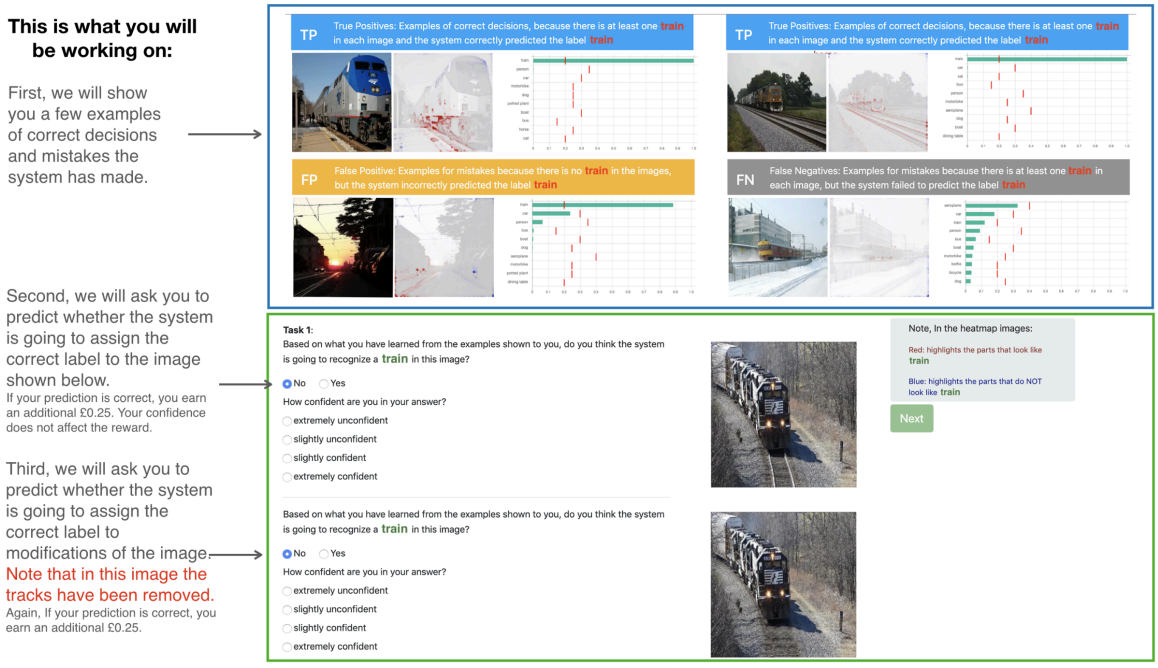


Figure C.3. Screenshot of instructions for the study interface.

Page 7 For each example, we show you the following information (Figure C.4).

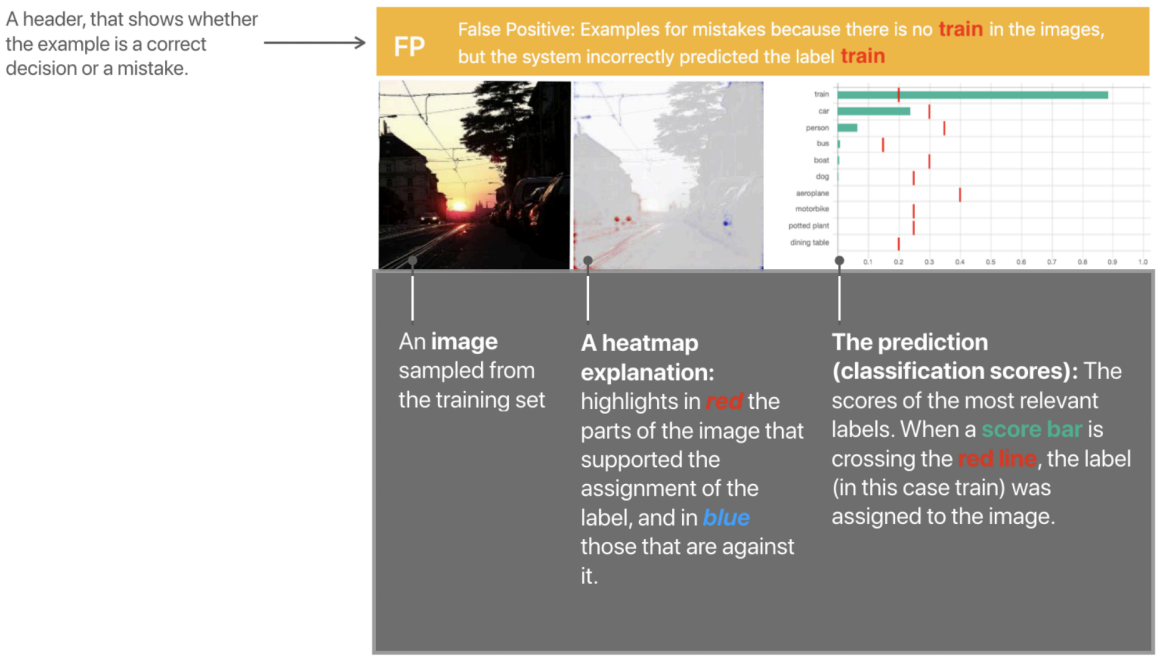
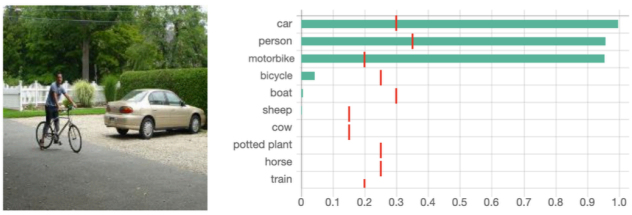


Figure C.4. Screenshot of instructions for example images.

Page 8

Before you start the task, please answer the following question(s): (you can refer to the previous pages) [see Figure C.5a and C.5b]

Ready? The total number of questions are **12**. Remember, for each correct answer, you will receive £0.5 extra!



(a) Questions: (1) Is this a False Negative (FN) example of bicycle? (2) Is this a False Negative (FN) example of motorbike?

(b) Questions: (1) The heatmap explanation suggests that the prediction of 'car' is supported by the presence of a rear window. (2) The heatmap explanation suggests that the prediction of 'car' is supported by the presence of a tire/wheel.

Figure C.5. Attention checks: Participants had to answer four question with *Yes* or *No*.

Appendix D

XAI Study II: Instructions, Preregistration, and Models

D.1 Instruction Videos and Screening Questions

Before participants could work on their task, they had to watch a tutorial video accompanied by a written summary. They then had to answer screening questions. The screening question of each stage could be attempted twice. When participants failed a screening, they were compensated for their time.

Introduction Tutorial for Peeky and Stretchy

All participants viewed the following video:

`https://f002.backblazeb2.com/file/iclr2022/Intro_Peeky_Stretchy.mp4`

Screening 1 After watching the video, they had to identify two Peekies and two Stretchies from four images correctly and had to select all correct statements from the following selection:

- ✓ "Peeky and Stretchy are made of 8 blocks: 4 for the legs, 4 for the spine."
- ✓ "The blocks that make up the animals can be more rounded or more rectangular."
- × "The animals always have the same background color."
- ✓ "Each animal can have a different color."

Introduction Tutorial for Machine Learning and Biases

All participants viewed the following video:

https://f002.backblazeb2.com/file/iclr2022/Second_Intro_ML.mp4

Screening 2 Participants had to select all correct statements from the following selection:

× "We know exactly which characteristics the trained machine learning system uses to distinguish between the two animals. It is the leg position relative to the spine. The system will not use any other characteristics."

✓ "After the learning phase, given an image, the machine learning system can predict which animal is shown in that image. To do that, it may use any combination of characteristics."

✓ "We trained a machine learning system by feeding it with thousands of images of the two animals."

Tutorial for Baseline Condition

Participants assigned to the baseline condition viewed the following video:

https://f002.backblazeb2.com/file/iclr2022/condition_BASE.mp4

Screening 3 After watching the video, they had to select all correct statements from the following selection:

✓ "Your task is to discover which characteristics the system is using."

✓ "The images are ordered on the horizontal axis according to the certainty of the system about its prediction (on the left - very certain Peeky, on the right - very certain Stretchy)."

× "If the system is using the background as characteristic, all "very certain Stretchy" images will have the same background color."

× "It is enough to look at one row of images to understand which characteristics are relevant for the system."

Tutorial for Invertible Neural Network Interpolations Condition

Participants assigned to the INN condition viewed the following video:

https://f002.backblazeb2.com/file/iclr2022/condition_INN.mp4

Screening 3 After watching the video, they had to select all correct statements from the following selection:

- ✓ "Your task is to discover which characteristics the system is using."
- ✓ "On each row, the system modifies the characteristics of the image. The characteristics are modified so that the system changes its prediction (whether the image is a Peeky or Stretchy)."
- × "A characteristic is only relevant for the system if it changes in every row."
- × "It is enough to look at a single row to understand which characteristics are relevant for the system."

Tutorial for Automatically Discovered Concepts Condition

Participants assigned to the CON condition viewed the following video:

`https://f002.backblazeb2.com/file/iclr2022/condition_CONCEPTS.mp4`

Screening 3 After watching the video, they had to select all correct statements from the following selection:

- ✓ "Your task is to discover which characteristics the system is using."
- × "The dark regions in the image highlight the concept that the system has learned."
- × "A concept always contains only one characteristic."
- ✓ "You have to look at several of these concepts to understand which characteristics are important for the system."

D.2 Architecture of the Invertible Neural Network

Leon Sixt developed the model and is his contribution. It is included here as a technical detail. The model is based on the Glow architecture [113] and contains 7 blocks. A block is a collection of 32 flow steps, followed by a down-sampling layer, and ends with a fade-out layer. A single flow step consists of *actnorm*, *invertible* 1×1 *convolution* and *affine coupling* layer. The down-sampling keeps all dimensions, e.g. a shape of (h, w, c) becomes $(h/2, w/2, 4c)$. The fade-out layer maps removes half of the channels. The out-faded channels are than mapped to a standard normal distribution to compute the unsupervised loss. For generating counterfactuals, the out-faded values are not thrown away but rather stored to be used when computing the inverse.

The model was trained using a supervised loss and an unsupervised objective. In total the model had 687 layers and 261 million parameters. The classifier used the output of layer 641. The remaining layers 642-687 were optimized using the standard unsupervised flow objective. For the first 641 layers, we also trained on the classifier's supervised loss.

Let φ denote the first 641 layers and $\mu : \mathbb{R}^n \mapsto \mathbb{R}^n$ the last. φ was trained on both, a supervised loss from the classifier $f(\boldsymbol{x})$ and an unsupervised loss from matching the prior distribution $\mathcal{N}(0, I)$ and the log determinante of the Jacobian. μ is only trained on the unsupervised loss:

$$\arg \min_{\theta_\varphi, \theta_\mu, \theta_f} L_{\text{un}}(\mu \circ \varphi(\boldsymbol{x})) + \beta \, L_{\text{sup}}(\boldsymbol{w}^T \varphi(\boldsymbol{x}) + b, \, y_{\text{true}}). \tag{D.1}$$

For the supervised loss L_{sup}, binary cross entropy was used. As unsupervised loss L_{un}, the commonly used standard flow loss obtained from the change of variables trick was used [48]. The unsupervised loss ensures that inverting the function results in realistic-looking images, which can also be seen as a regularization.

The layer 342 used for the concept explanations is an affine coupling layer.

D.3 Supervised MobileNet-V2 for Attribute Prediction

We used a MobileNet-V2 trained on the unbiased version of Two4Two to predict the attribute values from an image. As Figure D.1 shows, this model could predict each attribute with a marginal small error. Consequently, we could pass interpolated images to this model and quantify how much each attribute had changed as described in Section 4.5.5.

Attribute	Test MSE
Legs' Position	0.000891
Bending	0.000192
Background	0.000113
Color	0.000560
Rotation Pitch	0.000924
Rotation Roll	0.000562
Rotation Yaw	0.002243
Position X	0.000445
Position Y	0.000391
Shapes	0.001102

Figure D.1. Supervised model performance: Test performance of the supervised trained model MobileNet-V2 measured using a mean squared error (MSE).

D.4 Access to Code, Datasets, and Models

Two4Two has been published as an Open Source Github repository. The repository contains Links to the used dataset and model.

`https://github.com/mschuessler/two4two`

All additional code which is the contribution of Leon Sixt has been published as a seperate repository:

`https://github.com/berleon/do_users_benefit_from_interpretable_vision`

D.5 Study Preregistration for the Validation of Two4Two

The Preregistrations can also be viewed under the following URL:

`https://aspredicted.org/blind.php?x=/62X_15J`

1) Have any data been collected for this study already? No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study? This study investigates whether users identify biases learned by a neural network (NN). The neural networks task is to discriminate between two abstract animals (”Peeky” and ”Stretchy”). Each participant is presented with predictions of the system in a 10x5 image grid. After an initial tutorial phase, the participants have to find biases in the model. They do this by scoring different characteristics as relevant or irrelevant. The characteristics are: ”legs position relative to the spine (LEGS)”, ”object color (COLOR)”, ”background (BACK)”, ”rounded or rectangular shape of the blocks (SHAPE)”, and ”rotation and bending (ROT)”. The main research question is whether we succeeded in creating a model that contains at least one bias that is hard to detect, i.e. either COLOR or SHAPE should be harder to detect than LEGS.

> **HB**: *Participants can identify the biases in COLOR or SHAPE less frequently than LEGS.*

3) Describe the key dependent variable(s) specifying how they will be measured. Participant will answer the following questions:

- *LEGS*: How relevant is the legs position relative to the spine for the system?: Relevant or Irrelevant?

- *COLOR*: How relevant is the color of the animal for the system? Relevant or Irrelevant?
- *BACK*: How relevant is the background of the animal for the system? Relevant or Irrelevant?
- *SHAPE*: How relevant is the rounded or rectangular shape of the animal's blocks for the system? Relevant or Irrelevant?
- *ROT*: How relevant is the rotation and bending of the animal for the system? Relevant or Irrelevant?

The ground truth answer is that *LEGS*, *COLOR*, *SHAPE* are relevant while *BACK* and *ROT* are irrelevant. Our first dependent variable is the number of times the head position was selected as relevant. Our second dependent variable is the number of times the color of the animal was selected as relevant. Our third dependent variable is the number of times the rounded or rectangular shape of the animal's blocks was selected as relevant.

4) How many and which conditions will participants be assigned to? Our study follows a within-subject design and has only one condition. We first show the participants introductory videos about the two abstract animals, the machine learning system, and some guidance on how to interpret the predictions of the system. Each video is accompanied by a written summary. We then show the predictions of the system in a grid of images: 10 sorted rows of 5 images drawn from the validation set (50 original images). Each of the five columns represents the neural netwoks's logit range. Similarly rated images are assigned to the same column.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis. We will conduct two exact one-sided McNemar-tests with LEGS acting as our control: one between SHAPE and LEGS and a second between COLOR and LEGS. We will use a one-sided test as we expect that SHAPE and COLOR are harder to identify. The significance level of both tests will be Bonferroni adjusted to $\alpha = 0.025$.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations. We reject participants with low effort responses or who failed to understand the dataset, machine learning concept, or explanation method. We have implemented hard-coded exclusion criteria directly in the survey (implemented with Qualtrics and Prolific).

- did not finish experiment at all or not within 77 minutes

- did not watch tutorial videos completely (there are 3 videos) or failed a multiple-choice comprehension test twice (there are four such tests), unless participants explicitly ask us to retake the study
- using a device smaller than a tablet (min. 600 px in width or height)
- provided answers about relevant characteristics in under 30 seconds
- withdrawn data consent / returned task on Prolific
- circumvented Qualtrics protection against retaking the entire survey again (first complete submission will be counted)

We do not plan to exclude any participants who passed all of the above criteria unless the qualitative answers reveal a serious misunderstanding of the study instructions that the multiple choice tests did not cover.

7) How many observations will be collected or what will determine sample size? 50 participants from Prolific with the background:

- Fluent in English
- Hold an academic degree
- Prolific approval rate of at least 90%
- Did not participate in pilot studies
- Passed hard coded exclusion criteria (see 8).

We pay participants max. 8.00 GBP (6.00 GBP base salary + 2.00 GBP max bonus). For those failing any comprehension questions or not watching the video, we pay:

- First comprehension task: no compensation
- Second comprehension task: 0.5 GBP
- Third comprehension task: 1.75 GBP
- Failed to watch first video: no compensation
- Failed to watch second video: 1 GBP
- Failed to watch third video: 2 GBP

8) Anything else you would like to pre-register? We ask participants to answer three multiple choice comprehension tests in the form of true/false statements to ensure that they understood the task and the dataset. We also ask them to provide some free-text justification of why they chose a relevant / irrelevant rating to the questions in Section 3.

D.6 Study Preregistration for the Main Study

The Preregistrations can also be viewed under the following URL:

`https://aspredicted.org/blind.php?x=/7XN_77P`

1) Have any data been collected for this study already? No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study? This study investigates whether users identify biases learned by a neural network (NN). The neural networks task is to discriminate between two abstract animals (”Peeky” and ”Stretchy”). Each participant is presented one of three different explanation methods: baseline (B), counterfactuals obtained using invertible neural networks (CF) and prototypes (P). Each participant is randomly assigned to a method. After an initial tutorial phase, the participants have to find biases in the model. They do this by scoring different characteristics as relevant or irrelevant. The characteristics are: ”legs position relative to the spine (LEGS)”, ”object color (COLOR)”, ”background (BACK)”, ”rounded or rectangular shape of the blocks (SHAPE)”, and ”rotation and bending (ROT)”.

The main question of our study is whether the participants can correctly identify relevant and irrelevant attributes using these explanation methods (B, CF, P). This is reflected by two hypotheses:

H1: *Participants identify relevant and irrelevant attributes with less accuracy using P compared to B.*

H2: *Participants identify relevant and irrelevant attributes with higher accuracy using CF compared to B.*

3) Describe the key dependent variable(s) specifying how they will be measured. Participant will answer the following questions:

- *LEGS*: How relevant is the legs position relative to the spine for the system?: Relevant or Irrelevant?
- *COLOR*: How relevant is the color of the animal for the system? Relevant or Irrelevant?
- *BACK*: How relevant is the background of the animal for the system? Relevant or Irrelevant?

- *SHAPE*: How relevant is the rounded or rectangular shape of the animal's blocks for the system? Relevant or Irrelevant?
- *ROT*: How relevant is the rotation and bending of the animal for the system? Relevant or Irrelevant?

The ground truth answer is that *LEGS*, *COLOR*, *SHAPE* are relevant while *BACK* and *ROT* are irrelevant. Our dependent variable is the percentage of correctly answered questions per participant (accuracy, which is computed as (true positives + true negatives)/number of total answers).

4) How many and which conditions will participants be assigned to? We run a between-subject study, with randomly but equally assigned participants to 1 of 3 conditions. We first show introductory videos about the two abstract animals, the machine learning system, the explanation technique and some guidance on how to interpret the technique. Each video is accompanied by a written summary. We then show a grid of (10x5) images:

- B: NN predictions explained with 10 sorted rows of 5 images drawn from the validation set (50 original images). Each of the five columns represents a score range. Similarly rated images are assigned to the same column.
- CF: Same grid layout as B, but the NN is explained by counterfactual interpolations. Each row contains interpolations which change the prediction of the NN to fit the designated score. Original images are used as starting points but are not shown.
- P: We found concepts based on the work by Zhang et al. [263]. Each row shows a set of relevant concepts. We only used concepts correlated with at least r=0.2 with the model logit values. In total, we display 10 rows where each row contains a concept. Each row contains a set of 5 example images for which the concept is relevant.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis. We will compute the accuracy scores for each participant and then compare the accuracy scores between the conditions. We expect the data to be non-normally distributed, and will test this assumption using a Shapiro-Wilk test with a significance level of $\alpha = 0.05$. If our assumption is true, we plan to conduct a Kruskal-Wallis test, followed by post-hoc analysis using Wilcoxon's-rank-sum tests for focused comparison between the groups CF and B (expecting higher accuracy in CF) and P and B (expecting lower accuracy in P). If the data is normally distributed, we will conduct a one-way ANOVA with planned contrasts, if the following assumptions of ANOVAs are met:

> Homogeneity of the variance of the population (assessed with a Levene-Test with a significance level of $\alpha = 0.05$.)

If the homogeneity of variance assumption of ANOVA is violated (assessed with a Levene-Test with a significance level of $\alpha = 0.05$.), we plan to perform Welch's Anova.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations. We reject participants with low effort responses or who failed to understand the dataset, machine learning concept, or explanation method. We have implemented hard-coded exclusion criteria directly in the survey (implemented with Qualtrics and Prolific).

- did not finish experiment at all or not within 77 minutes
- did not watch the tutorial videos completely (there are 3 videos) or failed a multiple-choice comprehension test twice (there are four such tests), unless participants explicitly ask us to retake the study
- using a device smaller than a tablet (min. 600 px in width or height)
- provided answers about relevant characteristics in under 30 seconds
- withdrawn data consent / returned task on Prolific
- circumvented Qualtrics protection against retaking the entire survey again (first complete submission will be counted)

We do not plan to exclude any participants who passed all of the above criteria unless the qualitative answers reveal a serious misunderstanding of the study instructions that the multiple choice tests did not cover.

7) How many observations will be collected or what will determine sample size? No need to justify decision, but be precise about exactly how the number will be determined. 240 (80 per condition) participants from Prolific with the background:

- Fluent in English
- First, we sample participants with an academic degree. If we do not reach the desired participant number, which is likely given the limited availability of such subjects, we will supplement with participants with an academic degree in other subjects. All participants will be randomly and equally split into the 4 conditions.
- Prolific approval rate of at least 90%
- Did not participate in pilot studies
- Passed hard coded exclusion criteria (see 8).

We pay participants max. 6.50 GBP (4.50 GBP base salary + 2.00 GBP max bonus). For those failing any comprehension questions or not watching the video, we pay:

- First comprehension task: 0.5 GBP
- Second comprehension task: 1.75 GBP
- Third comprehension task: 3.50GBP
- Failed to watch first video: no compensation
- Failed to watch second video: 1 GBP
- Failed to watch third video: 2 GBP

8) Anything else you would like to pre-register? In a previous study, we collected 50 responses for the baseline condition only (Preregistration #75056)). We do not plan to use the data for this study. We ask participants to answer three multiple choice comprehension tests in the form of true/false statements to ensure that they understood the task and the dataset. We also ask them to provide some free-text justification of why they chose a relevant / irrelevant rating to the questions in Section 3. Additionally, we ask the participants about their machine learning expertise level. Participants can rate their expertise as: complete novice, some expertise, or expert in the topic. We plan to use descriptive statistics to see how accuracies change per condition for each expertise level and how expertise was distributed within our sample. We are also planning a qualitative thematic analysis of the open-text questions in our survey via open and axial coding, with the aim of understanding how participants integrated explanations in their reasoning about the relevance of attributes.

Appendix E

Gazing Heads Study: Experimental Procedure, System Performance and Statistical Results

In this part of the appendix, we provide additional details on our user study presented in Chapter II:

- Section E.1 provides more detail on the study procedure.
- Section E.2 describes the game participants played and the rationale for its design.
- Section E.3 reports additional details on the measurements obtained during the study in four Figures.
- Section E.4 reports on a visual attention analysis which reveals that focus areas received different amounts of attention depending on the task type.
- Section E.5 provides details on the thematic analysis we conducted to better understand participants' experiences with the Gazing Heads.
- Section E.6 reports on the accuracy of our eye-tracking solution and the calculated minimum accuracy for implementing our concept with webcam eye tracking.
- Section E.7 describes how we obtained latency measurement of our system.

E.1 Detailed Study Procedure

Upon arrival, participants were greeted and instructed that the study's goal was to evaluate two video conference systems. The two experimenters guided them to their rooms, adjusting

their seating height for the cameras and calibrating the eye-tracker. Once set up, participants filled in an initial questionnaire. It recorded basic demographics (gender, age, level of education), frequency of use of video conferencing system, and whether they pursue a knowledge worker occupation or education. Additionally, participants rated their agreements to five controversial statements, which were the potential topics for the group discussion. (Statements are listed in Section 6.5.2.)

We controlled the main part of the experiment from an operations room equipped with a server, allowing us to control the workstations to change between conditions, administer questionnaires, and receive live experimental data. We also used it to supervise the experiment and to talk to participants. For the main part of the study, we launched the first system on all stations simultaneously. We told participants to use the system for a few minutes to introduce themselves.

The *first measurement* was always a group discussion. We announced the statement that they should discuss, which we selected based on the agreement ratings collected earlier. We chose topics where ratings were most diverse. Participants were instructed to find a consensus on the topic within five minutes. However, rather than strictly interrupting participants after 5 minutes, we allowed the conversation to continue a little longer until we found a good opportunity to turn the video system off without an awkward interruption. The *second measurement* was always the game. Before commencing, participants listened to prerecorded audio instructions and read about them and their assigned roles. (Audio instructions are stored for 10 years in the digital research archival of the University of Heidelberg and are available upon request.) Once everyone confirmed they understood the game, we reactivated the video conferencing system. When participants reached the threshold of seven minutes, we allowed them to reach a consensus on the last set of items. Then they were told the game would continue using the other system. However, before that, participants needed to answer our UX questionnaire. Once completed, we activated the second system. Again participants could familiarise themselves with the system for a few minutes. The third measurement began by continuing the game with the second system. We interrupted participants again after roughly 7 minutes. Upon announcing the topic for the last discussion, the fourth measurement began and concluded after roughly 5 minutes by turning the system off again. Participants filled in the UX questionnaire a second time for the system they had just used. In addition, they filled in a questionnaire that compared the two systems directly. One question also asked them which system they would like to use for the final part of the experiment. Once the questionnaires were filled in, we reactivated the system the majority of the group chose. Using the system, we conducted a semi-structured interview based on an interview guideline. We asked participants about the advantages and disadvantages of both systems. Furthermore, we ensured that every interview addressed at least two

dimensions of the questionnaire, as this would enable us to understand the questionnaire ratings better. For some groups that mentioned that they were missing shoulders in Gazing Heads, we offered them the chance to try the system with shoulders by removing the green turtleneck. We concluded the interviews once the time slot for the experiment was up, and hence they varied in length. Finally, we helped participants to leave the workstations. They received their voucher and, after a short farewell in the hallway, left the building.

E.2 Details of Game Design

The survival game involves four players. Based on a majority vote, they have to decide on one out of three items to bring to the island. Once they reach a consensus, we present the next set of items. We instructed them that several items would be crucial for survival. Wanting to bring as many items as possible within the seven minutes they are allowed to play provided an incentive to reach an agreement quickly on each item. However, before commencing the game, each player was assigned a role undisclosed to the others. Three players are cooperative. Each received information about a different set of two items necessary for survival (six items in total.) We were inspired by Vertegaal and Ding [241], who also distributed information necessary for success among participants. In addition to these crucial items, we also assigned them a non-crucial item. Their secondary task was to convince the group to choose it at least once. They win the game if, at the moment time runs out, the group has chosen this item at least once in addition to all other items necessary for survival. The fourth player is uncooperative with the goal of jeopardising the survival plan and is given a lot of information about crucial items. In addition, we made them aware of a doom item. The uncooperative player wins if the group is convinced to take this item or fails to select all necessary items.

We chose this game design because the screen needed to show very little information, allowing players to still focus on their interlocutors. They were incentivised to do so since facial expressions might reveal who the uncooperative player is. At the same time, every player would eventually suggest an item non-crucial for survival to win the game, adding distrust to the social dynamics. We decided against using the survival of the team as a performance measure. Such measures are typically only sensitive to substantial manipulations of the experimental factors [153, 242], which we did not expect.

E.3 Questionnaire Items and Statistical Details

We provide a collection of Figures (E.1–E.4), which report additional details on the measurements obtained during the study. All Figures report significance values accompanied

Attribute pairs		Gazing Heads			Tiled View		
Left (−3)	Right (3)	Q1	Q2	Q3	Q1	Q2	Q3
Unsociable	Sociable	1	2	3	0	1	2
Impersonal	Personal	1	2	3	-1	1	2
Insensitive	Sensitive	1	1	2	0	1	2
Cold	Warm	1	2	2	0	1	2
Social presence		1	2	2	0	1	2

Figure E.1. Detailed results of the semantic differential: Questions were based on the smallest adequate set to measure social presence. These four attribute pairs were suggested by Short et al. [206]. The social presence is significantly higher for Gazing Heads ($p = 0.020$).

by additional information. Figure E.1–E.3 show the questionnaire items, together with a record of prior publications from which they were derived. Figure E.1-E.2 report the quantiles (Q1, Q3) and median (Q2) as descriptive statistics, while Figure E.3 shows what share of participants have preferred which system. Note that the UX questionnaire (Figure E.2) included statements assessing participants' perception and interpretation of three types of gazes: direct eye contact, third-party gazes, and "off-gazes" directed at no one. For each of those three gaze types, three statements were included. They are shown in the bottom section of the table. One asked whether such gazes were perceivable; another asked whether they helped to notice interlocutors' attention; and a last one inquired if they clarified who was addressed. Figure E.3 shows the results of the speech and gaze analysis previously presented in Figure 6.6 on page 101. The figure here shows different descriptive statistics that help understand the factorial repeated measures ANOVA results. We also show the F-value of the significance test.

		Gazing Heads			Tiled View			
Question (adopted from source)	measures	Q1	Q2	Q3	Q1	Q2	Q3	p-Value
It was exciting to follow the discussion [201]	Engagement	1.5	2	3	1	2	2	**0.044**
Turn-taking was difficult [23] *	Turn-Tak.	1*	2*	2*	0*	1*	2*	0.084
I was able to take control of the conversation when I wanted to [201]	Turn-Tak.	1	2	3	1	2	2	0.176
The conversation was highly interactive [201]	Engagement	1	2	3	0	2	2	<0.001
Sometimes I had the feeling I was excluded from the conversation *	UX Issue	1*	2*	3*	1*	2*	2*	0.790
I could not contribute anything to the solution we came up with [86]	Satisfaction	2*	2*	3*	1*	2*	3*	0.246
The system was distracting me from the conversation *	UX Issue	-.5*	1*	2*	2*	2*	3*	**<0.001**
During the experiment I had the feeling we were all in the same room [86]	Virtual Presence	-.5	1	2	-2	-1	1	**<0.001**
One does not get a good enough idea of how people at the other end are reacting [37] *	Social Presence	-1*	1*	2*	-1*	1*	2*	0.263
I couldn't get to know people very well if I only met them over this system [37] *	Social Presence	-1*	1*	2*	-1*	0*	2*	**0.010**
I was always aware of my partner's presence [86]	Social Presence	1	2	3	1	2	2	**0.002**
It was easy for me to notice when my conversation partners looked at me [86, 23]	Eye Contact	1	2	3	-2	-1	0	**<0.001**
I knew when I was being addressed by someone [23]	Being Addressed	1	2	3	-1	1	2	**<0.001**
I knew when someone was listening to me or paying attention to me [201, 23]	Being Attended to	2	2	3	-1	1	2	**<0.001**
It was easy to notice when my conversation partners looked at someone else (other than me) [86]	3rd-party Gaze	1	2	3	-2	-1	0	**<0.001**
I knew when I was not addressed (but instead someone else was) [23]	Addressed others	1	2	2	-1	0	1	**<0.001**
I knew when someone was listening or paying attention to someone else (other than me) [201]	Attended others	1	2	3	-1	0	1.5	**<0.001**
It was easy to notice when my conversation partners looked at no one (not at me or anybody)	Off Gaze	-2	0	1	-2	-1	1	0.708
I knew when someone was not following the conversation, was thinking about something or became distracted [86]	No Attention	0	1	1	0	1	2	0.289
I knew when someone was thinking about something [86]	Thinking	-1	1	1	-1	0	1	0.993

Figure E.2. Questions and detailed results of the UX questionnaire: Quantiles (Q1, Q2 Q3) and p-Values of participants ratings of 20 statements on a 7-point Likert scale ranging from "strongly disagree" (−3) to "strongly agree" (3). Note: * indicates that the rating had been multiplied with -1 to account for the inverse phrasing.

Question (adopted from source)	measures	Chose GH	Both Equal	Chose TV	p-Value
Which system would you recommend to your friends and colleagues?	Satisfaction	49%	19%	32%	**0.052**
For which system was turn-taking easier? [23]	Turn-Tak.	51%	30%	19%	$<$ **0.001**
Which system facilitated a more natural interaction with your conversation partners? [252]	Social Presence & Satisfaction	62%	19%	19%	$<$ **0.001**
With which system was the interaction more engaging/exciting? [201]	Engagement	87%	9%	4%	$<$ **0.001**
Which system was better for noticing if your conversation partners were paying attention to you or someone/something else? [201, 23]	Non-verb. Cues	91%	6%	3%	$<$ **0.001**
With which system was the interaction more social?	Social Presence	62%	25%	13%	$<$ **0.001**
Which system would you choose for a meeting where you intend to persuade other people? [37]	Social Presence	50%	22%	28%	**0.015**
Which system would you like to use for the interview?	Satisfaction	72%	8%	20%	$<$ **0.001**

Figure E.3. Detailed results of the comparative questionnaire: Participants' preferences regarding the two systems.

	Factor: System			Factor: Task		
	GH	TV	F and p-Value	Discuss	Game	F and p-Value
Turn Frequency per Minute	4.50 (1.1)	4.54 (0.7)	$F(1,18) = 0.028$ *$p = 0.869$*	2.72 (1.1)	6.32 (1.1)	$F(1,18) = 291.247$ **$p < 0.001$**
Turn Duration	16.26s (10.9)	15.36 (10.7)	$F(1,18) = 0.400$ *$p = 0.535$*	23.73 (10.1)	7.89 (1.9)	$F(1,18) = 82.083$ **$p < 0.001$**
Group Turn Freq./Minute	0.82 (0.8)	1.00 (1.0)	$F(1,18) = 5.046$ **$p = 0.037$**	0.23 (0.2)	1.60 (0.7)	$F(1,18) = 123.869$ **$p < 0.001$**
Turn Distribution (H)	1.93 (0.1)	1.93 (0.1)	$F(1,18) = 0.255$ *$p = 0.620$*	1.90 (0.1)	1.96 (0.0)	$F(1,18) = 14.431$ **$p = 0.001$**
Time one Person spoke	81.72% (10.3)	80.62% (10.9)	$F(1,18) = 1.140$ *$p = 0.300$*	90.40% (3.8)	72.0% (5.9)	$F(1,18) = 242.338$ **$p < 0.001$**
Simultaneous Speech	6.78% (6.1)	8.23% (6.9)	$F(1,18) = 3.949073$ *$p = 0.062$*	2.90% (2.7)	12.12% (6.0)	$F(1,18) = 101.303$ **$p < 0.001$**
Non-Int. Simult. Speech	6.64% (5.7)	7.50% (5.8)	$F(1,18) = 1.883$ *$p = 0.187$*	2.76% (2.2)	11.39% (4.9)	$F(1,18) = 135.425$ **$p < 0.001$**
Interruptive Simult. Speech	2.48% (2.2)	2.83% (2.4)	$F(1,18) = 1.263$ *$p = 0.276$*	1.04% (1.2)	4.28% (2.0)	$F(1,18) = 157.325$ **$p < 0.001$**
Sim. Speech Taking Control	31.08% (16.1)	26.67% (11.6)	$F(1,18) = 1.677$ *$p = 0.212$*	26.85% (18.4)	30.90% (7.7)	$F(1,18) = 2.771$ *$p = 0.113$*
Speaker Switches Overlaps	34.34% (17.2)	34.96 (17.0)	$F(1,18) = 0.059$ *$p = 0.810$*	24.65% (16.3)	44.64% (10.7)	$F(1,18) = 67.024$ **$p < 0.001$**
Switching Time	0.59s (0.8)	0.42s (0.6)	$F(1,18) = 2.037$ *$p = 0.171$*	0.85s (0.8)	0.16s (0.4)	$F(1,18) = 37.438$ **$p < 0.001$**
Focus Changes per Minute	58.76 (22.0)	60.42 (22.7)	$F(1,75) = 3.874$ *$p = 0.053$*	58.29 (23.4)	60.88 (21.2)	$F(1,75) = 2.632$ *$p = 0.109$*
Eye Contact per Minute	22.15 (13.6)	20.67 (11.8)	$F(1,75) = 7.480$ **$p = 0.008$**	29.46 (12.1)	13.36 (6.9)	$F(1,75) = 195.247$ **$p < 0.001$**
Eye Contact (% of Session)	18.9% (12.3)	17.6% (11.3)	$F(1,75) = 4.630$ **$p = 0.036$**	26.88% (10.32)	9.69% (5.12)	$F(1,75) = 304.942$ **$p < 0.001$**
Eye Contact Duration	0.52s (0.2)	0.51s (0.2)	$F(1,75) = 0.365$ *$p = 0.365$*	0.58s (0.2)	0.46s (0.1)	$F(1,75) = 83.649$ **$p < 0.001$**

Figure E.4. Speech and eye-gaze analysis. The first three columns compare our two systems, while averaging over the two tasks. The remaining three columns compare tasks, while averaging over the two systems. Note that one session was interrupted due to network problems. Hence, it needed to be excluded from the voice and eye-tracking analysis.

E.4 Visual Attention Analysis

We analysed what proportion of a session participants spent looking at the different focus areas. The focus area was added as the third predictor to the repeated measures ANOVA. Since Mauchly's test indicated that the assumption of sphericity had been violated for this predictor and all its interactions, the degrees of freedom were Greenhouse-Geisser corrected. Interaction effects were analysed using a Bonferroni post-hoc test. As expected, the proportion of time spent gazing was different among focus areas ($F(3.21, 241.09) = 169.495$, $p < 0.001$). More importantly, there was an interaction effect between the task participants worked on and the proportion of time they gazed at different focus areas ($F(3.21, 241) = 415$, $p < 0.001$). As shown in Figure E.6, for the discussion task, participants spend most of their time looking at the centre interlocutor ($M = 32.1\%$, $SD = 12.2\%$). The second and third largest proportions of time were spent looking at the left ($M = 25.9\%$, $SD = 11.8\%$) and the right interlocutor ($M = 24.0\%$, $SD = 10.3\%$) with no significant difference between them. The fourth largest proportion of time was spent looking off-screen ($M = 14.6\%$, $SD = 12.2\%$). The significantly smallest amount of time was spent looking at the empty content area ($M = 1.42\%$, $SD = 3.04\%$) and the remaining screen ($M = 0.98\%$, $SD = 1.21\%$) with no significant difference between them.

For the game, the main difference is that participants looked at the content area ($M = 46.2\%$, $SD = 15.1\%$) significantly more than any other focus area, which is unsurprising since it contained game-relevant information. The second, third and fourth largest proportion of time was spent looking at the centre ($M = 18.9\%$, $SD = 8.31\%$),
left ($M = 14.0\%$, $SD = 6.59\%$) and right interlocutor ($M = 13.4\%$, $SD = 6.79\%$). Again the centre interlocutor was gazed at significantly more than the left and right one. The left and right interlocutors received similar gaze times.
Participants looked off-screen ($M = 5.60\%$, $SD = 5.30\%$) significantly less than any face or the content areas but significantly more frequently than on the remaining screen showing no content ($M = 0.83\%$, $SD = 0.67\%$).

As a result the frequency of mutual eye was significantly ($F = 195$, $p = 0.08$) reduced by 54.6% during the game ($M = 13.380 min^{-1}$) in comparison to the discussion ($M = 29.460 min^{-1}$). The duration of mutual eye contact was also significantly ($F = 83.6$, $p < 0.001$) reduced by 20% during the game ($M = 0.458s$, $SD = 0.144s$) in comparison to the discussion ($M = 0.577s$, $SD = 0.193s$). The overall share of session spend gazing with mutual gazes was significantly ($F = 305$, $p < 0.001$) lower the game ($M = 9.69\%$, $SD = 5.12\%$) in comparison to the discussion ($M = 26.9\%$, $SD = 10.3\%$).

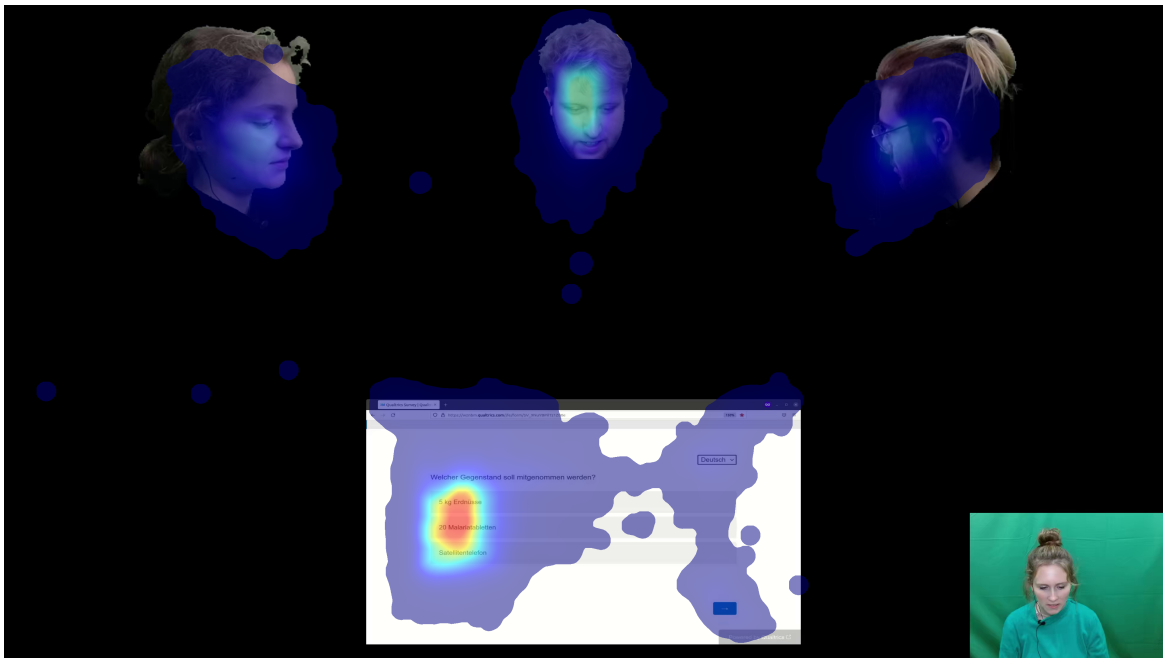


Figure E.5. Heatmap of visual attention: In the game task, the content area at the bottom of the screen received significantly more attention than the faces or any other area.

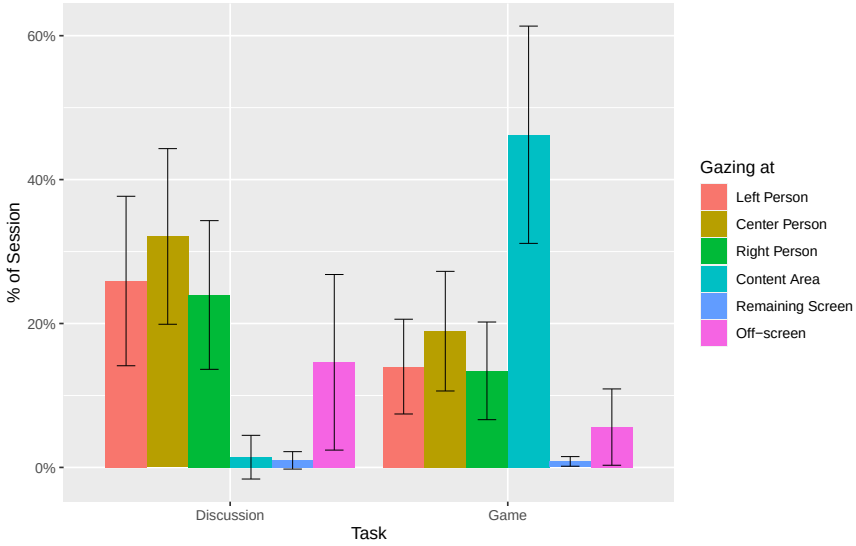


Figure E.6. Distribution of visual attention: The proportion of session time participants spent gazing at the different focus areas was distributed differently during the game and the discussion.

E.5 Thematic Analysis

For our analysis of participants' experiences with the Gazing Heads reported in group interviews, we chose thematic analysis to help us capture both the subjective experience and technical feedback. Our chosen method is thematic analysis, a widely used qualitative HCI research tool [29]. It is particularly valuable when one seeks to understand user experiences, perceptions, and behaviours. It allows us to identify, analyse, and report patterns (themes) within data, making it a practical and applicable tool for qualitative research. Figure E.7 outlines the domains. This is followed by Figure E.8, which lists the topics within those domains. Finally, Figure E.9 lists all codes with which interviews were coded.

Domain	**Summary**
Technical Feedback (TF)	This domain encompasses user reflections on the system's technical features, detailing positive and negative impacts on their experience. It also incorporates user suggestions for improvements and desired changes to system functionalities.
Experiential Feedback (EF)	This domain captures participants' subjective experiences and feelings as they interacted with the system during the experiment. In contrast to purely technical feedback, this domain focuses on how the system affected users on an emotional and experiential level.
Adoption Potential (AP)	This domain captures participants' reflections on the system's future potential, including their willingness to adopt it and its novelty compared to existing technologies. These reflections move beyond the immediate experimental experience and focus on forward-looking scenarios where the system could be useful and those where it may not be the right fit.

Figure E.7. Domain overview: The three domains used in the thematic analysis and their description.

Domain	Topic	Description
TF	Transitions	Visual perception of transitions compared to a normal conversation.
	Background	Visual background of the video conference system.
	Shoulders	Effects of the upper body being visible.
	Spatial Audio	Effects of Spatial Audio.
	Mirror View	Reflection on not having a mirror view.
EF	Immersion	Visual aspects of immersion and the same-room experience.
	Social Presence	Reduction of social distance and higher social engagement.
	Floating Heads	Social and visual effects of seeing only floating heads and no bodies.
	Turn Taking	Difference in turn-taking difficulty.
AP	Novelty	Novelty aspects and willingness to adopt the system.
	Use Cases	Future use case scenarios of the system.

Figure E.8. Topic overview: The 11 topics within the three domain and their descriptions.

Theme	Code	Description
Transitions	t_slow	Dwell time and/or animation to slow.
	t_distracting	Transitions are distracting.
	t_mature	Transitions are perceived as animations, not a head rotation.
	t_jittering	Issues with heads jittering due to tracking issues.
Background	bg_request	Suggestions for different background.
	not_bg_request	Request to leave background unchanged.
	bg_no_black	Expression of dislike for the black background.
Shoulder	shoulder_natural	Shoulders have a positive influence on naturalness, 3D illusion and immersion.
	not_shoulder_natural	Shoulders have a negative influence on naturalness and immersion.
	shoulder_worse_illusion	Shoulders have a negative influence 3D illusion.
	shoulder_hybrid	Shoulders make gaze-aware video conferencing feel more familiar.
Spatial Audio	audio_quality	Great audio quality.
	audio_spatial	Spatial audio recognised.
Mirror View	m_request	Request to add mirror view as a feature.
	not_m_request	Statements that mirror view is not needed.
	m_distract	Mirror views are or may be distracting.
	m_immersive	Increased immersion by omission of mirror view.
	m_control	Mirror view needed for control over own appearance.
Immersion	i_together_same_room	Feeling of being in the same room.
	i_circle	Noticing or commenting on the circular arrangement of users.
	i_arrangement	Noticing the circular arrangement is consistent.
	i_bg_seperate	Unified background increases immersion.
	i_not	Experience is not more immersive than normal video conferencing.
	i_table	Request to add a table to the scene.
Social Presence	sp_intimate	Experience is more personal.
	sp_social	Experience is more social.
	sp_engagement	Experience is more (socially) engaging.
	sp_presence	Feeling of increased presence of oneself and/or others.
Floating Heads	fh_reductive	Floating heads are too reductive.
	fh_reductive_mimic	Floating heads increase focus on mimic.
	fh_none_verbal_missing	Missing verbal cues from the upper body.
	fh_3D_illusion	Floating heads benefit 3D illusion.
	fh_unnatrual	Floating heads look unnatural.
Turn Taking	turn_taking	Perception of easier turn taking.
Novelty	new_not_used_to	Feeling unfamiliar with the system and needing time to get used to it.
	new_attoptation_willingness	Considerations of willingness to adopt the system.
	new_adoptation_easy	Opinion that the system is easy to adapt to.
	not_new_adoptation_easy	Opinion that system is hard to adapt to.
	new_hardware	Concerns about necessary hardware.
Use Cases	exploiting_potential	Reflections about use case which would benefit the most from this system.
	uc_leisure	Anticipated usage for leisure.
	not_uc_leisure	Anticipated non-usage for leisure.
	uc_formal	Anticipated usage for professional purposes.
	uc_friends	Anticipated usage for meeting friends and or family.
	uc_workgroup	Anticipated usage for small work group meetings.
	uc_presentation	Anticipated usage for presentations.
	uc_more_participants	Concerns about the usefulness with more than six participants.

Figure E.9. Coding scheme: Two authors developed and refined 47 codes collaboratively through multiple rounds of open coding, synchronisation, and closed coding.

E.6 Eye-Tracking Accuracy

We tracked participants' gaze using a Tobii Eye Tracker 5. It uses a near-infrared and colour sensor operating at an interlaced sampling rate of 133Hz and a non-interlaced rate of 33Hz. The maximum supported field of view in each direction is 40 degrees. It can recover lost gaze using neural head tracking. Gibaldi et al. [76] tested an older version of the Tobii eye tracker. They found it has at least an accuracy of 0.6° and an end-to-end latency of around 47 $\pm$ 4 ms when receiving the data via a local UDP connection. We found these specifications suitable for our design, and since the gaze elements in question were rather large, we encountered no issues with the eye-tracker's accuracy.

For the implementation of our concept with web cam eye tracking, we calculated a minimum accuracy using triangulation and the smallest margin for error in our interface. A gaze focused on centre interlocutor (Figure 6.5) can horizontally deviate by 6.75 cm on a 24 inch desk screen, and by 4.25 cm on 13 inch laptop before it leads to an error. Assuming a viewing distance of 70 cm for the desk monitor, and 50 cm for the laptop screen the required accuracy is $5.51°$ for the desktop, and $4.86°$ for the laptop.

E.7 Latency Measurement

For video conferencing, latency between two clients can be described as the total time it takes captured data from Client A to be processed and sent over the network to be represented on Client B. Hence, the total latency consists of the input latency (camera/microphone), processing time (segmentation, view stabilisation, camera transitions, compression), network delay, post-processing time (decompression, visualisation), and output delay (screen, headphones). We used an estimation measure to obtain our system's auditory and visual latency. We ran all clients in their standard configuration to simulate a realistic load situation. We arranged two clients such that their microphone and one of their cameras simultaneously captured a lab member clapping their hands. We modified one client to store the received data from the sender jointly with the unprocessed data captured from the input devices. This way, we could determine the differences in frames on the receiver between the raw input and the received data from the sender. Dividing the difference by the respective frame–rate yields the latency in seconds introduced by all processing and the network delay. It ignores the delay introduced by the input and output devices which is marginal.

Bibliography

[1] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y. Lim, and Mohan Kankanhalli. 2018. Trends and Trajectories for Explainable, Accountable and Intelligible Systems: An HCI Research Agenda. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/3173574.3174156.

[2] John P. Abraham, Brian D. Plourde, and Lijing Cheng. 2020. Using heat to kill SARS-CoV-2. *Reviews in Medical Virology*, 30, 5. doi: 10.1002/rmv.2115.

[3] A. Adadi and M. Berrada. 2018. Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052.

[4] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. 2018. Sanity Checks for Saliency Maps. In *Advances in Neural Information Processing Systems* (NeurIPS), 9505–9515. doi: 10.48550/arXiv.1810.03292.

[5] Julius Adebayo, Michael Muelly, Ilaria Liccardi, and Been Kim. 2020. Debugging Tests for Model Explanations. In *Advances in Neural Information Processing Systems* (NeurIPS). Vol. 34, 700–712. doi: 10.48550/arXiv.2011.05429.

[6] Alper T. Alan, Enrico Costanza, Sarvapali D. Ramchurn, Joel Fischer, Tom Rodden, and Nicholas R. Jennings. 2016. Tariff Agent: Interacting with a Future Smart Energy System at Home. *ACM Transactions on Computer-Human Interaction*. TOCHI 23, 4. doi: 10.1145/2943770.

[7] Ahmed Alqaraawi, Martin Schuessler, Philipp Weiß, Enrico Costanza, and Nadia Berthouze. 2020. Evaluating Saliency Map Explanations for Convolutional Neural Networks: A User Study. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (IUI). ACM, 263–274. doi: 10.1145/3377325.3377519.

[8] Christopher Anders, Plamen Pasliev, Ann-Kathrin Dombrowski, Klaus-Robert Müller, and Pan Kessel. 2020. Fairwashing Explanations with Off-Manifold Detergent. In *Proceedings of the 37th International Conference on Machine Learning* (ICML). Vol. 119, 314–323. doi: 10.48550/arXiv.2007.09969.

[9] Michael Argyle and Mark Cook. 1976. *Gaze and mutual gaze*. Cambridge University Press, Cambridge, UK. isbn: 978-0-521-20865-9.

[10] Michael Argyle and Jean A. Graham. 1976. The central Europe experiment: Looking at persons and looking at objects. *Environmental Psychology & Nonverbal Behavior*, 1, 1, 6–16. doi: 10.1007/BF01115461.

[11] Leila Arras, Ahmed Osman, and Wojciech Samek. 2021. CLEVR-XAI: A benchmark dataset for the ground truth evaluation of neural network explanations. *Information Fusion*, 14–40. doi: 10.1016/j.inffus.2021.11.008.

[12] Sebastian Bach, Alexander Binder, Grégoire Montavon, Frederick Klauschen, Klaus-Robert Müller, and Wojciech Samek. 2015. On Pixel-Wise Explanations for Non-Linear Classifier Decisions by Layer-Wise Relevance Propagation. *PLOS ONE*, 10, 7. doi: 10.1371/journal.pone.0130140.

[13] Jeremy N. Bailenson. 2021. Nonverbal overload: A theoretical argument for the causes of Zoom fatigue. *Technology, Mind, and Behavior*, 2, 1. doi: 10.1037/tmb0000030.

[14] David Bau, Bolei Zhou, Aditya Khosla, Aude Oliva, and Antonio Torralba. 2017. Network Dissection: Quantifying Interpretability of Deep Visual Representations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 6541–6549. doi: 10.1109/CVPR.2017.354.

[15] Christian F Baumgartner, Lisa M Koch, Kerem Can Tezcan, Jia Xi Ang, and Ender Konukoglu. 2018. Visual Feature Attribution Using Wasserstein GANs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 8309–8319. doi: 10.1109/CVPR.2018.00867.

[16] Piraye Bayman and Richard E. Mayer. 1984. Instructional Manipulation of Users' Mental Models for Electronic Calculators. *International Journal of Man-Machine Studies*, 20, 2, 189–199. doi: 10.1016/S0020-7373(84)80017-6.

[17] Gary Bente, Sabine Rüggenberg, Nicole C. Krämer, and Felix Eschenburg. 2008. Avatar-Mediated Networking: Increasing Social Presence and Interpersonal Trust in Net-Based Collaborations. *Human Communication Research*, 34, 2, 287–318. doi: 10.1111/j.1468-2958.2008.00322.x.

[18] Nienke Martine Bierhuizen, Wendy Powell, Tina Mioch, Omar Niamut, and Hans Stokking. 2022. Influence of Photorealism and Non-Photorealism on Connection in Social VR. In *Annual Review of Cybertherapy and Telemedicine*, 79–84. https://www.interactivemediainstitute.com/cypsy25/.

[19] Reuben Binns, Max Van Kleek, Michael Veale, Ulrik Lyngs, Jun Zhao, and Nigel Shadbolt. 2018. 'it's reducing a human being to a percentage': perceptions of justice in algorithmic decisions. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/3173574.3173951.

[20] Or Biran and Courtenay Cotton. 2017. Explanation and justification in machine learning: A survey. In *Explainable Artificial Intelligence – Workshop* (IJCAI). Vol. 8.

[21] Christopher M Bishop. 2006. *Pattern Recognition and Machine Learning*. *Information Science and Statistics*. Springer, New York, USA. isbn: 978-0-387-31073-2.

[22] G. Blakowski and R. Steinmetz. 1996. A media synchronization survey: Reference model, specification, and case studies. *IEEE Journal on Selected Areas in Communications*, 14, 5–35, 1. doi: 10.1109/49.481691.

[23] Martin Böcker and Lothar Mühlbach. 1993. Communicative Presence in Videocommunications. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* number 3. Vol. 37, 249–253. doi: 10.1177/154193129303700308.

[24] Judith Borowski. 2022. *Toward Understanding Visual Perception in Machines with Human Psychophysics*. Ph.D. Dissertation. Universität Tübingen. doi: 10.15496/publikation-74982.

[25] Judy Borowski, Roland S. Zimmermann, Judith Schepers, Robert Geirhos, Thomas S. A. Wallis, Matthias Bethge, and Wieland Brendel. 2021. Exemplary Natural Images Explain CNN Activations Better than State-of-the-Art Feature Visualization. In *Proceedings of the International Conference on Learning Representations* (ICLR). doi: 10.48550/arXiv.2010.12606.

[26] Benedict Du Boulay, Tim O'shea, and John Monk. 1999. The Black Box inside the Glass Box: Presenting Computing Concepts to Novices. *International Journal of Human-Computer Studies*, 51, 2, 265–277. doi: 10.1006/ijhc.1981.0309.

[27] Wieland Brendel and Matthias Bethge. 2018. Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. In *Proceedings of the International Conference on Learning Representations* (ICLR). doi: 10.48550/arXiv.1904.00760.

[28] Joel Bruckstein and Bob Veres. 2022. T3/Inside Information Advisor Software Survey. Tech. rep. Chap. Videoconferencing Tools/Services, 67. Retrieved July 25, 2024 from https://t3technologyhub.com/wp-content/uploads/2024/01/2024-T3_Inside-Information-Software-Survey.pdf.

[29] Emeline Brulé. 2020. Thematic Analysis in HCI. Accessed: 2022-09-13. (2020). doi: 10.58079/uaah.

[30] ITU-R Recommendation BT.1359-1. 1998. Relative timing of sound and vision for broadcasting. In https://www.itu.int/rec/R-REC-BT.1359-1-199811-I/en.

[31] Zana Buçinca, Phoebe Lin, Krzysztof Z. Gajos, and Elena L. Glassman. 2020. Proxy Tasks and Subjective Measures Can Be Misleading in Evaluating Explainable AI Systems. In *Proceedings of the 25th International Conference on Intelligent User Interfaces* (IUI), 454–464. doi: 10.1145/3377325.3377498.

[32] A. Bussone, S. Stumpf, and D. O'Sullivan. 2015. The Role of Explanations on Trust and Reliance in Clinical Decision Support Systems. In *Proceedings of the International Conference on Healthcare Informatics* (ICHI), 160–169. doi: 10.1109/ICHI.2015.26.

[33] Carrie J. Cai, Jonas Jongejan, and Jess Holbrook. 2019. The Effects of Example-based Explanations in a Machine Learning Interface. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI), 258–262. doi: 10.1145/3301275.3302289.

[34] Alex Campolo, Madelyn Sanfilippo, Meredith Whittaker, and Kate Crawford. 2017. AI Now 2017 Report. (2017). https://ainowinstitute.org/wp-content/uploads/2023/04/AI_Now_2017_Report.pdf.

[35] Chen Cao et al. 2022. Authentic Volumetric Avatars from a Phone Scan. *ACM Transactions on Graphics*. TOG 41, 4. doi: 10.1145/3528223.3530143.

[36] Valerie Caproni, Douglas Levine, Edgar O'neal, Peter McDonald, and Gray Garwood. 1977. Seating position, instructor's eye contact availability, and student participation in a small seminar. *The Journal of Social Psychology*, 103, 2, 315–316. doi: doi/10.1080/00224545.1977.9713335.

[37] Brain G. Champness. 1973. The assessment of user reactions to confravision: I. Design of the questionnaire. Tech. rep. E/73129/C. Communication Studies Group.

[38] Chun-Hao Chang, Elliot Creager, Anna Goldenberg, and D. Duvenaud. 2019. Explaining Image Classifiers by Counterfactual Generation. In *Proceedings of the International Conference on Learning Representations* (ICLR). doi: 10.48550/arXiv.1807.08024.

[39] Chaofan Chen, Oscar Li, Daniel Tao, Alina Barnett, Jonathan Su, and Cynthia Rudin. 2019. This Looks Like That: Deep Learning for Interpretable Image Recognition. In *Advances in Neural Information Processing Systems* (NeurIPS). Vol. 32, 801–812. doi: 10.48550/arXiv.1806.10574.

[40] Michael Chromik and Martin Schuessler. 2020. A Taxonomy for Human Subject Evaluation of Black-Box Explanations in XAI. In *Proceedings of the IUI workshop on Explainable Smart Systems and Algorithmic Transparency in Emerging Technologies* (IUI). Vol. 2582. http://ceur-ws.org/Vol-2582/paper9.pdf.

[41] Eric Chu, Deb Roy, and Jacob Andreas. 2020. Are Visual Explanations Useful? A Case Study in Model-in-the-Loop Prediction. *arXiv preprint*. doi: 10.48550/arXiv.2007.12248.

[42] Andy Cockburn, Carl Gutwin, and Alan Dix. 2018. HARK No More: On the Preregistration of CHI Experiments. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/3173574.3173715.

[43] Andy Crabtree, Lachlan Urquhart, and Jiahong Chen. 2019. Right to an Explanation Considered Harmful. SSRN Scholarly Paper ID 3384790. Social Science Research Network (SSRN). doi: 10.2139/ssrn.3384790.

[44] Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. 2008. The Effects of Transparency on Trust in and Acceptance of a Content-based Art Recommender. *User Modeling and User-Adapted Interaction*. UMUAI 18, 5, 455–496. doi: 10.1007/s11257-008-9051-3.

[45] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. ImageNet: A Large-Scale Hierarchical Image Database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 248–255. doi: 10.1109/CVPR.2009.5206848.

[46] Anind K. Dey and Alan Newberger. 2009. Support for Context-aware Intelligibility and Control. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 859–868. doi: 10.1145/1518701.1518832.

[47] Laurent Dinh, David Krueger, and Yoshua Bengio. 2015. NICE: Non-linear Independent Components Estimation. In *Workshop Track* (ICLR). doi: 10.48550/arXiv.1410.8516.

[48] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. 2017. Density Estimation Using Real NVP. In *Proceedings of the International Conference on Learning Representations* (ICLR). doi: 10.48550/arXiv.1605.08803.

[49] Alan Dix. 1992. Human Issues in the Use of Pattern Recognition Techniques. In *Neural Networks and Pattern Recognition in Human-Computer Interaction*. Ellis Horwood, 429–451. isbn: 978-0-13-626995-3. eprint: https://alandix.com/academic/papers/neuro92/.

[50] Jonathan Dodge, Q. Vera Liao, Yunfeng Zhang, Rachel K. E. Bellamy, and Casey Dugan. 2019. Explaining Models: An Empirical Study of How Explanations Impact Fairness Judgment. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI), 275–285. doi: 10.1145/3301275.3302310.

[51] Ann-Kathrin Dombrowski, Maximillian Alber, Christopher Anders, Marcel Ackermann, Klaus-Robert Müller, and Pan Kessel. 2019. Explanations can be manipulated and geometry is to blame. In *Advances in Neural Information Processing Systems* (NeurIPS), 13589–13600. doi: 10.48550/arXiv.1906.07983.

[52] Finale Doshi-Velez and Been Kim. 2017. Towards A Rigorous Science of Interpretable Machine Learning. *arXiv preprint*. doi: 10.48550/arXiv.1702.08608.

[53] Filip Karlo Došilović, Mario Brčić, and Nikica Hlupić. 2018. Explainable artificial intelligence: A survey. In *Proceedings of the 41st International Convention on Information and Communication Technology, Electronics and Microelectronics* (MIPRO), 0210–0215. doi: 10.23919/MIPRO.2018.8400040.

[54] Starkey Duncan. 1972. Some Signals and Rules for Taking Speaking Turns in Conversations. *Journal of Personality and Social Psychology*, 23, 2, 283–292. doi: 10.1037/h0033031.

[55] Starkey Duncan Jr, George Niederehe, Starkey Duncan, and George Niederehe. 1974. On signalling that it's your turn to speak. *Journal of Experimental Social Psychology*, 10, 3, 234–247. doi: 10.1016/0022-1031(74)90070-5.

[56] Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. 2003. The Role of Trust in Automation Reliance. *International Journal of Human-Computer Studies*, 58, 6, 697–718. doi: 10.1016/S1071-5819(03)00038-7.

[57] Upol Ehsan, Pradyumna Tambwekar, Larry Chan, Brent Harrison, and Mark O. Riedl. 2019. Automated rationale generation: a technique for explainable ai and its effects on human perceptions. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI), 263–274. doi: 10.1145/3301275.3302316.

[58] Viktor Eisenstadt, Christian Espinoza-Stapelfeld, Ada Mikyas, and Klaus-Dieter Althoff. 2018. Explainable Distributed Case-Based Support Systems: Patterns for Enhancement and Validation of Design Recommendations. In *Proceedings of the International Conference on Case-Based Reasoning* (ICCBR), 78–94. doi: 10.1007/978-3-030-01081-2_6.

[59] Nadia El Bekri, Jasmin Kling, and Marco F. Huber. 2019. A Study on Trust in Black Box Models and Post-hoc Explanations. In *Proceedings of the 14th International Conference on Soft Computing Models in Industrial and Environmental Applications* (SOCO), 35–46. doi: 10.1007/978-3-030-20055-8_4.

[60] Radwa ElShawi, Youssef Sherif, Mouaz Al□Mallah, and Sherif Sakr. 2020. Interpretability in healthcare: A comparative study of local machine learning interpretability techniques. *Computational Intelligence*, 37, 4, 1633–1650. doi: 10.1111/coin.12410.

[61] Hartmut Esser. 1986. Über die Teilnahme an Befragungen. de. *ZUMA Nachrichten*, 10, 18, 38–47. https://www.ssoar.info/ssoar/handle/document/21030.

[62] Patrick Esser, Robin Rombach, and Bjorn Ommer. 2020. A Disentangling Invertible Interpretation Network for Explaining Latent Representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 9223–9232. doi: 10.1109/CVPR42600.2020.00924.

[63] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2007. The PASCAL Visual Object Classes Challenge 2007 (VOC2007). http://host.robots.ox.ac.uk/pascal/VOC/voc2007/.

[64] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. 2012. The PASCAL Visual Object Classes Challenge 2012 (VOC2012). http://host.robots.ox.ac.uk/pascal/VOC/voc2012/.

[65] Li Fei-Fei, Asha Iyer, Christof Koch, and Pietro Perona. 2007. What do we perceive in a glance of a real-world scene? *Journal of Vision*, 7, 1. doi: 10.1167/7.1.10.

[66] Shi Feng and Jordan Boyd-Graber. 2019. What Can AI Do for Me?: Evaluating Machine Learning Interpretations in Cooperative Play. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI), 229–239. doi: 10.1145/3301275.3302265.

[67] Andy P. Field, Jeremy Miles, and Zoë Field. 2012. Sage, London, UK. isbn: 978-1-4462-0046-9.

[68] Ruth Fong, Mandela Patrick, and Andrea Vedaldi. 2019. Understanding Deep Networks via Extremal Perturbations and Smooth Masks. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (ICCV). doi: 10.1109/ICCV.2019.00304.

[69] ITU-T Recommendation G.114. 2003. One-way transmission time, 2–3. https://www.itu.int/rec/T-REC-G.114-200305-I/en.

[70] Guy Gafni, Justus Thies, Michael Zollhofer, and Matthias Nießner. 2021. Dynamic Neural Radiance Fields for Monocular 4D Facial Avatar Reconstruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 8645–8654. doi: 10.1109/CVPR46437.2021.00854.

[71] Yaroslav Ganin, Daniil Kononenko, Diana Sungatullina, and Victor Lempitsky. 2016. Deepwarp: Photorealistic image resynthesis for gaze manipulation. In *Computer Vision* (ECCV), 311–326. doi: 10.1007/978-3-319-46475-6_20.

[72] Fatih Gedikli, Dietmar Jannach, and Mouzhi Ge. 2014. How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72, 4, 367–382. doi: 10.1016/j.ijhcs.2013.12.007.

[73] Joey George, Akmal Mirsadikov, Misty Nabors, and Kent Marett. 2022. What do users actually look at during 'zoom'meetings? Discovery research on attention, gender and distraction effects. In *Proceedings of the 55th Hawaii International Conference on System Sciences* (HICSS), 4779–4787. isbn: 978-0-9981331-5-7. https://hdl.handle.net/10125/79919.

[74] Amirata Ghorbani, Abubakar Abid, and James Zou. 2019. Interpretation of Neural Networks Is Fragile. In *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI) number 01. Vol. 33, 3681–3688. doi: 10.1609/aaai.v33i01.33013681.

[75] Amirata Ghorbani, James Wexler, James Y Zou, and Been Kim. 2019. Towards Automatic Concept-based Explanations. In *Advances in Neural Information Processing Systems* (NeurIPS), 9277–9286. doi: 10.48550/arXiv.1902.03129.

[76] Agostino Gibaldi, Mauricio Vanegas, Peter J. Bex, and Guido Maiello. 2017. Evaluation of the Tobii EyeX Eye tracking controller and Matlab toolkit for research. *Behavior Research Methods*, 49, 923–946, 3. doi: 10.3758/s13428-016-0762-9.

[77] Jon Gillick, Wesley Deng, Kimiko Ryokai, and David Bamman. 2021. Robust Laughter Detection in Noisy Environments. In *Proceedings of the Annual Conference of the International Speech Communication Association* (INTERSPEECH), 2481–2485. doi: 10.21437/Interspeech.2021-353.

[78] Leilani H Gilpin, David Bau, Ben Z Yuan, Ayesha Bajwa, Michael Specter, and Lalana Kagal. 2018. Explaining explanations: An overview of interpretability of machine learning. In *Proceedings of the IEEE 5th International Conference on data science and advanced analytics* (DSAA), 80–89. doi: 10.1109/DSAA.2018.00018.

[79] Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Counterfactual Visual Explanations. In *Proceedings of the 36th International Conference on Machine Learning* (ICML). Vol. 97, 2376–2384. doi: 10.48550/arXiv.1904.07451.

[80] Philip-William Grassal, Malte Prinzler, Titus Leistner, Carsten Rother, Matthias Nießner, and Justus Thies. 2022. Neural Head Avatars from Monocular RGB Videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 18653–18664. doi: 10.1109/CVPR52688.2022.01810.

[81] Sachin Grover, Chiara Pulice, Gerardo I. Simari, and V. S. Subrahmanian. 2019. BEEF: Balanced English Explanations of Forecasts. *IEEE Transactions on Computational Social Systems*, 6, 2, 350–364. doi: 10.1109/tcss.2019.2902490.

[82] Riccardo Guidotti, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. A Survey of Methods for Explaining Black Box Models. *ACM Computing Surveys*. CSUR 51, 5. doi: 10.1145/3236009.

[83] David Gunning and David Aha. 2019. DARPA's Explainable Artificial Intelligence (XAI) Program. *AI Magazine*, 40, 2, 44–58. doi: 10.1609/aimag.v40i2.2850.

[84] Jennifer X. Haensel, Tim J. Smith, and Atsushi Senju. 2022. Cultural differences in mutual gaze during face-to-face interactions: A dual head-mounted eye-tracking study. *Visual Cognition*, 30, 1-2, 100–115. doi: 10.1080/13506285.2021.1928354.

[85] Maaike Harbers, Karel van den Bosch, and John-Jules Meyer. 2010. Design and Evaluation of Explainable BDI Agents. In *Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology* (WI-IAT). doi: 10.1109/wi-iat.2010.115.

[86] Jörg Hauber, Holger Regenbrecht, Mark Billinghurst, and Andy Cockburn. 2006. Spatiality in Videoconferencing: Trade-offs between Efficiency and Social Presence. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW), 413–422. doi: 10.1145/1180875.1180937.

[87] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep Residual Learning for Image Recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 770–778. doi: 10.1109/CVPR.2016.90.

[88] Muchen He, Beibei Xiong, and Kaseya Xia. 2021. Are you looking at me? Eye gazing in web video conferences. Tech. rep. eprint: http://courses.ece.ubc.ca/518/previous/hit2020W/papers/group8_99750_14423175_ACM_Conference_Proceedings_Primary_Article_Template%20(2).pdf.

[89] Zhenyi He, Keru Wang, Brandon Yushan Feng, Ruofei Du, and Ken Perlin. 2021. GazeChat: Enhancing Virtual Conferences with Gaze-Aware 3D Photos. In *Proceedings of the 34th Annual Symposium on User Interface Software and Technology* (UIST), 769–782. doi: 10.1145/3472749.3474785.

[90] Melanie Heck, Christian Becker, and Viola Deutscher. 2023. Webcam Eye Tracking for Desktop and Mobile Devices: A Systematic Review. In *Proceedings of the Hawaii International Conference on System Sciences* (HICSS), 6820–6829. doi: 10.24251/HICSS.2023.825.

[91] Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW), 241–250. doi: 10.1145/358916.358995.

[92] Robert R Hoffman, Shane T Mueller, Gary Klein, and Jordan Litman. 2018. Metrics for explainable AI: Challenges and prospects. *arXiv preprint*. doi: 10.48550/arXiv.1812.04608.

[93] Fred Hohman, Andrew Head, Rich Caruana, Robert DeLine, and Steven M. Drucker. 2019. Gamut: A Design Probe to Understand How Data Scientists Understand Machine Learning Models. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/3290605.3300809.

[94] Chih-Fan Hsu, Yu-Shuen Wang, Chin-Laung Lei, and Kuan-Ta Chen. 2019. Look at Me! Correcting Eye Gaze in Live Video Communication. *ACM Transactions on Multimedia Computing, Communications, and Applications*. TOMM 15, 2. doi: 10.1145/3311784.

[95] Amanda Hutton, Alexander Liu, and Cheryl Martin. 2012. Crowdsourcing Evaluations of Classifier Interpretability. In *Proceedings of the AAAI Spring Symposium*. eprint: https://aaai.org/papers/04267-4267-crowdsourcing-evaluations-of-classifier-interpretability/.

[96] Frederik Hvilshøj, Alexandros Iosifidis, and Ira Assent. 2021. ECINN: Efficient Counterfactuals from Invertible Neural Networks. *arXiv preprint*. doi: 10.48550/arXiv.2103.13701.

[97] Jörn-Henrik Jacobsen, Arnold W.M. Smeulders, and Edouard Oyallon. 2018. i-RevNet: Deep Invertible Networks. In *Proceedings of the International Conference on Learning Representations* (ICLR). doi: 10.48550/arXiv.1802.07088.

[98] Jerald M Jellison and William John Ickes. 1974. The power of the glance: Desire to see and be seen in cooperative and competitive situations. *Journal of Experimental Social Psychology*, 10, 5, 444–450. doi: 10.1016/0022-1031(74)90012-2.

[99] Jason Jerald and Mike Daily. 2002. EyeGaze: Enabling Eye Contact over Video. In *Proceedings of the symposium on Eye tracking research & applications* (ETRA), 77. doi: 10.1145/507072.507088.

[100] Jeya Vikranth Jeyakumar, Joseph Noor, Yu-Hsi Cheng, Luis Garcia, and Mani Srivastava. 2020. How Can I Explain This to You? An Empirical Study of Deep Neural Network Explanation Methods. In *Advances in Neural Information Processing Systems* (NeurIPS). Vol. 33, 4211–4222. https://proceedings.neurips.cc/paper_files/paper/2020/file/2c29d89cc56cdb191c60db2f0bae796b-Paper.pdf.

[101] Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C. Lawrence Zitnick, and Ross Girshick. 2017. CLEVR: A Diagnostic Dataset for Compositional Language and Elementary Visual Reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR), 2901–2910. doi: 10.1109/CVPR.2017.215.

[102] Andrew Jones, Magnus Lang, Graham Fyffe, Xueming Yu, Jay Busch, Ian McDowall, Mark Bolas, and Paul Debevec. 2009. Achieving Eye Contact in a One-to-Many 3D Video Teleconferencing System. *ACM Transactions on Graphics*. TOG 28, 3. doi: 10.1145/1531326.1531370.

[103] Harmanpreet Kaur, Harsha Nori, Samuel Jenkins, Rich Caruana, Hanna Wallach, and Vaughan Jennifer Wortman. 2020. Interpreting Interpretability: Understanding Data Scientists' Use of Interpretability Tools for Machine Learning. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/3313831.3376219.

[104] Frank C. Keil. 2006. Explanation and Understanding. *Annual Review of Psychology*, 57, 1, 227–254. doi: 10.1146/annurev.psych.57.102904.190100.

[105] Adam Kendon. 1967. Some functions of gaze-direction in social interaction. *Acta Psychologica*, 26, 22–63. doi: 10.1016/0001-6918(67)90005-4.

[106] David E. Kieras and Susan Bovair. 1984. The Role of a Mental Model in Learning to Operate a Device. *Cognitive Science*, 8, 3, 255–273. doi: 10.1207/s15516709cog0803_3.

[107] Been Kim, Rajiv Khanna, and Oluwasanmi O Koyejo. 2016. Examples are not enough, learn to criticize! Criticism for Interpretability. In *Advances in Neural Information Processing Systems* (NIPS). Vol. 29, 2288–2296. https://proceedings.neurips.cc/paper/2016/file/5680522b8e2bb01943234bce7bf84534-Paper.pdf.

[108] Been Kim, Martin Wattenberg, Justin Gilmer, Carrie Cai, James Wexler, Fernanda Viegas, et al. 2018. Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV). In *Proceedings of the 35th International Conference on Machine Learning* (ICML), 2673–2682. doi: 10.48550/arXiv.1711.11279.

[109] Hyeongwoo Kim et al. 2018. Deep Video Portraits. *ACM Transactions on Graphics*. TOG 37, 4. doi: 10.1145/3197517.3201283.

[110] Kibum Kim, John Bolton, Audrey Girouard, Jeremy Cooperstock, and Roel Vertegaal. 2012. TeleHuman: Effects of 3D Perspective on Gaze and Pose Estimation with a Life-size Cylindrical Telepresence Pod. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 2531–2540. doi: 10.1145/2207676.2208640.

[111] Sunnie S. Y. Kim, Nicole Meister, Vikram V. Ramaswamy, Ruth Fong, and Olga Russakovsky. 2022. HIVE: Evaluating the Human Interpretability of Visual Explanations. In *Computer Vision* (ECCV), 280–298. doi: 10.1007/978-3-031-19775-8_17.

[112] Sunnie S. Y. Kim, Elizabeth Anne Watkins, Olga Russakovsky, Ruth Fong, and Andrés Monroy-Hernández. 2023. "Help Me Help the AI": Understanding How Explainability Can Support Human-AI Interaction. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/3544548.3581001.

[113] Diederik P. Kingma and Prafulla Dhariwal. 2018. Glow: Generative Flow with Invertible 1x1 Convolutions. In *Advances in Neural Information Processing Systems* (NeurIPS). Vol. 31, 10236–10245. doi: 10.48550/arXiv.1807.03039.

[114] René F. Kizilcec. 2016. How Much Information?: Effects of Transparency on Trust in an Algorithmic Interface. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 2390–2395. doi: 10.1145/2858036.2858402.

[115] Jesper Kjeldskov, Jacob H. Smedegård, Thomas S. Nielsen, Mikael B. Skov, and Jeni Paay. 2014. EyeGaze: enabling eye contact over video. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (AVI), 105–112. doi: 10.1145/2598153.2598165.

[116] Jon Kleinberg, Himabindu Lakkaraju, Jure Leskovec, Jens Ludwig, and Sendhil Mullainathan. 2018. Human Decisions and Machine Predictions. *The Quarterly Journal of Economics*, 133, 1, 237–293. doi: 10.1093/qje/qjx032.

[117] Chris L. Kleinke. 1986. Gaze and eye contact: A research review. *Psychological Bulletin*, 100, 1, 78–100. doi: 10.1037/0033-2909.100.1.78.

[118] Jürgen Koenemann and Nicholas J. Belkin. 1996. A Case for Interaction: A Study of Interactive Information Retrieval Behavior and Effectiveness. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 205–212. doi: 10.1145/238386.238487.

[119] Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. 2020. Concept Bottleneck Models. In *Proceedings of the 37th International Conference on Machine Learning* (ICML). Vol. 119, 5338–5348. doi: 10.48550/arXiv.2007.04612.

[120] Pigi Kouki, James Schaffer, Jay Pujara, John O'Donovan, and Lise Getoor. 2019. Personalized Explanations for Hybrid Recommender Systems. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI), 379–390. doi: 10.1145/3301275.3302306.

[121] Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. 2015. Principles of Explanatory Debugging to Personalize Interactive Machine Learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces* (IUI), 126–137. doi: 10.1145/2678025.2701399.

[122] Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. 2012. Tell Me More?: The Effects of Mental Model Soundness on Personalizing an Intelligent Agent. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/2207676.2207678.

[123] Todd Kulesza, Weng-Keen Wong, Simone Stumpf, Stephen Perona, Rachel White, Margaret M. Burnett, Ian Oberst, and Andrew J. Ko. 2009. Fixing the Program My Computer Learned: Barriers for End Users, Challenges for the Machine. In *Proceedings of the 14th International Conference on Intelligent User Interfaces* (IUI), 187–196. doi: 10.1145/1502650.1502678.

[124] Vivian Lai and Chenhao Tan. 2019. On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (FAccT), 29–38. doi: 10.1145/3287560.3287590.

[125] Himabindu Lakkaraju, Stephen H. Bach, and Jure Leskovec. 2016. Interpretable Decision Sets: A Joint Framework for Description and Prediction. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), 1675–1684. doi: 10.1145/2939672.2939874.

[126] Himabindu Lakkaraju, Ece Kamar, Rich Caruana, and Jure Leskovec. 2019. Faithful and Customizable Explanations of Black Box Models. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (AIES), 131–138. doi: 10.1145/3306618.3314229.

[127] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine*, 94, 42–53. doi: 10.1016/j.artmed.2019.01.001.

[128] Sebastian Lapuschkin, Stephan Wäldchen, Alexander Binder, Grégoire Montavon, Wojciech Samek, and Klaus-Robert Müller. 2019. Unmasking Clever Hans Predictors and Assessing What Machines Really Learn. *Nature Communications*, 10, 1. doi: 10.1038/s41467-019-08987-4.

[129] Jason Lawrence et al. 2021. Project Starline: A High-Fidelity Telepresence System. TOG 40, 6. doi: 10.1145/3478513.3480490.

[130] Matthew L Leavitt and Ari Morcos. 2020. Towards falsifiable interpretability research. In *ML Retrospectives, Surveys & Meta-Analyses – Workshop* (NeurIPS). doi: 10.48550/arXiv.2010.12016.

[131] John D. Lee and Katrina A. See. 2004. Trust in Automation: Designing for Appropriate Reliance. *Human Factors*, 46, 1, 50–80. doi: 10.1518/hfes.46.1.50_30392.

[132] Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. 2009. Why and Why Not Explanations Improve the Intelligibility of Context-aware Intelligent Systems. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 2119–2128. doi: 10.1145/1518701.1519023.

[133] Shanchuan Lin, Andrey Ryabtsev, Soumyadip Sengupta, Brian Curless, Steve Seitz, and Ira Kemelmacher-Shlizerman. 2021. Real-Time High-Resolution Background Matting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 8758–8767. doi: 10.1109/CVPR46437.2021.00865.

[134] Zachary C. Lipton. 2018. The Mythos of Model Interpretability. *Communications of the ACM*, 61, 10, 36–43. doi: 10.1145/3233231.

[135] Shusen Liu, Bhavya Kailkhura, Donald Loveland, and Yong Han. 2019. Generative Counterfactual Introspection for Explainable Deep Learning. In *Proceedings of the IEEE Global Conference on Signal and Information Processing* (GlobalSIP). doi: 10.1109/GlobalSIP45357.2019.8969491.

[136] Matthew Lombard and Theresa Ditton. 1997. At the Heart of It All: The Concept of Presence. *Journal of Computer-Mediated Communication*, 3, 2. doi: 10.1111/j.1083-6101.1997.tb00072.x.

[137] Vincent Lully, Philippe Laublet, Milan Stankovic, and Filip Radulovic. 2018. Enhancing explanations in recommender systems with knowledge graphs. *Procedia Computer Science*, 137, 211–222. doi: 10.1016/j.procs.2018.09.020.

[138] Scott M Lundberg and Su-In Lee. 2017. A Unified Approach to Interpreting Model Predictions. In *Advances in Neural Information Processing Systems* (NIPS). Vol. 30, 4765–4774. doi: 10.48550/arXiv.1705.07874.

[139] Ian Scott MacKenzie. 2013. *Human-Computer Interaction: An Empirical Research Perspective*. (1st ed.). Morgan Kaufmann Publishers Inc., San Francisco, USA. isbn: 978-0-12-405865-1.

[140] Radek Mackowiak, Lynton Ardizzone, Ullrich Köthe, and Carsten Rother. 2021. Generative Classifiers as a Basis for Trustworthy Image Classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 2970–2980. doi: 10.1109/CVPR46437.2021.00299.

[141] Philipp Mayring and Thomas Fenzl. 2019. Qualitative Inhaltsanalyse. In *Handbuch Methoden der empirischen Sozialforschung*. Springer, Wiesbaden, Germany, 633–648. isbn: 978-3-658-21308-4. doi: 10.1007/978-3-658-21308-4_42.

[142] Silvan Mertes, Tobias Huber, Katharina Weitz, Alexander Heimerl, and Elisabeth Andre. 2022. GANterfactual — Counterfactual Explanations for Medical Non-experts Using Generative Adversarial Learning. *Frontiers in Artificial Intelligence*, 5. doi: 10.3389/frai.2022.825565.

[143] Milagros Miceli, Martin Schuessler, and Tianling Yang. 2020. Between Subjectivity and Imposition: Power Dynamics in Data Annotation for Computer Vision. *Proceedings of the ACM on Human-Computer Interaction*, 4, CSCW2. doi: 10.1145/3415186.

[144] Milagros Miceli, Tianling Yang, Laurens Naudts, Martin Schuessler, Diana Serbanescu, and Alex Hanna. 2021. Documenting Computer Vision Datasets: An Invitation to Reflexive Data Practices. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency* (FAccT). ACM, 161–172. isbn: 978-1-4503-8309-7. doi: 10.1145/3442188.3445880.

[145] Martijn Millecamp, Nyi Nyi Htun, Cristina Conati, and Katrien Verbert. 2019. To Explain or not to Explain: the Effects of Personal Characteristics when Explaining Music Recommendations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI), 397–407. doi: 10.1145/3301275.3302313.

[146] Tim Miller. 2019. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, 267. doi: 10.1016/j.artint.2018.07.007.

[147] Tim Miller, Piers Howe, and Liz Sonenberg. 2017. Explainable AI: Beware of Inmates Running the Asylum. In *Explainable Artificial Intelligence (XAI) – Workshop* (IJCAI). doi: 10.48550/arXiv.1712.00547.

[148] Yao Ming, Panpan Xu, Huamin Qu, and Liu Ren. 2019. Interpretable and Steerable Sequence Learning via Prototypes. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), 903–913. doi: 10.1145/3292500.3330908.

[149] G. O. Mohler, M. B. Short, Sean Malinowski, Mark Johnson, G. E. Tita, Andrea L. Bertozzi, and P. J. Brantingham. 2015. Randomized Controlled Field Trials of Predictive Policing. *Journal of the American Statistical Association*, 110, 512, 1399–1411. doi: 10.1080/01621459.2015.1077710.

[150] Sina Mohseni, Jeremy E Block, and Eric Ragan. 2021. Quantitative Evaluation of Machine Learning Explanations: A Human-Grounded Benchmark. In *Proceedings of the 26th International Conference on Intelligent User Interfaces* (IUI), 22–31. doi: 10.1145/3397481.3450689.

[151] Christoph Molnar. 2022. *Interpretable Machine Learning: A Guide For Making Black Box Models Explainable*. Independently published, Munich, GER. isbn: 979-8-4114-6333-0. eprint: https://christophm.github.io/interpretable-ml-book/index.html.

[152] Christoph Molnar, Giuseppe Casalicchio, and Bernd Bischl. 2020. Interpretable Machine Learning – A Brief History, State-of-the-Art and Challenges. In *Workshop Track* (ECML-PKDD), 417–431. doi: 10.1007/978-3-030-65965-3_28.

[153] Andrew Monk, John McCarthy, Leon Watts, and Owen Daly-Jones. 1996. Measures of Process. In *CSCW requirements and evaluation*, 125–139. doi: 10.1007/978-1-4471-3056-7_9.

[154] Grégoire Montavon, Sebastian Lapuschkin, Alexander Binder, Wojciech Samek, and Klaus-Robert Müller. 2017. Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognition*, 65, 211–222. doi: 10.1016/j.patcog.2016.11.008.

[155] Neville Moray. 1999. Mental Models in Theory and Practice. *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application*, 223–258. isbn: 978-0-262-07188-8.

[156] Masahiro Mori, Karl F. MacDorman, and Norri Kageki. 2012. The Uncanny Valley [From the Field]. *IEEE Robotics & Automation Magazine*, 19, 2, 98–100. doi: 10.1109/MRA.2012.2192811.

[157] Khalil Ibrahim Muhammad, Aonghus Lawlor, and Barry Smyth. 2016. A Live-User Study of Opinionated Explanations for Recommender Systems. In *Proceedings of the 21st International Conference on Intelligent User Interfaces* (IUI), 256–260. doi: 10.1145/2856767.2856813.

[158] Lothar Muhlbach, Martin Bocker, and Angela Prussog. 1995. Telepresence in Videocommunications: A Study on Stereoscopy and Individual Eye Contact. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 37, 2, 290–305. doi: 10.1518/001872095779064582.

[159] Ferdinand Müller, Elsa Kirchner, and Martin Schüßler. 2021. Ein „KI-TÜV“ für Europa? Eckpunkte einer horizontalen Regulierung algorithmischer Entscheidungssysteme. In *GRUR Junge Wissenschaft Intelligente Systeme – Intelligentes Recht*. Vol. 2020/21. Nomos Verlagsgesellschaft, 85–106. isbn: 978-3-8487-8142-3. https://www.nomos-shop.de/nomos/titel/intelligente-systeme-intelligentes-recht-id-99401/.

[160] Jack Muramatsu and Wanda Pratt. 2001. Transparent Queries: Investigation Users' Mental Models of Search Engines. In *Proceedings of the 24th Annual International ACM Conference on Research and Development in Information Retrieval* (SIGIR), 217–224. doi: 10.1145/383952.383991.

[161] Menaka Narayanan, Emily Chen, Jeffrey He, Been Kim, Sam Gershman, and Finale Doshi-Velez. 2018. How Do Humans Understand Explanations from Machine Learning Systems? An Evaluation of the Human-Interpretability of Explanation. *arXiv preprint*. doi: 10.48550/arXiv.1802.00682.

[162] Meike Nauta, Jan Trienes, Shreyasi Pathak, Elisa Nguyen, Michelle Peters, Yasmin Schmitt, Jörg Schlötterer, Maurice van Keulen, and Christin Seifert. 2023. From Anecdotal Evidence to Quantitative Evaluation Methods: A Systematic Review on Evaluating Explainable AI. *ACM Computing Surveys*. CSUR 55, 13s. doi: 10.1145/3583558.

[163] Ian Neath and Aimee Surprenant. 2002. *Human Memory*. (2 edition ed.). Thomson/Wadsworth, Belmont, USA. isbn: 978-0-534-59562-3.

[164] An Nguyen, Aditya Kharosekar, Matthew Lease, and Byron Wallace. 2018. An Interpretable Joint Graphical Model for Fact-Checking From Crowds. In *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI) number 1. Vol. 32, 1511–1518. doi: 10.1609/aaai.v32i1.11487.

[165] David T. Nguyen and John Canny. 2007. MultiView: Improving Trust in Group Video Conferencing Through Spatial Faithfulness. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 1465–1474. doi: 10.1145/1240624.1240846.

[166] Robert C Nickerson, Upkar Varshney, and Jan Muntermann. 2013. A method for taxonomy development and its application in information systems. *European Journal of Information Systems*, 22, 3, 336–359. doi: 10.1057/ejis.2012.26.

[167] Weili Nie, Yang Zhang, and Ankit Patel. 2018. A Theoretical Explanation for Perplexing Behaviors of Backpropagation-based Visualizations. In *Proceedings of the 35th International Conference on Machine Learning* (ICML), 3809–3818. doi: 10.48550/arXiv.1805.07039.

[168] Donald Norman. 2014. On the Relationship between Conceptual and Mental Models. In *Mental Models*. Psychology Press, New York, USA. Chap. 1. isbn: 978-0-89859-242-9.

[169] Conor Nugent, Pádraig Cunningham, and Dónal Doyle. 2005. The Best Way to Instil Confidence Is by Being Right. In *Proceedings of the International Conference on Case-Based Reasoning* (ICCBR), 368–381. doi: 10.1007/11536406_29.

[170] Ingrid Nunes and Dietmar Jannach. 2017. A Systematic Review and Taxonomy of Explanations in Decision Support and Recommender Systems. *User Modeling and User-Adapted Interaction*. UMUAI 27, 3-5, 393–444. doi: 10.1007/s11257-017-9195-0.

[171] Ingrid Nunes, Simon Miles, Michael Luck, Simone Barbosa, and Carlos Lucena. 2014. Pattern-based Explanation for Automated Decisions. In *Proceedings of the 21st European Conference on Artificial Intelligence* (ECAI), 669–674. doi: 10.3233/978-1-61499-419-0-669.

[172] Ken-Ichi Okada, Fumihiko Maeda, Yusuke Ichikawaa, and Yutaka Matsushita. 1994. Multiparty videoconferencing at virtual social distance: MAJIC design. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW), 385–393. doi: 10.1145/192844.193054.

[173] Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature Visualization. *Distill*, 2, 11. doi: 10.23915/distill.00007.

[174] Sergio Orts-Escolano et al. 2016. Holoportation: Virtual 3D Teleportation in Real-Time. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology* (UIST), 741–754. doi: 10.1145/2984511.2984517.

[175] Kazuhiro Otsuka. 2018. Behavioral Analysis of Kinetic Telepresence for Small Symmetric Group-to-Group Meetings. *IEEE Transactions on Multimedia*, 20, 6, 1432–1447. doi: 10.1109/TMM.2017.2771396.

[176] Kazuhiro Otsuka. 2016. MMSpace: Kinetically-augmented telepresence for small group-to-group conversations. In *IEEE Virtual Reality (VR)*, 19–28. doi: 10.1109/VR.2016.7504684.

[177] Tomislav Pejsa, Julian Kantor, Hrvoje Benko, Eyal Ofek, and Andrew Wilson. 2016. Room2Room: Enabling Life-Size Telepresence in a Projected Augmented Reality Environment. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW), 1716–1725. doi: 10.1145/2818048.2819965.

[178] Rolf Porst and Christa von Briel. 1995. *Wären Sie vielleicht bereit, sich gegebenenfalls noch einmal befragen zu lassen? Oder: Gründe für die Teilnahme an Panelbefragungen*. *ZUMA-Arbeitsbericht*. Vol. 1995/04. Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim, GER. isbn: 978-3-924220-20-7.

[179] Samira Pouyanfar, Saad Sadiq, Yilin Yan, Haiman Tian, Yudong Tao, Maria Presa Reyes, Mei-Ling Shyu, Shu-Ching Chen, and S. S. Iyengar. 2018. A Survey on Deep Learning: Algorithms, Techniques, and Applications. *ACM Computing Surveys*. CSUR 51, 5. doi: 10.1145/3234150.

[180] Shenhan Qian, Tobias Kirschstein, Liam Schoneveld, Davide Davoli, Simon Giebenhain, and Matthias Nießner. 2024. GaussianAvatars: Photorealistic Head Avatars with Rigged 3D Gaussians. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 20299–20309. doi: 10.48550/arXiv.2312.02069.

[181] Vikram V. Ramaswamy, Sunnie S. Y. Kim, Ruth Fong, and Olga Russakovsky. 2023. Overlooked Factors in Concept-Based Explanations: Dataset Choice, Concept Learnability, and Human Capability. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 10932–10941. doi: 10.48550/arXiv.2207.09615.

[182] Tajul Rosli Razak, Jonathan M. Garibaldi, Christian Wagner, Amir Pourabdollah, and Daniele Soria. 2018. Interpretability and Complexity of Design in the Creation of Fuzzy Logic Systems — A User Study. In *IEEE Symposium Series on Computational Intelligence* (SSCI). doi: 10.1109/ssci.2018.8628924.

[183] H. Regenbrecht, L. Müller, S. Hoermann, T. Langlotz, M. Wagner, and M. Billinghurst. 2014. Eye-to-Eye Contact for Life-Sized Videoconferencing. In *Proceedings of the 26th Australian Computer-Human Interaction Conference* (OzCHI), 145–148. doi: 10.1145/2686612.2686632.

[184] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (KDD), 1135–1144. doi: 10.1145/2939672.2939778.

[185] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Anchors: High-Precision Model-Agnostic Explanations. In *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI) number 1. Vol. 32, 1527–1535. doi: 10.1609/aaai.v32i1.11491.

[186] René Riedl. 2022. On the stress potential of videoconferencing: definition and root causes of Zoom fatigue. *Electronic Markets*, 32, 1, 153–177. doi: 10.1007/s12525-021-00501-3.

[187] Giuseppe Riva, Brenda K. Wiederhold, and Fabrizia Mantovani. 2021. Surviving COVID-19: The Neuroscience of Smart Working and Distance Learning. *Cyberpsychology, Behavior and Social Networking*, 24, 2, 79–85. doi: 10.1089/cyber.2021.0009.

[188] Shane L Rogers, Rebecca Broadbent, Jemma Brown, Allan Fraser, and Craig P Speelman. 2022. Realistic Motion Avatars are the Future for Social Interaction in Virtual Reality. *Frontiers in Virtual Reality*, 2. doi: 10.3389/frvir.2021.750729.

[189] Robin Rombach, Patrick Esser, and Björn Ommer. 2020. Making Sense of CNNs: Interpreting Deep Representations and Their Invariances with INNs. In *Computer Vision* (ECCV), 647–664. doi: 10.1007/978-3-030-58520-4_38.

[190] Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1, 5, 206–215. doi: 10.1038/s42256-019-0048-x.

[191] Kimiko Ryokai, Elena Durán López, Noura Howell, Jon Gillick, and David Bamman. 2018. Capturing, Representing, and Interacting with Laughter. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/3173574.3173932.

[192] Shunsuke Saito, Gabriel Schwartz, Tomas Simon, Junxuan Li, and Giljoo Nam. 2024. Relightable Gaussian Codec Avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 130–141. doi: 10.48550/arXiv.2312.03704.

[193] Mark Sandler, Andrew G. Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. 2018. MobileNetV2: Inverted Residuals and Linear Bottlenecks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 4510–4520. doi: 10.1109/CVPR.2018.00474.

[194] Masahiro Sato, Budrul Ahsan, Koki Nagatani, Takashi Sonoda, Qian Zhang, and Tomoko Ohkuma. 2018. Explaining Recommendations Using Contexts. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces* (IUI), 659–664. doi: 10.1145/3172944.3173012.

[195] Axel Sauer and Andreas Geiger. 2021. Counterfactual Generative Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR). doi: 10.48550/arXiv.2101.06046.

[196] James Schaffer, John O'Donovan, James Michaelis, Adrienne Raglin, and Tobias Höllerer. 2019. I Can Do Better than Your AI: Expertise and Explanations. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI), 240–251. doi: 10.1145/3301275.3302308.

[197] Martin Schuessler, Luca Hormann, Raimund Dachselt, Andrew Blake, and Carsten Rother. 2024. Gazing Heads: Investigating Gaze Perception in Video-Mediated Communication. *ACM Transactions on Computer-Human Interaction*. TOCHI 31, 3. doi: 10.1145/3660343.

[198] Martin Schuessler and Philipp Weiß. 2019. Minimalistic Explanations: Capturing the Essence of Decisions. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (CHI). ACM, LBW2810:1–LBW2810:6. doi: 10.1145/3290607.3312823.

[199] Martin Schuessler, Philipp Weiß, and Leon Sixt. 2021. Two4Two: Evaluating Interpretable Machine Learning - A Synthetic Dataset For Controlled Experiments. In *Responsible AI – Workshop* (ICLR). doi: 10.48550/arXiv.2105.02825.

[200] Karl Schulz, Leon Sixt, Federico Tombari, and Tim Landgraf. 2020. Restricting the Flow: Information Bottlenecks for Attribution. In *Proceedings of the International Conference on Learning Representations* (ICLR). doi: 10.48550/arXiv.2001.00396.

[201] Abigail J. Sellen. 1992. Speech patterns in video-mediated conversations. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 49–59. doi: 10.1145/142750.142756.

[202] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2017. Grad-CAM: Visual Explanations From Deep Networks via Gradient-Based Localization. In *Proceedings of the IEEE International Conference on Computer Vision* (ICCV). doi: 10.1109/ICCV.2017.74.

[203] Hua Shen and Ting-Hao Kenneth Huang. 2020. How Useful Are the Machine-Generated Interpretations to General Users? A Human Evaluation on Guessing the Incorrectly Predicted Labels. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* (HCOMP). Vol. 8, 168–172. doi: 10.1609/hcomp.v8i1.7477.

[204] Soo Yun Shin, Ezgi Ulusoy, Kelsey Earle, Gary Bente, and Brandon Van Der Heide. 2022. The effects of self-viewing in video chat during interpersonal work conversations. *Journal of Computer-Mediated Communication*, 28, 1. doi: 10.1093/jcmc/zmac028.

[205] Ben Shneiderman. 2016. Opinion: The Dangers of Faulty, Biased, or Malicious Algorithms Requires Independent Oversight. In *Proceedings of the National Academy of Sciences* number 48. Vol. 113, 13538–13540. doi: 10.1073/pnas.1618211113.

[206] John Short, Ederyn Williams, and Bruce Christie. 1976. *The social psychology of telecommunications*. Wiley. isbn: 978-0-471-01581-9.

[207] Aliaksandr Siarohin, Stéphane Lathuilière, Sergey Tulyakov, Elisa Ricci, and Nicu Sebe. 2019. First Order Motion Model for Image Animation. In *Advances in Neural Information Processing Systems* (NeurIPS). Vol. 32, 7137–7147. https://proceedings.neurips.cc/paper/2019/file/31c0b36aef265d9221af80872ceb62f9-Paper.pdf.

[208] K Simonyan, A Vedaldi, and A Zisserman. 2014. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. In *Proceedings of the International Conference on Learning Representations* (ICLR). doi: 10.48550/arXiv.1312.6034.

[209] Sumedha Singla, Brian Pollack, Junxiang Chen, and Kayhan Batmanghelich. 2020. Explanation by Progressive Exaggeration. In *Proceedings of the International Conference on Learning Representations* (ICLR). doi: 10.48550/arXiv.1911.00483.

[210] David Sirkin, Gina Venolia, John Tang, George Robertson, Taemie Kim, Kori Inkpen, Mara Sedlins, Bongshin Lee, and Mike Sinclair. 2011. Motion and Attention in a Kinetic Videoconferencing Proxy. In *Proceedings of the IFIP Conference on Human-Computer Interaction* (INTERACT), 162–180. doi: 10.1007/978-3-642-23774-4_16.

[211] Leon Sixt, Maximilian Granz, and Tim Landgraf. 2020. When Explanations Lie: Why Many Modified BP Attributions Fail. In *Proceedings of the 37th International Conference on Machine Learning* (ICML), 9046–9057. doi: 10.48550/arXiv.1912.09818.

[212] Leon Sixt, Martin Schuessler, Oana-Iuliana Popescu, Philipp Weiß, and Tim Landgraf. 2022. Do Users Benefit From Interpretable Vision? A User Study, Baseline, And Dataset. In *Proceedings of the International Conference on Learning Representations* (ICLR). https://openreview.net/forum?id=v6s3HVjPerv. doi: 10.48550/arXiv.2204.11642.

[213] Leon Sixt, Martin Schuessler, Philipp Weiß, and Tim Landgraf. 2021. Interpretability Through Invertibility: A Deep Convolutional Network With Ideal Counterfactuals And Isosurfaces. Rejected submission to the International Conference on Learning Representations (ICLR). (2021). https://openreview.net/forum?id=8YFhXYe1Ps.

[214] Dylan Slack, Sophie Hilgard, Emily Jia, Sameer Singh, and Himabindu Lakkaraju. 2020. Fooling LIME and SHAP: Adversarial Attacks on Post Hoc Explanation Methods. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (AIES), 180–186. doi: 10.1145/3375627.3375830.

[215] Daniel Smilkov, Nikhil Thorat, Been Kim, Fernanda Viégas, and Martin Wattenberg. 2017. SmoothGrad: removing noise by adding noise. In *Visualization for Deep Learning – Workshop* (ICML). doi: 10.48550/arXiv.1706.03825.

[216] Deborah Hinderer Sova and Jakob Nielsen. 2010. Determining the Appropriate Incentives. In *234 Tips and Tricks for Recruiting Users as Participants in Usability Studies*. New Riders, 31–44. isbn: 978-0-13-210753-2. https://www.nngroup.com/reports/how-to-recruit-participants-usability-studies/.

[217] Martin Spindler, Martin Schuessler, Marcel Martsch, and Raimund Dachselt. 2014. Pinch-Drag-Flick vs. Spatial Input: Rethinking Zoom & Pan on Mobile Displays. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 1113–1122. doi: 10.1145/2556288.2557028.

[218] Jost Tobias Springenberg, Alexey Dosovitskiy, Thomas Brox, and Martin Riedmiller. 2015. Striving for Simplicity: The All Convolutional Net. In *Workshop Track* (ICLR). doi: 10.48550/arXiv.1412.6806.

[219] Aaron Springer, Victoria Hollis, and Steve Whittaker. 2017. Dice in the Black Box: User Experiences with an Inscrutable Algorithm. In *Proceedings of the AAAI Spring Symposium*. doi: 10.48550/arXiv.1812.03219 .

[220] Aaron Springer and Steve Whittaker. 2019. Progressive Disclosure: Empirically Motivated Approaches to Designing Effective Transparency. In *Proceedings of the 24th International Conference on Intelligent User Interfaces* (IUI), 107–120. doi: 10.1145/3301275.3302322.

[221] Sven Stadtmüller and Rolf Porst. 2005. *Zum Einsatz von Incentives bei postalischen Befragungen. GESIS-How-to*. Vol. 14. Zentrum für Umfragen, Methoden und Analysen (ZUMA), Mannheim, GER. https://nbn-resolving.org/urn:nbn:de:0168-ssoar-201465.

[222] William Steptoe, Robin Wolff, Alessio Murgia, Estefania Guimaraes, John Rae, Paul Sharkey, David Roberts, and Anthony Steed. 2008. Eye-Tracking for Avatar Eye-Gaze and Interactional Analysis in Immersive Collaborative Virtual Environments. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW), 197–200. doi: 10.1145/1460563.1460593.

[223] Peter Stone et al. 2016. "Artificial Intelligence and Life in 2030." One Hundred Year Study on Artificial Intelligence: Report of the 2015-2016 Study Panel. Tech. rep. Stanford University. doi: 10.48550/arXiv.2211.06318.

[224] Anant Subramanian, Danish Pruthi, Harsh Jhamtani, Taylor Berg-Kirkpatrick, and Eduard Hovy. 2018. SPINE: SParse Interpretable Neural Embeddings. In *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI) number 1. Vol. 32, 4921–4928. doi: 10.1609/aaai.v32i1.11935.

[225] Jian Sun and Holger Regenbrecht. 2007. Implementing Three-Party Desktop Videoconferencing. In *Proceedings of the 19th Australian Computer-Human Interaction Conference* (OZCHI), 95. doi: 10.1145/1324892.1324910.

[226] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic Attribution for Deep Networks. In *Proceedings of the 34th International Conference on Machine Learning* (ICML), 3319–3328. doi: 10.48550/arXiv.1703.01365.

[227] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (CVPR). doi: 10.1109/CVPR.2015.7298594.

[228] Silero Team. 2021. Silero VAD: pre-trained enterprise-grade Voice Activity Detector (VAD), Number Detector and Language Classifier. https://github.com/snakers4/silero-vad. (2021).

[229] Jamie Teevan and Abigal Sellen. 2021. New Future of Work: Meeting and collaborating in a remote and hybrid world with Jaime Teevan and Abigail Sellen. Microsoft Research Podcast. (2021). Retrieved Feb. 22, 2023 from https://www.microsoft.com/en-us/research/podcast/new-future-of-work-meeting-and-collaborating-in-a-remote-and-hybrid-world-with-jaime-teevan-and-abigail-sellen/.

[230] Stefano Teso and Kristian Kersting. 2019. Explanatory Interactive Machine Learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* (AIES), 239–245. doi: 10.1145/3306618.3314293.

[231] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. 2017. Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems, Version 2. (2017). https://standards.ieee.org/content/dam/ieee-standards/standards/web/documents/other/ead_v2.pdf.

[232] Nava Tintarev and Judith Masthoff. 2007. A Survey of Explanations in Recommender Systems. In *Proceedings of the IEEE International Conference on Data Engineering Workshop* (ICDEW), 801–810. doi: 10.1109/ICDEW.2007.4401070.

[233] Nava Tintarev and Judith Masthoff. 2012. Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*. UMUAI 22, 4–5, 399–439. doi: 10.1007/s11257-011-9117-5.

[234] Chun-Hua Tsai and Peter Brusilovsky. 2019. Evaluating Visual Explanations for Similarity-Based Recommendations: User Perception and Performance. In *Proceedings of the ACM Conference on User Modeling, Adaptation and Personalization* (UMAP), 22–30. doi: 10.1145/3320435.3320465.

[235] Katherine M. Tsui, Munjal Desai, and Holly A. Yanco. 2012. Towards Measuring the Quality of Interaction: Communication through Telepresence Robots. In *Proceedings of the Workshop on Performance Metrics for Intelligent Systems* (PerMIS), 101. doi: 10.1145/2393091.2393112.

[236] Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. 2007. How It Works: A Field Study of Non-technical Users Interacting with an Intelligent System. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 31–40. doi: 10.1145/1240624.1240630.

[237] European Union. 2024. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 on Artificial Intelligence and amending certain Union legislative acts. Official Journal of the European Union, L 168. Article 9. (2024). https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=OJ:L_202401689#d1e3277-1-1.

[238] Rick van der Kleij, Jan Maarten Schraagen, Peter Werkhoven, and Carsten K. W. De Dreu. 2009. How Conversations Change Over Time in Face-to-Face and Video-Mediated Communication. *Small Group Research*, 40, 4, 355–381. doi: 10.1177/1046496409333724.

[239] Nadya Vasilyeva, Daniel Wilkenfeld, and Tania Lombrozo. 2015. Goals Affect the Perceived Quality of Explanations. In *Proceedings of the 37th Annual Meeting of the Cognitive Science Society* (CogSci), 2469–2474. isbn: 978-0-9911967-2-2.

[240] Jennifer Wortman Vaughan and Hanna Wallach. 2021. A Human-Centered Agenda for Intelligible Machine Learning. In *Machines We Trust: Perspectives on Dependable AI*, 123–138. doi: 10.7551/mitpress/12186.003.0014.

[241] Roel Vertegaal and Yaping Ding. 2002. Explaining Effects of Eye Gaze on Mediated Group Conversations: Amount or Synchronization? In *Proceedings of the ACM Conference on Computer Supported Cooperative Work* (CSCW), 41. doi: 10.1145/587078.587085.

[242] Roel Vertegaal, Gerrit Van Der Veer, and Harro Vons. 2000. Effects of Gaze on Multiparty Mediated Communication. In *Proceedings of Graphics Interface 2000*. Vol. Montréal, 8 pages, 366.40 KB. doi: 10.20380/GI2000.14.

[243] Roel Vertegaal, Ivo Weevers, Changuk Sohn, and Chris Cheung. 2003. GAZE-2: Conveying Eye Contact in Group Video Conferencing Using Eye-Controlled Camera Direction. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI), 521–528. doi: 10.1145/642611.642702.

[244] Sandra Wachter, Brent Mittelstadt, and Luciano Floridi. 2017. Why a Right to Explanation of Automated Decision-Making Does Not Exist in the General Data Protection Regulation. *International Data Privacy Law*, 7, 2, 76–99. doi: 10.1093/idpl/ipx005.

[245] Sandra Wachter, Brent Mittelstadt, and Christopher Russell. 2017. Counterfactual Explanations Without Opening The Black Box: Automated Decisions And The GDPR. *SSRN Electronic Journal*, 31, 2, 841–887. doi: 10.2139/ssrn.3063289.

[246] C. Wah, S. Branson, P. Welinder, P. Perona, and S. Belongie. 2011. The Caltech-UCSD Birds-200-2011 Dataset. CNS-TR-2011-001.

[247] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y. Lim. 2019. Designing Theory-Driven User-Centric Explainable AI. In *Proceedings of the Conference on Human Factors in Computing Systems* (CHI). doi: 10.1145/3290605.3300831.

[248] Ting-Chun Wang, Arun Mallya, and Ming-Yu Liu. 2021. One-Shot Free-View Neural Talking-Head Synthesis for Video Conferencing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 10034–10044. doi: 10.1109/CVPR46437.2021.00991.

[249] Oscar Michael Watson. 1970. *Proxemic behavior: A cross-cultural study*. Vol. 8. Walter de Gruyter, 127. isbn: 978-3-11-098102-5.

[250] Katharina Weitz, Dominik Schiller, Ruben Schlagowski, Tobias Huber, and Elisabeth André. 2019. "Do you trust me?": Increasing User-Trust by Integrating Virtual Agents in Explainable AI Interaction Design. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (IVA). doi: 10.1145/3308532.3329441.

[251] Halbert White. 1980. A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 48, 4, 817–838. doi: 10.2307/1912934.

[252] Bob G. Witmer, Christian J. Jerome, and Michael J. Singer. 2005. The Factor Structure of the Presence Questionnaire. *Presence*, 14, 3, 298–312. doi: 10.1162/105474605323384654.

[253] Tom Nuno Wolf, Fabian Bongratz, Anne-Marie Rickmann, Sebastian Pölsterl, and Christian Wachinger. 2024. Keep the Faith: Faithful Explanations in Convolutional Neural Networks for Case-Based Reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI) number 6. Vol. 38, 5921–5929. doi: 10.1609/aaai.v38i6.28406.

[254] Erroll Wood, Tadas Baltrušaitis, Louis-Philippe Morency, Peter Robinson, and Andreas Bulling. 2018. GazeDirector: Fully Articulated Eye Gaze Redirection in Video. *Computer Graphics Forum*. CGF 37, 2, 217–225. doi: 10.1111/cgf.13355.

[255] Bichen Wu et al. 2020. Visual Transformers: Token-based Image Representation and Processing for Computer Vision. *arXiv preprint*. doi: 10.48550/arXiv.2006.03677.

[256] Yang Xu, Chenguang Yu, Jingjiang Li, and Yong Liu. 2012. Video Telephony for End-Consumers: Measurement Study of Google+, IChat, and Skype. In *Proceedings of the Internet Measurement Conference* (IMC), 371–384. doi: 10.1145/2398776.2398816.

[257] Mengjiao Yang and Been Kim. 2019. Benchmarking Attribution Methods with Relative Feature Importance. *arXiv preprint*. doi: 10.48550/arXiv.1907.09701.

[258] Rayoung Yang and Mark W. Newman. 2013. Learning from a Learning Thermostat: Lessons for Intelligent Systems for the Home. In *Proceedings of the ACM International Joint Conference on Pervasive and Ubiquitous Computing* (UbiComp), 93–102. doi: 10.1145/2493432.2493489.

[259] Svetlana Yarosh and Panos Markopoulos. 2010. Design of an Instrument for the Evaluation of Communication Technologies with Children. In *Proceedings of the 9th International Conference on Interaction Design and Children* (IDC), 266–269. doi: 10.1145/1810543.1810587.

[260] Alex Yu, Vickie Ye, Matthew Tancik, and Angjoo Kanazawa. 2021. pixelnerf: Neural radiance fields from one or few images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR), 4578–4587. doi: 10.1109/CVPR46437.2021.00455.

[261] Markus Zanker. 2012. The Influence of Knowledgeable Explanations on Users' Perception of a Recommender System. In *Proceedings of the 6th ACM conference on Recommender systems* (RecSys), 269–272. doi: 10.1145/2365952.2366011.

[262] Matthew D. Zeiler and Rob Fergus. 2014. Visualizing and Understanding Convolutional Networks. In *Computer Vision* (ECCV), 818–833. doi: 10.1007/978-3-319-10590-1_53.

[263] Ruihan Zhang, Prashan Madumal, Tim Miller, Krista A Ehinger, and Benjamin IP Rubinstein. 2021. Invertible Concept-based Explanations for CNN Models with Non-negative Concept Activation Vectors. In *Proceedings of the AAAI Conference on Artificial Intelligence* (AAAI) number 13. Vol. 35, 11682–11690. doi: 10.1609/aaai.v35i13.17389.

[264] Bolei Zhou, Yiyou Sun, David Bau, and Antonio Torralba. 2018. Interpretable Basis Decomposition for Visual Explanation. In *Computer Vision* (ECCV), 122–138. doi: 10.1007/978-3-030-01237-3_8.