BRIDGING THE GAP BETWEEN AI AND HUMAN UNDERSTANDING

[a]Abhishek Kumar Jha, [b]Praveen Kumar Pandey, [c]Vishesh Poswal

[a]B Com Student, School of Commerce and Management , Lingaya's Vidyepeeth , Faridabad

[b]Research Assistant, Faculty of Management Studies, Manav Rachna International Institute of Research and Studies, Faridabad

[c]Assistant Professor, Aggarwal College, Ballabhgarh

ABSTRACT

In current era, a huge amount of data or information of our personal life is being collected by through various sources. These sources help an Artificial intelligence (AI) to know about user preferences through various algorithm. The aim of Artificial intelligence (AI) aimsto gain new knowledge for making decision properly. Artificial intelligence has taken a vital and crucial role in life ofevery person who is connected to any. We know that response of any human to a conversation contains many constraints like feelings, personal judgement, thinking and many more. If these constraints emerged with artificial intelligence, then it will be one of the greatest discoveries in the history. Because it will enhance our way of thinking in each sector and field. It will allow to find new knowledge into their understanding. In this research paper we build an informative step which will makes a step closer to our main objective which is bridging the gap between Artificial intelligence and human understanding, as well as compare the existing paper which is already been researched.

Keywords: constraints, algorithm, discovery, human understanding, Enhance, Preferences

INTRODUCTION

In today's era, developments in electronics and computers such as sensors, natural language system, neural network, intelligent retrieval, knowledge engineering have led to the creation of huge amounts of data (Liu, 2014), (J. R. F. Díaz, 2011). Thesemountains of information often emerge and grow at a rapid rate, are complex in nature, have complex processes, and are large in size beyond the capabilities of modern technology to capture (Tufféry, 2011), manage, and process information in a timely manner. Recently, he envisions the word "big data" to cover this phenomenon in all its aspects with a special concept (Gantz & Reinsel, 2012). Moreover, the task of presenting and/or representing information in a clear, understandable and powerful way is not easy because this information is often fragmented and comes from many sources. (Ball, 2012). Additionally, when data is combined, new data will emerge that may have different and unpredictable patterns.

Another thing to consider is the different sizes of the material, here called dimensions. For example, information created through complex processes is often available immediately but retained for long periods of timetherefore, this knowledge can be frustrating and often has not yet been explored as the basis for creating new knowledge. However, it can be said that as the product grows, existing tools must also evolve. The new technology then combines the sources of the material, including the results of human analysis and interpretation (Kim & Lee, 2014) (Theron & Fontanillo, 2015). The combination of artificial intelligence and natural intelligence is a method suitable for the search for information by creating a holistic view (Timaran, 2005) (Bertini & Lalanne, 2009). In this sense, communication with visual information is an important tool that allows reaching more information and ideas. BT. and then turn the machine to valuable items that partners cannot easily see, in the process of searching for patterns and events (Y. Wang & Li, 2014). Perceiving patterns can often form the basis of a new vision, so data visualization (DV) helps understand the characteristics of data big and small. The DV phase is often important for generating data analysis data when the main assumptions of the data need to be made first. In fact, although modern and powerful data processing tools exist, only some of them can be used without prior knowledge about the data, and this is important in terms of choosing the best models, data evaluation methods and algorithms. important. In this case, DV can help create the initial idea of the model of the material. Additionally, visualizing information is especially useful for non-expert users and will have an impact on vision (Gibson, 2014). In this paper, we describe two data mining methods that represent two independent learning methods using computer processing and using visual analysis using human reasoning, that is, integrating wisdom with. This integration can be done through the knowledge discoveryprocess in the database. This study briefly reviews methods for integrating intelligence and intelligence and provides an overview of data mining and data visualization.

DATA ANALYZING: FROM THE HUMAN BEHAVIOR

Every person in the world is surrounded by Artificial Intelligence, from the beginning they connected to smart devices through network. Billions of people think or find it efficient and effective to use Artificial Intelligence to make their work or daily life more convenient, which is surprising. Although, we have seen or came across that social media and many websites uses our personal data in the form of cookies and uses them as to understand the human behavior. With the help of these cookies AI creates an interface from various algorithms. These algorithms run various tests to get the knowledge of what a user actually wants. Many studies have stated that the AI and Machine Learning have obtained understanding into human behavior and psychology without concern over social media. The pattern of like sets you up with the content you consume, resulting in targeted ads and personalized content.

Artificial Intelligence merge with the neural networks plays a crucial role in identifying the relationships in a set of data via a process that imitate how the human brain works. The data collected over the years contains everything from the usage maps or patterns to the details. In simple words, more you spend time online, the larger your digital footprint becomes.

THE GAP BETWEEN AI SYSTEMS AND HUMAN UNDERSTANDING

AI and human share some resemblance, but there are several differences between those two. AI system can do tasks which generally requires human knowledge, although they lack emotional intelligence, creativity which a human possess. Human understanding is a complex process of genetic code for experiencing of different environments which allows human to understand and interact with the surroundings and world which an AI cannot make a replica. As we know that AI continues to develop, it is important to understand the unique strength and limitations of both AI and human understanding.

Below given differences will give light on the better understanding of AI and human understanding:

RESEARCH QUESTION AND OBJECTIVES

Some questions which arise when we talk about AI and human:

- What is Artificial Intelligence?
- What is human intelligence?
- What AI cannot do without – The "Human" Factor?
- What future holds?
- Do you think artificial intelligence will dominate human intelligence?
- How can we ensure the responsible use of artificial intelligence and prevent potential negative consequences?

Characteristics	**Artificial Intelligence**	**Human Understanding**
Origin	AI is an innovation created by human intelligence; its early development is credited to Norbert Weiner who theorized on feedback mechanisms while the father of AI is John McCarthy for coining the term and organizing the first conference on research projects regarding machine intelligence.	Human beings are created with the innate ability to think, reason, recall, etc.
Speed	As compared to humans, computers can process more information at a faster rate. For instance, if the human mind can solve a math problem in 5 minutes, AI can solve 10 problems in a	Human mind is slower than AI.

	minute.	
Decision Making	AI is highly objective in decision making as it analyses based on purely gathered data.	However, humans' decisions may be influenced by subjective elements which are not based on figures alone.
Accuracy	AI often produces accurate results as it functions based on a set of programmed rules.	As for human intelligence, there is usually a room for "human error" as certain details may be missed at one point or the other.
Energy Used	Modern computers only generally use 2 watts.	The human brain uses about 25 watts
Adaptation	AI takes much more time to adapt to new changes.	Human intelligence can be flexible in response to the changes to its environment. This makes people able to learn and master various skills.

Above are some questions which generally come in mind when we discuss about Artificial Intelligence.

The main objective of this research **is Bridging the Gap between Artificial Intelligence**.

THEORITICAL FRAMEWORK

Now let's discuss some theories and models on human cognition and understanding:

Cognition is actually the mental state process and action of gaining or acquiring knowledge and understanding through thought, experience, and the senses. In simple words it is the mental processes relating to the input and storage of information and how that information is them used to guide our behavior.

Cognition theory tells that the human mind is like a computer which is constantly processing and encoding data. According to this theory, when a person experiences stimuli their mind will simply look toward to understand this information.Models of human cognition which hold that information processing begin in series of various stages. Models of human cognition are based on the qualitative analysis and empirical rules, but providingaccurate digital representation of the process of cognition is difficult. The cognition model is to make prediction regarding the behavior of IT (information technology) interaction between human and machine.

HUMAN COGNITION HELPS IN DEVELOPING AI

We know that human are biological organisms that have limited lifetime, also a limited amount of which we carry around inside our heads, and which have limited capacity for communication. But there are many characteristics which makes us better than AI which is expressing our thoughts which contains various emotions. We can certainly use our minds to develop various algorithms which improve the efficiency of artificial intelligence. AI can do may data-based task more efficiently than any human but human intelligence can work on creative, emotional and complex work. Since it is proved that artificial intelligence cannot replace human intelligence entirely. Although AI has potential to bring as much positivity in the society, including enhanced productivity, improved health system and enhanced education system in every field. There are many technologies which created by humans which provide support or improve their results.

EXPLAINABILITY

AI is set of frameworks and tools which helps us to understand and interpret the prediction which is made by machines intelligence. With the helps of various chains of data we can improve AI performance, and help ourselves with many tasks and works. AI can help humans to understand and explains machine language with deep learning and neural network. Explainability is important because being able to interpret a machine model increases trust in the model, which is essential in scenarios like financial, health care and life and death decisions.

INTERPRETABILITY

Interpretability in artificial intelligence refers to the study of how to understand the decision of machine learning systems whose decisions are easily understood, or interpretable. The machine model itself itself has come to the source of knowledge captured by these models. There are some methods from which we can evaluate the interpretability method these are: application grounded, human-grounded, and functional grounded. This evaluation is basically done or perform by various experiments within a real-life application.

TRANSPARENCY

AI transparency helps ensure that every stakeholder or users can clearly understand the working of an AI system, including how does it makes decision and analyze data. The more we analyze the data the more we get clarity on AI system. Transparency in AI requires a multi-faced approach. Using various algorithms, provide clear documentation, and helps to maintain the record of data sources and models. There are basically the levels in transparency and these are: algorithmic transparency, interaction transparency, social transparency. Transparency helps to gain public trust and confidence in AI system. The trust is crucial, considering the potential impact of AI on sensitive areas like healthcare, finance, and criminal justice.

AI TECHNOLOGIES AND APPROACHES

AI technology have been growing rapidly and developing into various aspects. Like machine learning techniques which helps to in some application like image recognition, speech recognition and many more. And natural language processing allows machines to comprehend, interpret and generate human language. This AI technique has played a virtual role while assisting user through many chatbots, tools through a network with flawless decisions. Same as there many applications of AI in various industries some of them are:

- **HEALTHCARE**
  In healthcare system, AI played a very crucial role in development of personalized treatments. As we know there are some applications which provide treatment to patient with the help of machine language and provide medicine to the user.
- **FINANCE**
  AI has changed the financial network with help of technologies like trading, fraud detection, and personal financial advice. This technology has the ability to analyze the large financial data to empower various institutions to make informed decisions.
- **RETAIL**
  For the retail industry, AI has opened many doors to target marketing, demand forecasting, and customer service chatbots. These advancements have enhanced the overall shopping experience for consumers while analyzing operations for retailers.
- **TRANSPORTATION**
  AI has played a very crucial role for the transportation sector which have helped the automobile vehicle, traffic operations and provide a much safer and efficient transportation system.

As business has grow more in recent years with the help of AI which provide a better understanding to users. AI provide many strategies which is helpful for business in many circumstances, it also gives you an early advantage in the coverage. Embrace the future technologies with AI to led a better future for upcoming generation.

APPROACHES AND TECHNIQUES FOR ENHANCING AI'S UNDERSTANDING

Enhancing AI's understanding same as human concept is a complex area which is an ongoing area of research. There are several techniques and approaches have gained the liberty for achieving goal. Some of the methods are:

- **MACHINE LEARNING ALGORITHM**
  Machine learning algorithms are some mathematical model mapping methods used to learn or uncover the underlying pattern embedded in the data. ML algorithm is a set of mathematical processes or techniques by which an artificial intelligence (AI) system conducts its tasks. these tasks include gleaning important insights, patterns and predictions about the future from input data the algorithm is trained on. a data science professional feeds an ml algorithm training data so it can learn from that data to enhance its decision-making capabilities and produce desired outputs.

  ML is a subset of ai and computer science. its use has expanded in recent years along with other areas of ai, such as deep learning algorithms used for big data and natural language processing for speech recognition. what makes ml algorithms important is their ability to sift through thousands of data points to produce data analysis outputs more efficiently than humans.
- **NATURAL LANGUAGE PROCESSING**

Natural language processing (NLP) is a branch of artificial intelligence that enables computers to comprehend, generate, and manipulate human language. Natural language processing has the ability to interrogate the data with natural language text or voice. Most of the users have already interacted with the NLP without realizing it. Basically NLP is the core of technology behind the virtual assistants, such Siri, Cortana or Alexa. When we ask questions to them, NLP help them to not on

- **COMPUTER VISION**
  Computer vision equips machine with the ability to interpret visual information from the world. The technique has revolutionized industries like healthcare, automotive, and robotics, enabling task such as facial recognition, object detection, and autonomous driving. It is generally a field of computer science which focuses on enabling computer to identify and understand objects and people in images and videos. Like others types of AI, computer vision seeks to perform and automate tasks that replicate human capabilities.

- **DEEP LEARNING**
  Deep learning takes ML to a higher level by employing neural networks with multiple layers to process complex data representations. It has propelled AI achievements, such as beating human champions in games like chess and go and enhancing image and speech recognition systems. It teaches computers to process data in a way that is inspired by the human brain. Deep learning models can recognize complex patterns in pictures, text, sounds, and other data to produce accurate insights and predictions.

LITREATURE REVIEW

There are several studies have been conducted for understand the relationship of AI and human understanding (for example, Khan Jasim Mohammed Asif 2023; Hayder Mahdi A Al Dawood 2021; Juan C. ALVARADO-PÉREZa, Diego H. PELUFFO-ORDÓÑEZb, Roberto THERONc 2015; Imke van Heerden 2021).

There are various studies which helps the other studies to understand algorithms, methodology, and strategies. Khan Jasim Mohammed Asif research emphasizes the performance analysis with incorporating traditional metrics for predictive accuracy and interpretability specific metrics, ensures that interpretable machine learning models achieve satisfactory predictive performance while providing transparent explanation. The study of Hayder Mahdi A Al Dawood emphasizes the results which indicate the need for standardized rules and an oversight body to ensure that ethics and accountability are implemented. Therefore considering establishing local and regional standards and certification for the ethics of AI, such as other certifications as ISO. The study of Juan C. ALVARADO-PÉREZa, Diego H. PELUFFO-ORDÓÑEZb, Roberto THERONc emphasizes the human knowledge and machine learning in which he elaborate the the performance which is strongly dependent on the nature of the data as well as the right use of the data processing techniques. It is important to remark that due to the high demand of computational load, integration approaches are often implemented on parallel and distributed architectures such as computer clusters and grids. The study of Imke van Heerden emphasizes on drawing the expertise in the humanities, primarily literature theory, to contribute to the development of computer science, specifically AI writing. To achieve human level creativity, machine-generated literature has to overcome various obstacles.

RESEARCH METHODOLOGY

The research methodology which is used in the study is mainly depend on the secondary data mostly collected through the various sources like Google scholar using the keywords such as 'machine language', 'artificial intelligence', 'natural language processing', 'human understanding'. To data is collected from the published papers between 2010 to 2023.

Additionally a qualitative approach has been used in this paper to determine the findings of the research papers. After the analysis the conclusion is determined on the basis of the other results from papers. There are several terminologies were used to understand the results more effectively. At last this research methodology is aims to provide the best data on the basis of several exploration of data through various respectable sources.

CHALLENGES AND ETHICAL CONSIDERATIONS

Achieving the human level understanding with AI is the aim or ultimate goal of many developers, but it's still a distant goal in technical challenges. There are some hurdles which should to consider is, Replicating human senses and perception, Mastering commonsense reasoning and grounding, Handling natural language with human fluency, Embracing open-ended learning and adaption, integrating emotions and social intelligence.

The sensory input, from vision to touch, is incredibly rich. Current AI systems, specially visions based excel at specific task. AI system struggle with his commonsense reasoning, often vast amounts of labeled data, explicit rules, and programming for even basic trick.

While AI has made a long step in natural language processing, it still stumbles with sarcasm, ambiguity, and several expression. Unlike most AI models trained on specific datasheets, human continuously learn and adapt throughout their lives. Which requires innovative approaches to learning algorithms and knowledge representations. AI currently lacks the ability to react to emotional states, and ability to truly understand and interact with human on a meaningful level.

Addressing the technical challenges will require many studies and innovative ideas on computer science.

HUMAN-AI COLLABORATION

Human-AI collaboration (HAC) is rapidly evolving as a critical set of ideas for enhancing human capabilities and achieving greater levels of understanding and problem-solving across the several domains. Human strengths is reasoning and judgment, creative and intuition, social intelligence, ethics and values. AI strengths is data processing, automation and precision, objectivity.

By combining these strengths HAC offers several benefits like enhanced decision-making, improved problem solving, efficient workflows, innovations and advancements. Although apart from the potential there are certain challenges which is mutual understanding, human-AI interaction design, bias and fairness, job displacements. Human-AI collaboration is a tangible approach shaping the future of various fields. By finding the unique strength, HAC contains various potential qualities like enhancing human capabilities, tackling complex challenges, and driving human capabilities.When a human uses and ai's output, they often 1 to understand why a model give a certain outputhuman AI collaboration leverages the strength of both human and AI to achieve shared objectives. By combining human intoon and creativity with a eyes computational power and data analysis we can unlock remarkable opportunities for innovation across industries. Transparency and ethical consideration are essential for successful collaboration. Although some perceive this development as potential threat, capable of replacing humans in their jobs others invision it as the rise of new era where human and AI collaborate to unlock extraordinary opportunities innovations and break through that would not be possible otherwise. Mean while, AI system can offer computational power to process fast amount of data and extract valuable insights and patterns that may remain hidden from human perceptions.

APPLICATIONS AND IMPACT

Some Present Cases Studies:

- **Medical diagnosis:** AI assists doctors in analyzing medical images and patient data for faster and more accurate diagnoses.
- **Scientific research:** AI helps scientist analyze massive datasets and identify the promising researches.
- **Cyber security:** AI-powered systems detect and prevent cyber threats, while human expertise remains crucial for complex incident response.
- **Climate change mitigation:** AI models predict climate patterns and inform strategies for renewable energy development and environmental protection.

The impact of AI and human understanding will enhance human information processing capabilities, enabling analysis of vast data sets, pattern recognition, and identification of complex relationships. Which can lead to breakthrough in science, medicine, and other fields. AI powered system can personalize learning, recommendation, and experiences which is based on individual preferences and needs.

Which can improve educational outcomes outcome, enhance user engagements, and tailor services to better suit individuals. Overall, the impact of human understanding and AI is a dynamic and evolving interplay with both positive and negative potential.

FUTURE DIRECTIONS

The future directions of the relationship between AI and human understanding promises to be exciting and transformative marked by both collaboration and ongoing challenges there are some potential pathway we might observe:

DEEPNING COLLABORATION:

- **Hybrid Intelligence**: Humans and AI will increasingly work together in hybrid intelligence system which helps them each other's strength. Human will provide context, institution, and ethical grounding, while AI will handle data analysis, pattern recognition, and automation.

- **Co-Creation of Knowledge:** Humans and AI will collaborate to create many new theories and solve many problems. AI can generate many theories as well as explore vast data, while humans can guide the many explorations, result, and refine the process in their own way and simpler process.

ADDRESSING CHALLENGES

- **Explainable Ai:** developing truly explainable AI system which remains a challenge. We need many advancements in algorithms and communication strategies to make AI decision making process transparent and understandable to human which should not be complex.
- **Bias and fairness:** AI system can inherit and amplify human business if not carefully designed and monitored continuous effort are needed to identify and mitigate business in data continuous effort are needed to identify and mitigate business in data algorithms, and application to ensure fairness and ethical use.

AI is predicted to grow increasingly pervasive as technology develops, revolutionizing including healthcare, banking, and transportation. The work market will change as a result of AI-driven automation, necessitating new position and skills. The future of AI is likely to be centered around human AI collaboration rather than a scenario where AI completely replaces human. Combining the strength of AI (data processing, pattern recognition) with a human creativity, empathy, and institutional institution can lead to powerful outcomes in various domains.

CONCLUSION

Bridging the gap between AI and human understanding is not a just destination but an ongoing journey with collaborative fuel by collaboration, innovation and a very commitment responsible development. While challenges like explain ability bias, and the evolution of human skill which is evolving through the ages, the potential rewards are vast like an error of intelligence empowered decision making, breakthrough across diverse field. Humans and AI are not just basically a competitor it is a partners. They both force their each other's strengths, we can see the cook relation between the AI and human understanding AI decision making is a crucial for building trust and ensuring ethical use of various data which helps them to enabling the effective human oversight. Continuous innovation in the AI techniques will help to clear the path for the upcoming future in the collaboration of AI.

By identifying the data, algorithm, application which is essential for ensuring the fairness and ethical Ness of the data which will directly help in responsible development and practices of the AI advancement. While AI do certain tasks, which is critical thinking creativity, social skills in humans to thrive in an evolving landscape. Both human and AI continuously learn and adapt in the dynamic environment of the computer science by sharing of knowledge and ongoing research on bridging the gap between AI and human understanding will increase the relationship between the computer science and human understanding.

In last bridging the gap between AI and human understanding is not just a technological challenge but also a human demanding vision in the ethical consideration for continuous learning. Bye giving first prioritization collaboration to the explain ability of AI will help them to navigate this journey for the purpose of the future where human and AI grow together for shaping the next generation, also which should be increased with an understanding between the AI and human will help them to grow the future with empathy and share progresses.

BIBLIOGRAPHY

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., & Devin, M. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems.

Araujo, T., Aresta, G., Castro, E., Rouco, J., Aguiar, P., Eloy, C., Campilho, A. (2017). Classification of breast cancer histology images using convolutional neural networks.

Artificial Intelligence Technology and Engineering Applications. (2017). ACES JOURNAL, 32, 5th ser., 381-386.

Bach, S., Binder, A., Montavon, G., Klauschen, F., Müller, K.-R., & Samek, W. (2015). On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation.

Baumgartner, C. F., Koch, L. M., Tezcan, K. C., Ang, J. X., & Konukoglu, E. (2017). *Visual feature attribution using wasserstein gans*. Paper presented at Proceedings of the IEEE computer society conference on computer vision and pattern recognition.

Bengio, Y. (2009). Learning deep architectures for AI. *Foundations and Trends in Machine Learning*, **2**, 1–127.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **35**, 1798–1828.

Berkes, P., & Wiskott, L. (2006). On the analysis and interpretation of inhomogeneous quadratic forms as receptive fields. *Neural Computation*, **18**, 1868–1895.

Biffi, C., Oktay, O., Tarroni, G., Bai, W., De Marvao, A., Doumou, G., … Rueckert, D. (2018). *Learning interpretable anatomical features through deep generative models: Application to cardiac remodeling*. Paper presented at International conference on medical image computing and computer-assisted intervention (pp. 464–471).

Bologna, G., & Hayashi, Y. (2017). Characterization of symbolic rules embedded in deep dimlp networks: A challenge to transparency of deep learning. *Journal of Artificial Intelligence and Soft Computing Research*, **7**, 265–286.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721-1730).

Doran, D., Schulz, S., & Besold, T. R. (2017). What does explainable AI really mean? A new conceptualization of perspectives.

Doshi-Velez, F., & Kim, B. (2017). Towards a rigorous science of interpretable machine learning.

Doshi-Velez, F., & Kim, B. (2019). Anchors: High-precision model-agnostic explanations. In Proceedings of the AAAI Conference on Artificial Intelligence, 33(01), 1527-1535.

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). *Visualizing higher-layer features of a deep network*. University of Monetreal Technical Report Nr. 1341

Esteva, A., Kuprel, B., Novoa, R. A., Ko, J., Swetter, S. M., Blau, H. M., & Thrun, S. (2017). Dermatologist-level classification of skin cancer with deep neural networks. *Nature,* **542**, 115–118.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, **349**, 273–278.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., … Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems (NIPS)* (pp. 2672–2680). Montreal, Canada: Neural Information Processing Systems Foundation.

Guillaume, S. (2001). Designing fuzzy inference systems from data: An interpretability-oriented review. *IEEE Transactions on Fuzzy Systems*, **9**, 426–443.

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). Metrics for explainable AI: Challenges and prospects.

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM*, **48**, 71–74.

Holzinger, A. (2014). Trends in interactive knowledge discovery for personalized medicine: Cognitive science meets machine learning. *IEEE Intelligent Informatics Bulletin*, **15**, 6–14.

Holzinger, A. (2016). Interactive machine learning for health informatics: When do we need the human-in-the-loop? *Brain Informatics,* **3**, 119–131.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science,* **349**, 255–260.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in neural information processing systems* (pp. 5574–5584). Long Beach, CA: Neural Information Processing Systems Foundation.

Komura, D., & Ishikawa, S. (2018). Machine learning methods for histopathological image analysis. *Computational and Structural Biotechnology Journal*, **16**, 34–42.

Lacave, C., & Diez, F. J. (2002). A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review*, **17**, 107–127.

Lakkaraju, H., Bach, S. H., & Leskovec, J. (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 (pp. 1675-1684).

Lakkaraju, H., Kamar, E., Caruana, R., & Leskovec, J. (2017). Interpretable and explorable approximations of black box models.

Lee, H., Grosse, R., Ranganath, R., & Ng, A. Y. (2009). *Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations*. Paper presented at 26th Annual international conference on machine learning (ICML '09) (pp. 609–616).

Lipton, Z. C. (2016). The mythos of model interpretability. In Proceedings of the 2016 ICML Workshop on Human Interpretability in Machine Learning (pp. 1-6).

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space.

Miller, T., Howe, P., & Sonenberg, L. (2017) Explainable AI: Beware of inmates running the asylum or: How i learnt to stop worrying and love the social and behavioural sciences.

Montavon, G., Samek, W., & Müller, K.-R. (2017). Methods for interpreting and understanding deep neural networks.

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In D. D. Lee, M.

Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.), *Advances in neural information processing systems 29 (NIPS 2016)* (pp. 3387–3395).

Barcelona, Spain: Neural Information Processing Systems Foundation.

Pawlowski, N., Brock, A., Lee, M. C., Rajchl, M., & Glocker, B. (2017). Implicit weight uncertainty in neural networks.

Poulin, B., Eisner, R., Szafron, D., Lu, P., Greiner, R., Wishart, D. S., … Anvik, J. (2006). Visual explanation of evidence with additive classifiers. In *National conference on artificial intelligence* (pp. 1822–1829). Cambridge, MA: MIT Press

Ribeiro, M. T., Singh, S., & Guestrin, C. (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 (pp. 1135-1144).

Ribeiro, M. T., Singh, S., & Guestrin, C. (2018). Anchors: High-precision model-agnostic explanations. ACM Transactions on Intelligent Systems and Technology (TIST), 9(3), 1-35.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature*, **323**, 533–536.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, **61**, 85–117.

Seeböck, P., Waldstein, S. M., Klimscha, S., Bogunovic, H., Schlegl, T., Gerendas, B. SLangs, G. (2018). Unsupervised identification of disease marker candidates in retinal oct imaging data.

Singh, D., Merdivan, E., Psychoula, I., Kropf, J., Hanke, S., Geist, M., & Holzinger, A. (2017). Human activity recognition using recurrent neural networks. In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine learning and knowledge extraction: Lecture notes in computer science LNCS 10410* (pp. 267–274).

Sundararajan, M., Taly, A., & Yan, Q. (2017). Axiomatic attribution for deep networks.

Zhou, B., Khosla, A., Lapedriza, A., Oliva, A., & Torralba, A. (2016). *Learning deep features for discriminative localization*. Paper presented at Proceedings of the IEEE conference on computer vision and pattern recognition (2921–2929).