



# City Research Online

## City St George's, University of London

**Citation:** Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P. & Holzinger, A. (2018). Explainable AI: The new 42?. Paper presented at the CD-MAKE 2018, 27-30 Aug 2018, Hamburg, Germany. doi: 10.1007/978-3-319-99740-7\_21

This is the accepted version of the paper.

This version of the publication may differ from the final published version. To cite this item please consult the publisher's version.

**Permanent repository link:** <https://openaccess.city.ac.uk/id/eprint/20659/>

**Link to published version:** [https://doi.org/10.1007/978-3-319-99740-7\\_21](https://doi.org/10.1007/978-3-319-99740-7_21)

**Copyright and Reuse:** Copyright and Moral Rights remain with the author(s) and/or copyright holders. Copies of full items can be used for personal research or study, educational, or not-for-profit purposes without prior permission or charge, unless otherwise indicated, provided that the authors, title and full bibliographic details are credited, a hyperlink and/or URL is given for the original metadata page and the content is not changed in any way. For full details of reuse please refer to [City Research Online policy](#).

# Explainable AI: the new 42?

Randy Goebel<sup>1</sup>, Ajay Chander<sup>2</sup>, Katharina Holzinger<sup>3</sup>, Freddy Lecue<sup>4,5</sup>  
Zeynep Akata<sup>6,7</sup>, Simone Stumpf<sup>8</sup>, Peter Kieseberg<sup>3,9</sup>, Andreas Holzinger<sup>10,11</sup>

<sup>1</sup> Alberta Machine Intelligence Institute  
University of Alberta, Edmonton, Canada  
`rgoebel@ualberta.ca`

<sup>2</sup> Fujitsu Labs of America, US  
`achander@us.fujitsu.com`

<sup>3</sup> SBA-Research, Vienna, Austria  
`kholzinger@sba-research.org`

<sup>4</sup> INRIA, Sophia Antipolis, France

<sup>5</sup> Accenture Labs, Dublin, Ireland  
`freddy.lecue@inria.fr`

<sup>6</sup> Amsterdam Machine Learning Lab, University of Amsterdam, The Netherlands

<sup>7</sup> Max Planck Institute for Informatics, Saarbruecken, Germany  
`z.akata@uva.nl`

<sup>8</sup> City, University of London, UK  
`Simone.Stumpf.1@city.ac.uk`

<sup>9</sup> University of Applied Sciences St. Pölten, Austria  
`Peter.Kieseberg@fhstp.ac.at`

<sup>10</sup> Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics &  
Documentation, Medical University Graz, Austria

<sup>11</sup> Institute of Interactive Systems and Data Science & Computer Media  
Graz University of Technology, Austria  
`a.holzinger@hci-kdd.org`

**Abstract.** Explainable AI is not a new field. Since at least the early exploitation of C.S. Pierce’s abductive reasoning in expert systems of the 1980s, there were reasoning architectures to support an explanation function for complex AI systems, including applications in medical diagnosis, complex multi-component design, and reasoning about the real world. So explainability is at least as old as early AI, and a natural consequence of the design of AI systems. While early expert systems consisted of handcrafted knowledge bases that enabled reasoning over narrowly well-defined domains (e.g., INTERNIST, MYCIN), such systems had no learning capabilities and had only primitive uncertainty handling. But the evolution of formal reasoning architectures to incorporate principled probabilistic reasoning helped address the capture and use of uncertain knowledge.

There has been recent and relatively rapid success of AI/machine learning solutions arising from neural network architectures. A new generation of neural methods now scale to exploit the practical applicability of statistical and algebraic learning approaches in arbitrarily high dimensional spaces. But despite their huge successes, largely in problems which can be cast as classification problems, their effectiveness is still limited by

their un-debuggability, and their inability to “explain” their decisions in a human understandable and reconstructable way. So while AlphaGo or DeepStack can crush the best humans at Go or Poker, neither program has any internal model of its task; its representations defy interpretation by humans, there is no mechanism to explain their actions and behaviour, and furthermore, there is no obvious instructional value . . . the high performance systems can not help humans improve.

Even when we understand the underlying mathematical scaffolding of current machine learning architectures, it is often impossible to get insight into the internal working of the models; we need explicit modeling and reasoning tools to explain how and why a result was achieved. We also know that a significant challenge for future AI is contextual adaptation, i.e., systems that incrementally help to construct explanatory models for solving real-world problems. Here it would be beneficial not to exclude human expertise, but to augment human intelligence with artificial intelligence.

**Keywords:** artificial intelligence, machine learning, Explainability, Explainable AI

## 1 Introduction

Artificial intelligence (AI) and machine learning (ML) have recently been highly successful in many practical applications (e.g., speech recognition, face recognition, autonomous driving, recommender systems, image classification, natural language processing, automated diagnosis, . . . ), particularly when components of those practical problems can be articulated as data classification problems. Deep learning approaches, including the more sophisticated reinforcement learning architectures, exceed human performance in many areas [18],[24], [17],[6].

However, an enormous problem is that deep learning methods turn out to be uninterpretable “black boxes,” which create serious challenges, including that of interpreting a predictive result when it may be confirmed as incorrect. For example, consider Figure 1, which presents an example from the Nature review by LeCun, Bengio, and Hinton [15]. The figure incorrectly labels an image of a dog lying on a floor and half hidden under a bed as “A dog sitting on a hardwood floor.” To be sure, the coverage of their image classification/prediction model is impressive, as is the learned coupling of language labels. But the reality is that the dog is *not* sitting.

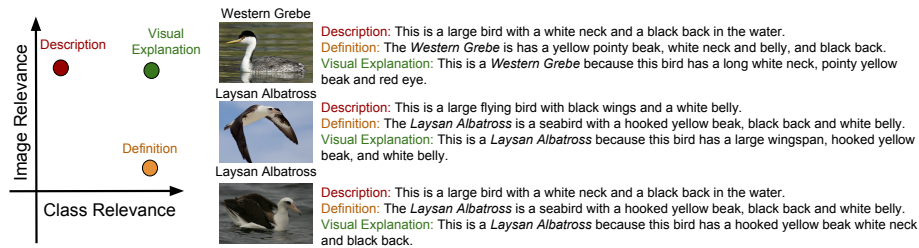
The first problem is the naive but popular remedy about how to debug the predictive classifier to correct the error: augment the original labeled training set with more carefully crafted inputs to distinguish, say, a sitting from a laying dog might improve the incorrect output. This may or may not correct the problem, and doesn’t address the resource challenge of recreating the original learned model.

The transparency challenge gets much more complex when the output predictions are not obviously wrong. Consider medical or legal reasoning, where one typically seeks not just an answer or output (e.g., a diagnostic prediction of



A **dog** is standing on a hardwood floor.

**Fig. 1.** Segment of an example from LeCun, Bengio, Hinton, Science [15]



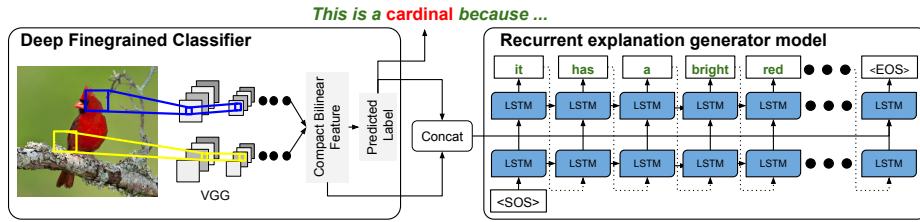
**Fig. 2.** The goal is to generate *explanations* that are both image relevant and class relevant. In contrast, *descriptions* are image relevant, but not necessarily class relevant, and *definitions* are class relevant but not necessarily image relevant.

prostate cancer would require some kind of explanation or structuring of evidence used to support such a prediction). In short, false positives can be disastrous.

Briefly, the representational and computational challenge is about *how* to construct more explicit models of what is learned, in order to support explicit computation that produces a model-based explanation of a predicted output.

However, this is one of the historical challenges of AI: what are appropriate representations of knowledge that demonstrate some veracity with the domain being captured? What reasoning mechanisms offer the basis for conveying a computed inference in terms of that model?

The reality of practical applications of AI and ML in sensitive areas (such as the medical domain) reveals an inability of deep learned systems to communicate effectively with their users. So emerges the urgent need to make results and machine decisions transparent, understandable and explainable [10], [9], [11]. The big advantage of such systems would include not only explainability, but deeper understanding and replicability [8]. Most of all, this would increase acceptance and trust, which is mandatory in safety-critical systems [12], and desirable in many applications (e.g., in medical robotics [19], Ambient Assisted Living [23], Enterprise decision making [4], etc.). First steps have been taken towards making these systems understandable to their users, by providing textual and visual explanations [13], [22] (see Figures 2 and 3).



**Fig. 3.** A joint classification and explanation model [7]. Visual features are extracted using a fine-grained classifier before sentence generation; unlike other sentence generation models, condition sentence generation on the predicted class label. A discriminative loss function encourages generated sentences to include class specific attributes.

## 2 Current State-of-the-Art

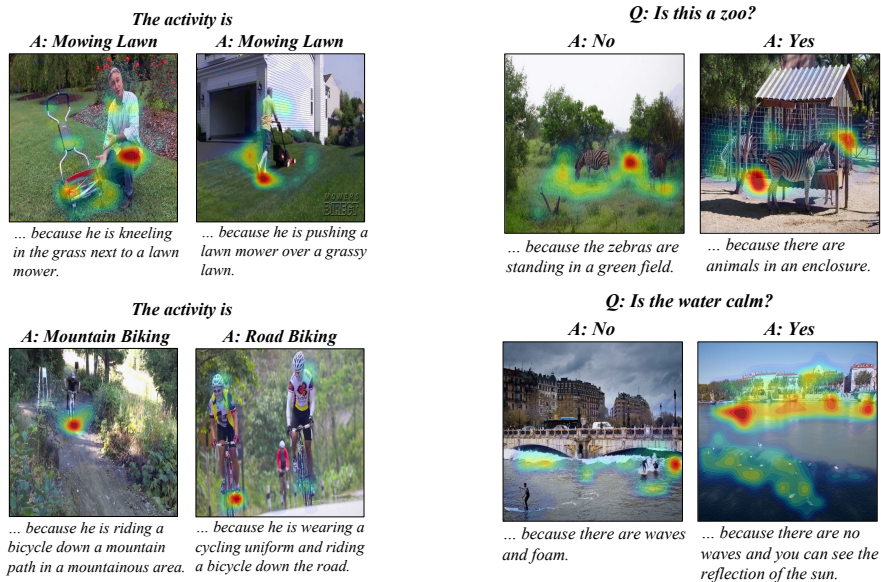
Explaining decisions is an integral part of human communication, understanding, and learning, and humans naturally provide both deictic (pointing) and textual modalities in a typical explanation. The challenge is to build deep learning models that are also able to explain their decisions with similar fluency in both visual and textual modalities (see Figure 2). Previous machine learning methods for explanation were able to provide a text-only explanation conditioned on an image in context of a task, or were able to visualize active intermediate units in a deep network performing a task, but were unable to provide explanatory text grounded in an image.

Existing approaches for deep visual recognition are generally opaque and do not output any justification text; contemporary vision-language models can describe image content but fail to take into account class-discriminative image aspects which justify visual predictions.

Hendriks et al. [7] propose a new model (see Figure 3) that focuses on the discriminating properties of the visible object, jointly predicts a class label, and explains why the predicted label is appropriate for the image. The idea relies on a loss function based on sampling and reinforcement learning, which learns to generate sentences that realize a global sentence property, such as class specificity. This produces a fine-grained bird species classification dataset, and shows that an ability to generate explanations which are not only consistent with an image but also more discriminative than descriptions produced by existing captioning methods.

Although, deep models that are both effective and explainable are desirable in many settings, prior explainable models have been unimodal, offering either image-based visualization of attention weights or text-based generation of post-hoc justifications. Park et al. [21] propose a multimodal approach to explanation, and argue that the two modalities provide complementary explanatory strengths.

Two new datasets are created to define and evaluate this task, and use a model which can provide joint textual rationale generation and attention visu-



**Fig. 4.** Left: ACT-X qualitative results: For each image the PJ-X model provides an answer and a justification, and points to the evidence for that justification. Right: VQA-X qualitative results: For each image the PJ-X model provides an answer and a justification, and points to the evidence for that justification.

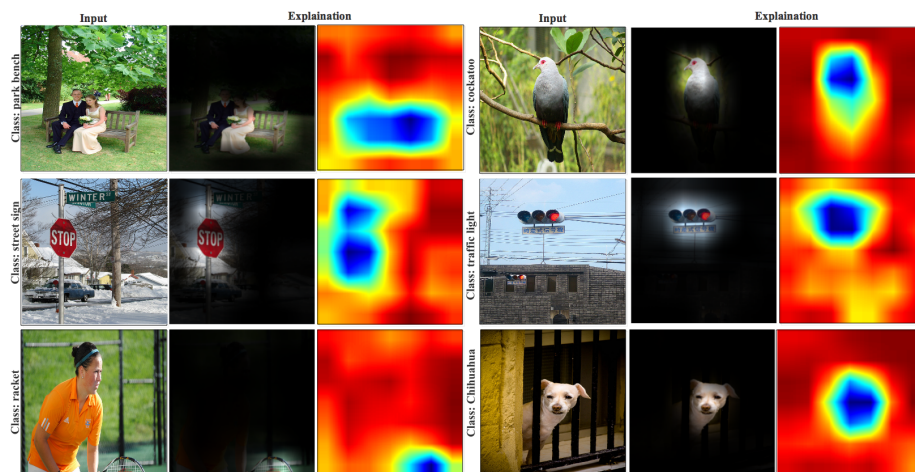
alization (see Figure 4). These datasets define visual and textual justifications of a classification decision for activity recognition tasks (ACT-X) and for visual question answering tasks (VQA-X). They quantitatively show that training with the textual explanations not only yields better textual justification models, but also better localizes the evidence that supports the decision.

Qualitative cases also show both where visual explanation is more insightful than textual explanation, and vice versa, supporting the hypothesis that multimodal explanation models offer significant benefits over unimodal approaches. This model identifies visual evidence important for understanding each human activity. For example to classify “mowing lawn” in the top row of Figure 4 the model focuses both on the person, who is on the grass, as well as the lawn mower. This model can also differentiate between similar activities based on the context, e.g. “mountain biking” or “road biking.”

Similarly, when asked “Is this a zoo?” the explanation model is able to discuss what the concept of “zoo” represents, i.e., “animals in an enclosure.” When determining whether the water is calm, which requires attention to specific image regions, the textual justification discusses foam on the waves.

Visually, this attention model is able to point to important visual evidence. For example in the top row of Figure 2, for the question “Is this a zoo?” the visual explanation focuses on the field in one case, and on the fence in another.

There are also other approaches to explanation that formulate heuristics for creating what have been called “Deep Visual Explanation” [1]. For example, in the application to debugging image classification learned models, we can create a heat map filter to explain where in an image a classification decision was made. There are an arbitrary number of methods to identify differences in learned variable distributions to create such maps; one such is to compute a Kullback-Leibler (KL) divergence gradient, experiments with which are described in [2], and illustrated in (see Figure 5). In that figure, the divergence for each input image and the standard VGG image classification predictor is rendered as a heat map, to provide a visual explanation of which portion of an image was used in the classification.



**Fig. 5.** Explaining the decisions made by the VGG-16 (park bench, street sign, racket, cockatoo, traffic light and chihuahua), our approach highlights the most discriminative region in the image.

### 3 Conclusion and Future Outlook

We may think of an explanation in general as a filter on facts in a context [3]. An effective explanation helps the explainer cross a *cognitive valley*, allowing them to update their understanding and beliefs [4]. AI is becoming an increasingly ubiquitous co-pilot for human decision making. So AI learning systems will require explicit attention to the construction of problem domain models and companion reasoning mechanisms which support general explainability.

Figure 6 provides one example of how we might bridge the gaps between digital inference and human understanding. Deep Tensor [16] is a deep neural

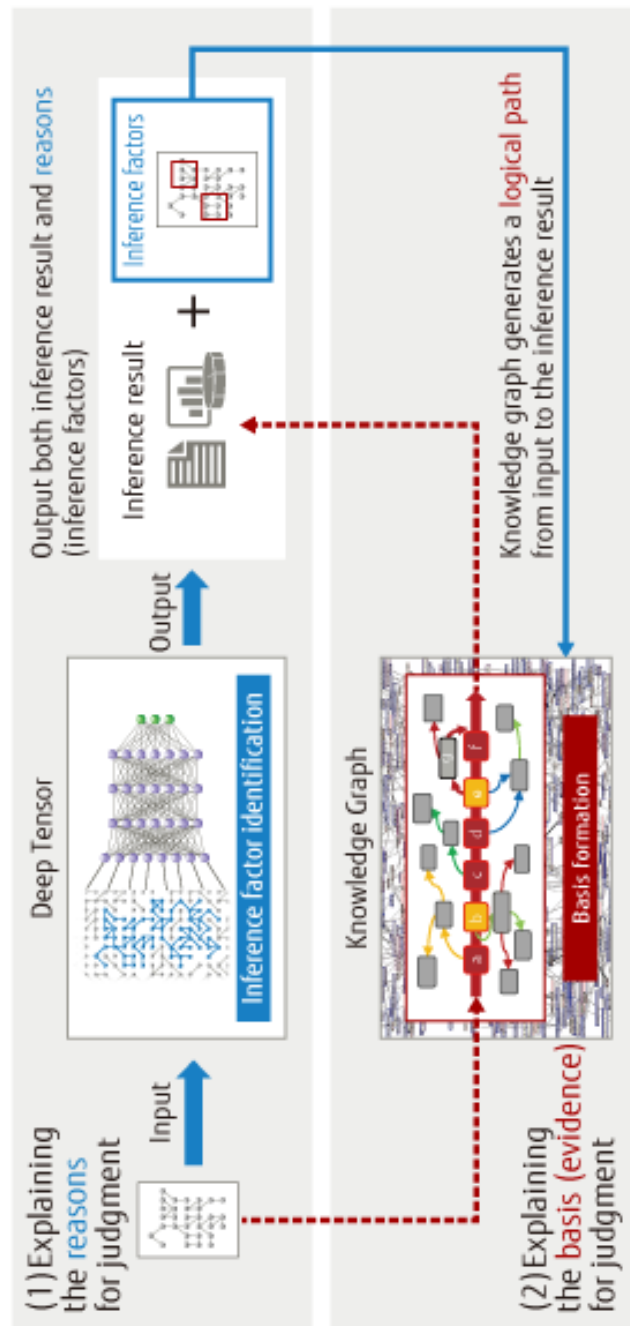


Fig. 6. Explainable AI with Deep Tensor and a knowledge graph

network that is especially suited to datasets with meaningful graph-like properties. The domains of biology, chemistry, medicine, and drug design offer many such datasets where the interactions between various entities (mutations, genes, drugs, disease) can be encoded using graphs. Let's consider a Deep Tensor network that learns to identify biological interaction paths that lead to disease. As part of this process, the network identifies *inference factors* that significantly influenced the final classification result. These influence factors are then used to filter a knowledge graph constructed from publicly available medical research corpora. In addition, the resulting interaction paths are further constrained by known logical constraints of the domain, biology in this case. As a result, the classification result is presented (explained) to the human user as an annotated interaction path, with annotations on each edge linking to specific medical texts that provide supporting evidence.

Explanation in AI systems is considered to be critical across all areas where machine learning is used. There are examples which combine multiple architectures, e.g., combining logic-based system with classic stochastic systems to derive human-understandable semantic explanations [14]. Another example is in the case of transfer learning [20], where learning complex behaviours from small volumes of data is also in strong needs of explanation of efficient, robust and scalable transferability [5].

## Acknowledgements

The authors thanks their colleagues from local and international institutions for their valuable feedback, remarks and critics on this introduction to the MAKE-Explainable-AI workshop.

## References

1. Housam Khalifa Bashier Babiker and Randy Goebel. An introduction to deep visual explanation. *NIPS 2017 - Workshop Interpreting, Explaining and Visualizing Deep Learning*, 2017.
2. Housam Khalifa Bashier Babiker and Randy Goebel. Using kl-divergence to focus deep visual explanation. *CoRR*, abs/1711.06431, 2017.
3. Ajay Chander and Ramya Srinivasan. Evaluating explanations. In *Joint Proceedings of the IFIP Cross-Domain Conference for Machine Learning and Knowledge Extraction (IFIP CD-MAKE 2018)*, 2018.
4. Ajay Chander, Ramya Srinivasan, Suhas Chelian, Jun Wang, and Kanji Uchino. Working with beliefs: AI transparency in the enterprise. In *Joint Proceedings of the ACM IUI 2018 Workshops co-located with the 23rd ACM Conference on Intelligent User Interfaces (ACM IUI 2018)*, 2018.
5. Jiaoyan Chen, Freddy Lecue, Jeff Z. Pan, Ian Horrocks, and Huajun Chen. Transfer learning explanation with ontologies. In *Principles of Knowledge Representation and Reasoning: Proceedings of the Eleventh International Conference, KR 2018, 30 October-2 November 2018, Tempe, Arizona (USA)*, page to appear, 2018.

6. Andre Esteva, Brett Kopley, Roberto A. Novoa, Justin Ko, Susan M. Swetter, Helen M. Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, 542(7639):115–118, 2017.
7. Lisa Anne Hendricks, Zeynep Akata, Marcus Rohrbach, Jeff Donahue, Bernt Schiele, and Trevor Darrell. Generating visual explanations. In *European Conference of Computer Vision - ECCV 2016*, 2016.
8. Andreas Holzinger, Chris Biemann, Constantinos S. Pattichis, and Douglas B. Kell. What do we need to build explainable ai systems for the medical domain? *arXiv:1712.09923*, 2017.
9. Andreas Holzinger, Bernd Malle, Peter Kieseberg, Peter M. Roth, Heimo Mller, Robert Reihls, and Kurt Zatloukal. Towards the augmented pathologist: Challenges of explainable-ai in digital pathology. *arXiv:1712.06657*, 2017.
10. Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M. Pintea, and Vasile Palade. Towards interactive machine learning (iml): Applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In *Springer Lecture Notes in Computer Science LNCS 9817*, pages 81–95. Springer, Heidelberg, Berlin, New York, 2016.
11. Andreas Holzinger, Markus Plass, Katharina Holzinger, Gloria Cerasela Crisan, Camelia-M. Pintea, and Vasile Palade. A glass-box interactive machine learning approach for solving np-hard problems with the human-in-the-loop. *arXiv:1708.01104*, 2017.
12. Katharina Holzinger, Klaus Mak, Peter Kieseberg, and Andreas Holzinger. Can we trust machine learning results? artificial intelligence in safety-critical decision support. *ERCIM News*, 112(1):42–43, 2018.
13. Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, pages 126–137. ACM, 2015.
14. Freddy Lécué and Jiewen Wu. Semantic explanations of predictions. *CoRR*, abs/1805.10587, 2018.
15. Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521:436 EP –, 05 2015.
16. Koji Maruhashi, Masaru Todoriki, Takuya Ohwa, Keisuke Goto, Yu Hasegawa, Hiroya Inakoshi, and Hirokazu Anai. Learning multi-way relations via tensor decomposition with neural networks. In *The Thirty-Second AAAI Conference on Artificial Intelligence AAAI-18*, pages 3770–3777, 2018.
17. Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.
18. Matej Moravčík, Martin Schmid, Neil Burch, Viliam Lisý, Dustin Morrill, Nolan Bard, Trevor Davis, Kevin Waugh, Michael Johanson, and Michael Bowling. Deepstack: Expert-level artificial intelligence in heads-up no-limit poker. *Science*, 2017.
19. Shane OSullivan, Andreas Holzinger, Kurt Zatloukal, Paulo Saldiva, Muhammad I. Sajid, and Wichmann Dominic. Machine learning enhanced virtual autopsy. *Autopsy Case Report*, 7(4):3–7, 2017.
20. Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

21. Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.
22. Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 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*, pages 1135–1144. ACM, 2016.
23. Deepika Singh, Erinc Merdivan, Ismini Psychoula, Johannes Kropf, Sten Hanke, Matthieu Geist, and Andreas Holzinger. Human activity recognition using recurrent neural networks. In Andreas Holzinger, Peter Kieseberg, A Min Tjoa, and Edgar Weippl, editors, *Machine Learning and Knowledge Extraction, CD-MAKE, Lecture Notes in Computer Science LNCS 10410*, pages 267–274. Springer International, Cham, 2017.
24. Yaniv Taigman, Ming Yang, Marc’ Aurelio Ranzato, and Lior Wolf. Deepface: Closing the gap to human-level performance in face verification. pages 1701–1708, 2014.