Explainable AI: The New 42?

Randy Goebel[1](✉), Ajay Chander[2], Katharina Holzinger[3], Freddy Lecue[4,5],
Zeynep Akata[6,7], Simone Stumpf[8], Peter Kieseberg[3,9],
and Andreas Holzinger[10,11]

[1] Alberta Machine Intelligence Institute, University of Alberta, Edmonton, Canada
rgoebel@ualberta.ca
[2] Fujitsu Labs of America, Sunnyvale, USA
achander@us.fujitsu.com
[3] SBA-Research, Vienna, Austria
kholzinger@sba-research.org
[4] INRIA, Sophia Antipolis, France
freddy.lecue@inria.fr
[5] Accenture Labs, Dublin, Ireland
[6] Amsterdam Machine Learning Lab,
University of Amsterdam, Amsterdam, The Netherlands
[7] Max Planck Institute for Informatics, Saarbruecken, Germany
z.akata@uva.nl
[8] City, University of London, London, UK
Simone.Stumpf.1@city.ac.uk
[9] University of Applied Sciences St. Pölten, St. Pölten, Austria
Peter.Kieseberg@fhstp.ac.at
[10] Holzinger Group HCI-KDD, Institute for Medical Informatics, Statistics and
Documentation, Medical University Graz, Graz, Austria
[11] Institute of Interactive Systems and Data Science,
Graz University of Technology, Graz, 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 arises from neural network architectures. A new generation of neural methods now scale to exploit the practical applicability

A. Holzinger et al. (Eds.): CD-MAKE 2018, LNCS 11015, pp. 295–303, 2018.
https://doi.org/10.1007/978-3-319-99740-7_21

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 [6,17,18,24].

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 Fig. 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.

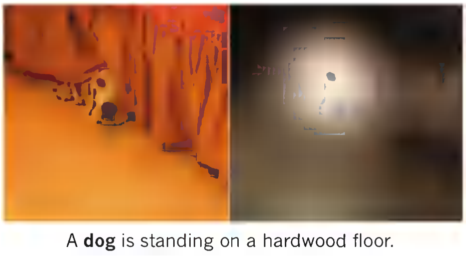


Fig. 1. Segment of an example from LeCun, Bengio, Hinton, Science [15]

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 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.

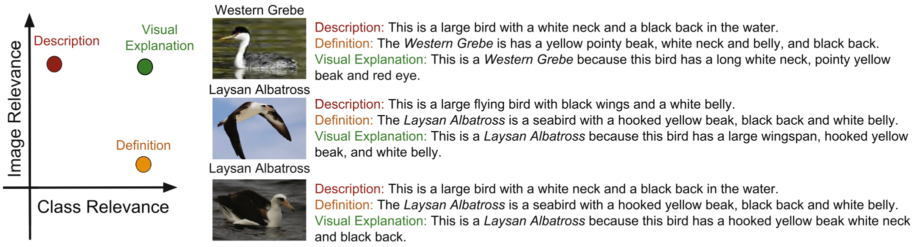
Briefly, the representational and computational challenge is about *how* to construct more explicit models of what is learned, in order to support explicit computation that produces a model-based explanation of a predicted output.

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 [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 Figs. 2 and 3).

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 Fig. 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.



Description: This is a large bird with a white neck and a black back in the water.
Definition: The *Western Grebe* is has a yellow pointy beak, white neck and belly, and black back.
Visual Explanation: This is a *Western Grebe* because this bird has a long white neck, pointy yellow beak and red eye.

Description: This is a large flying bird with black wings and a white belly.
Definition: The *Laysan Albatross* is a seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a large wingspan, hooked yellow beak, and white belly.

Description: This is a large bird with a white neck and a black back in the water.
Definition: The *Laysan Albatross* is a seabird with a hooked yellow beak, black back and white belly.
Visual Explanation: This is a *Laysan Albatross* because this bird has a hooked yellow beak white neck and black back.

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.

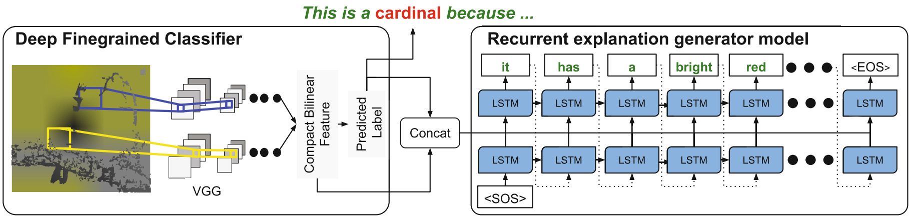
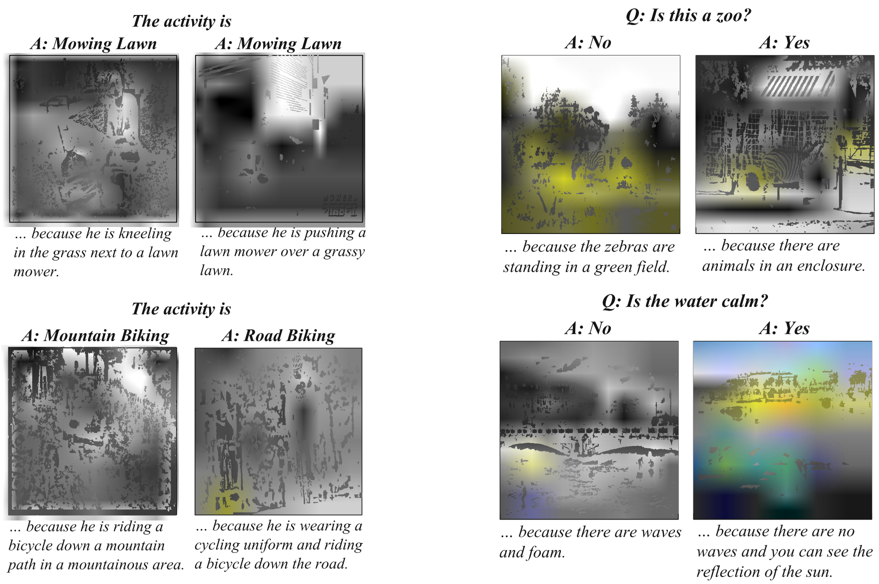

This is a cardinal because ...
Deep Finegrained Classifier
VGG
Compact Bilinear Feature
Predicted Label
Concat
Recurrent explanation generator model
it
has
a
bright
red
<EOS>
LSTM
<SOS>

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.

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 Fig. 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



The activity is

A: Mowing Lawn — *... because he is kneeling in the grass next to a lawn mower.*

A: Mowing Lawn — *... because he is pushing a lawn mower over a grassy lawn.*

Q: Is this a zoo?

A: No — *... because the zebras are standing in a green field.*

A: Yes — *... because there are animals in an enclosure.*

The activity is

A: Mountain Biking — *... because he is riding a bicycle down a mountain path in a mountainous area.*

A: Road Biking — *... because he is wearing a cycling uniform and riding a bicycle down the road.*

Q: Is the water calm?

A: No — *... because there are waves and foam.*

A: Yes — *... because there are no waves and you can see the reflection of the sun.*

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.

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 visualization (see Fig. 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 Fig. 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 Fig. 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 Fig. 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.

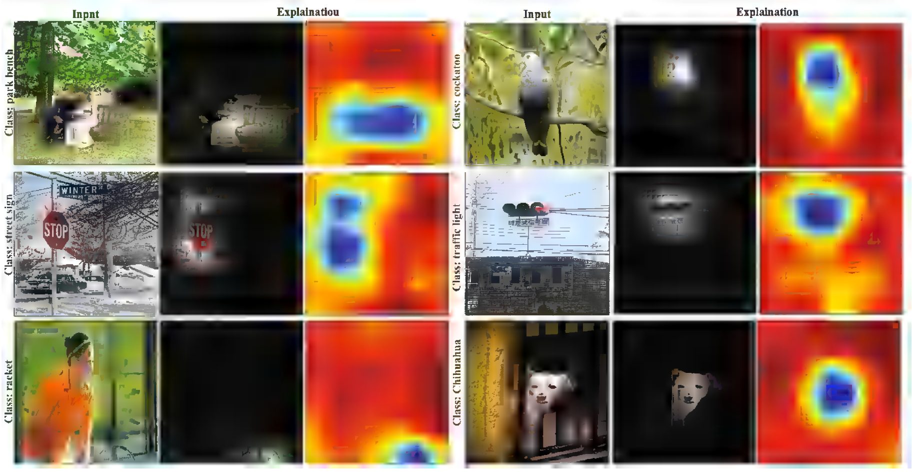


Fig. 5. Explaining the decisions made by the VGG-16 (park bench, street sign, racket, cockatoo, traffic light and chihuahua), our approach highlights the most discriminative region in the image.

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

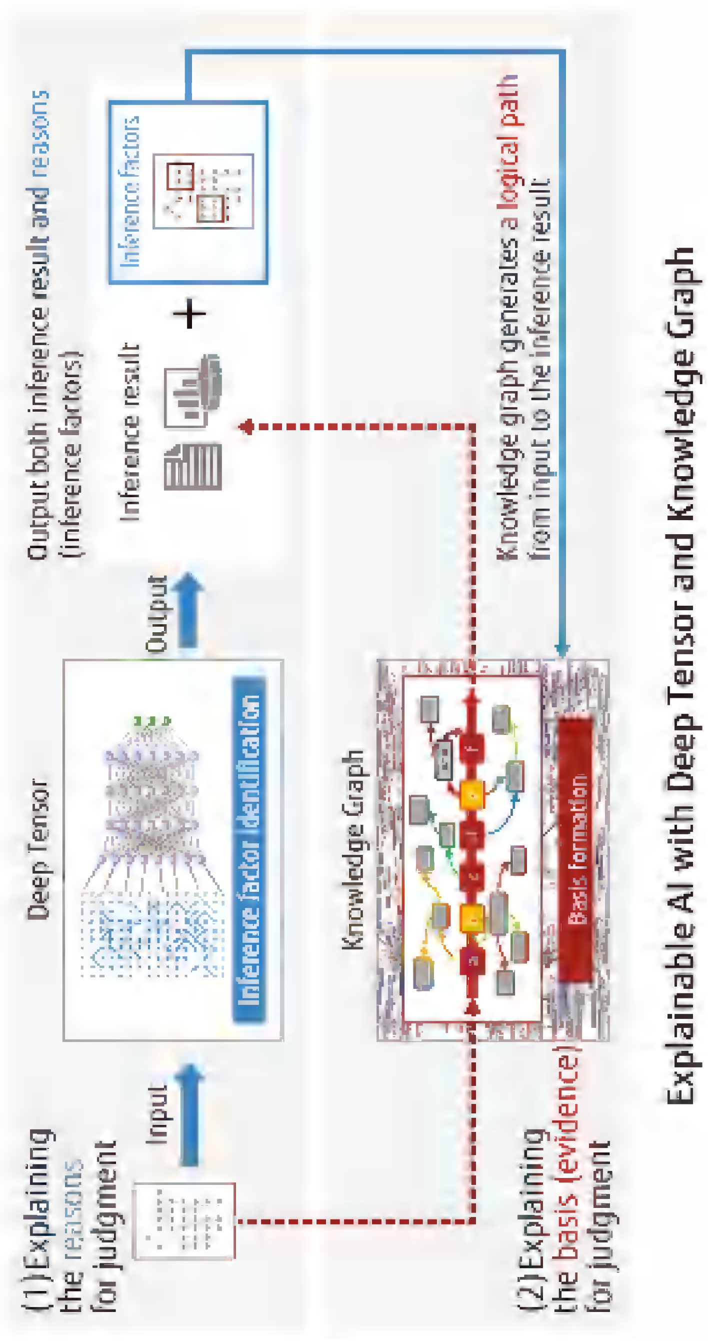

(1) Explaining the reasons for judgment
Input
Deep Tensor
Inference factor identification
Output
Output both inference result and reasons (inference factors)
Inference result
+
Inference factors
Knowledge Graph
Basis formation
Knowledge graph generates a logical path from input to the inference result
(2) Explaining the basis (evidence) for judgment
Explainable AI with Deep Tensor and Knowledge Graph

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

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 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. Babiker, H.K.B., Goebel, R.: An introduction to deep visual explanation. In: NIPS 2017 - Workshop Interpreting, Explaining and Visualizing Deep Learning (2017)
2. Babiker, H.K.B., Goebel, R.: Using KL-divergence to focus deep visual explanation. CoRR, abs/1711.06431 (2017)
3. Chander, A., Srinivasan, R.: Evaluating explanations. In: Joint Proceedings of the IFIP Cross-Domain Conference for Machine Learning and Knowledge Extraction (IFIP CD-MAKE 2018) (2018)
4. Chander, A., Srinivasan, R., Chelian, S., Wang, J., Uchino, K.: 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. Chen, J., Lecue, F., Pan, J.Z., Horrocks, I., Chen, H.: 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) (2018, to appear)

6. Esteva, A., et al.: Dermatologist-level classification of skin cancer with deep neural networks. Nature **542**(7639), 115–118 (2017)
7. Hendricks, L.A., et al.: Generating visual explanations. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 3–19. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_1
8. Holzinger, A., Biemann, C., Pattichis, C.S., Kell, D.B.: What do we need to build explainable AI systems for the medical domain? arXiv:1712.09923 (2017)
9. Holzinger, A., et al.: Towards the augmented pathologist: challenges of explainable-AI in digital pathology. arXiv:1712.06657 (2017)
10. Holzinger, A., et al.: Towards interactive Machine Learning (iML): applying ant colony algorithms to solve the traveling salesman problem with the human-in-the-loop approach. In: Buccafurri, F., Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-ARES 2016. LNCS, vol. 9817, pp. 81–95. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-45507-5_6
11. Holzinger, A., et al.: A glass-box interactive machine learning approach for solving NP-hard problems with the human-in-the-loop. arXiv:1708.01104 (2017)
12. Holzinger, K., Mak, K., Kieseberg, P., Holzinger, A.: Can we trust machine learning results? Artificial intelligence in safety-critical decision support. ERCIM News **112**(1), 42–43 (2018)
13. Kulesza, T., Burnett, M., Wong, W.-K., Stumpf, S.: Principles of explanatory debugging to personalize interactive machine learning. In: Proceedings of the 20th International Conference on Intelligent User Interfaces, pp. 126–137. ACM (2015)
14. Lécué, F., Wu, J.: Semantic explanations of predictions. CoRR, abs/1805.10587 (2018)
15. LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436 (2015)
16. Maruhashi, K., et al.: Learning multi-way relations via tensor decomposition with neural networks. In: The Thirty-Second AAAI Conference on Artificial Intelligence AAAI-18, pp. 3770–3777 (2018)
17. Mnih, V., et al.: Human-level control through deep reinforcement learning. Nature **518**(7540), 529–533 (2015)
18. Moravčík, M.: Deepstack: expert-level artificial intelligence in heads-up no-limit poker. Science **356**(6337), 508–513 (2017)
19. O'Sullivan, S., et al.: Machine learning enhanced virtual autopsy. Autopsy Case Rep. **7**(4), 3–7 (2017)
20. Pan, S.J., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
21. Park, D.H., et al.: Multimodal explanations: justifying decisions and pointing to the evidence. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2018)
22. Ribeiro, M.T., Singh, S., Guestrin, C.: 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, pp. 1135–1144. ACM (2016)
23. Singh, D., et al.: Human Activity recognition using recurrent neural networks. In: Holzinger, A., Kieseberg, P., Tjoa, A.M., Weippl, E. (eds.) CD-MAKE 2017. LNCS, vol. 10410, pp. 267–274. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-66808-6_18
24. Taigman, Y., Yang, M., Ranzato, M.A., Wolf, L.: Deepface: Closing the gap to human-level performance in face verification, pp. 1701–1708 (2014)