EDITORIAL

Some germinal roots of AI and their impact on Computer Assisted Radiology and Surgery (CARS)

Heinz U. Lemke[1]

It may be of interest to the reader of this IJCARS Special Issue containing CARS 2019 papers with a focus on AI methods and tools, to have a brief glance at the historic context in which some of the developments toward intelligent machines have taken place.

Alan Turing

Going back to the roots of the now very popular and general term Artificial Intelligence (AI), the more neutral expression relating to Machine Intelligence (MI) was used as an umbrella term during its first years of the existence of this significant information technology. It is well-acknowledged that Alan Turing's historic paper in 1950 on "Computing Machinery and Intelligence" [1], outlining what is now called the Turing Test, was the starting point for the science and an increasing body of myths about the thinking machine, which Turing referred to in his paper.

At those times and even now, it was and is much easier to define in unambiguous terms what constitutes a machine (such as a Turing or von Neumann machine) than what is the essence of thinking in a verifiable manner. Turing took a short cut by means of a metaphor which he called the "imitation game," as a way of describing a situational model with several entities (a machine and some human protagonists) and actions (goal-driven intelligent interaction between the entities) which in today's terms is equivalent to emulate human thinking on a computer or a simulation of a virtual reality demanding complex decision making by either a human or a machine.

The situational model in the imitation game envisaged by Turing consists of two strictly separated rooms and three people: a man (A), a woman (B), and an interrogator (C) who may be of either sex. The interrogator stays in a room apart from the other two people. The ideal arrangement is to have a teleprinter communicating between the two rooms.

> The object of the game for the interrogator is to determine which of the other two is the man and which is the woman. It is A's object in the game to try and cause C to make the wrong identification. The object of the game for the third player (B) is to help the interrogator. What will happen when a machine takes the part of A in this game? Will the interrogator still decide incorrectly as many times if the role is performed by a machine?

These questions posed by Turing and their elaboration in the context of postulated "Contrary Views" and aspects of "Learning Machines" finally resulted in his observation in 1950:

> We may hope that machines will eventually compete with men in all purely intellectual fields… We can only see a short distance ahead, but we can see plenty there that needs to be done.

Even though Alan Turing predicted the possibility that thinking machines would come to be a reality in 50 years (apparently in 1952 he corrected this to 100 years), in the year 2019, we are still facing the same issues that he outlined in his paper, in particular, on digital computers, machine learning, and what constitutes human thinking and decision making in general.

Maurice Wilkes

An interesting statement came from one of Turing's contemporaries in Cambridge, Maurice Wilkes (head of the Cambridge University Mathematical Laboratory from 1945 to

✉ Heinz U. Lemke
   hulemke@cars-int.org

1   International Foundation for Computer Assisted Radiology and Surgery (IFCARS), Kussaberg, Germany

1980 and the second recipient of the Turing award in 1967), when he observed in 1953 that,

> If ever a machine is made to pass the (Turing) test, it will be hailed as one of the crowning achievements of technical progress, and rightly so.

The result of Maurice Wilkes's work over the period from 1945 to 1980 may best be described as pioneering contribution toward enabling technologies for the (intelligent) machines that Turing and von Neumann envisaged, in particular, stored program computers, microprogramming, macros, CAD modeling, capability-based computers, local area networks such as the Cambridge Ring, and mainframe-satellite computer connections (similar to server–client systems in today's terms). However, Maurice Wilkes was always very careful when it comes to making prognostic statements, perhaps exceptionally somewhat later in 1992, when he observed more cautiously that [2]:

> It is difficult to escape the conclusion that, in the 40 years that have elapsed since 1950, no tangible progress has been made toward realizing machine intelligence in the sense that Turing had envisaged. Perhaps the time has come to face the possibility that it never will be realized with a digital computer.

There are computer scientists such as John McCarthy and Marvin Minsky (the inventors of the term "Artificial Intelligence" in 1955 and Turing award winners of 1971 and 1969, respectively) and many others, who would thoroughly disagree with Maurice Wilkes's conclusion. Nevertheless, it is worth noting that there is not only Alan Turing's observation "plenty there that needs to be done," but also the question addressing the potential economic, social and ethical implications of Machine Intelligence.

Finally, regarding developing computer programs, Maurice Wilkes is associated with the observation [3]: "… It would be more logical first to choose a data structure appropriate to the problem, and then to look around for, or construct with a kit of tools provided, a language suitable for manipulating the structure." Something that translates very well into the modeling and simulation themes in the context of CARS in our times.

Joseph Weizenbaum

Even though the "Eliza" program [4] developed by Joseph Weizenbaum in 1966 (Professor of Computer Science at MIT from 1963 to 1988) was celebrated by some of its first users of this program as a breakthrough for Artificial Intelligence (for example, by implementing a powerful list processing tool based on an extension of LISP, allowing for the manipulation of graphs) and having past the Turing test, he himself, for many reasons, was rather skeptical of his own work as well as of some other AI pioneers (probably a reason why he never received the Turing award). In particular, he observed that while AI may eventually be possible, we should never allow computers to make important decisions, because computers will always lack human qualities such as compassion and wisdom.

For these and related views on the ultimate limits of computation, in the 1970s and well beyond, Joseph Weizenbaum found himself on the outside of the mainstream of AI [3]. "He criticized his colleagues for overselling AI and for not reaching their professed goals in a reasonable time span. Promises had been made by the profession that were not being fulfilled, and he had the temerity to tell the world of their shortcomings," an observation not too far away from the remarks made by Maurice Wilkes.

Other significant contributions toward machine intelligence

Following the work of the early AI pioneers, in a long series of workshops on Machine Intelligence and associated book volumes [5] extending over a period of about 35 years, many interesting mathematical methods and IT tools were conceived, in particular with reference to natural language processing (NLP) and cognition problems. The gradual translation of these methods and tools into health care applications started in the early 1970s with the AI in medicine pioneer Edward H. Shortliffe, who developed the clinical expert system MYCIN, one of the first rule-based artificial intelligence systems to enable machine-assisted medical decision making.

It is worth noting that many papers [5] included in the book series focused on modeling methods related to graph theoretic concepts and their applications in complex decision making. During this period, graphs were gradually attributed with some type of uncertainty quantifications by means of Bayesian conditional probabilities [6], resulting in a powerful method for complex situation modeling. Judea Pearl (Turing award winner 2011) [7] continued his drive to provide a comprehensive framework (Ladder of Causation) in which uncertainty is revealed to be a much more complex problem than had been hitherto thought. Of particular relevance in the context of this editorial is his analysis of Turing's binary classification of thinking and non-thinking entities, e.g., humans or machines, as compared to Pearl's three tier ladder of causation, consisting of the graduated and increasing intelligence levels of observing (association), doing (intervention), and imagining (counterfactuals).

Whether the abstraction based on three categories of intelligence levels in the ladder of causation is appropriate or not, 10 or more, or even only Turing's two levels, is an

interesting question, which may have to be addressed when designing intelligent systems relating to computer assisted radiology and surgery.

A brief summary of the last 70-year history of machine-assisted medical decision making and the role of computer modeling, with a plea to a formal uncertainty quantification (UQ) discipline (possibly derived from other domains such as nuclear security), is given by Begoli et al. [8]. Here, UQ appears to become a key methodological issue for the future advances in AI.

The interested reader who may want to learn about opinions where high-level machine intelligence may take us in the next 20–30 years, for example, what will be the role of the cognitive science and cognitive architectures, as well as relational reasoning and graph networks, is referred to references [9, 10], respectively.

Battaglia et al. [10] in 2018, reemphasized in their position paper on "Relational inductive biases, deep learning, and graph networks," the importance of structured representations of knowledge and computations, and in particular, systems that operate on graphs, much in line of what J. Pearl has been promoting. Their outline on a graph networks' (GN) framework gives an exhaustive up-to-date summary of the role of neural networks such as convolutional neural networks (CNNs), recurrent neural networks RNNs, multi-layer perceptrons (MLPs), message-passing neural networks MPNNs, non-local neural networks NLNNs, and others.

In summary, the authors of Ref. [10] claim that a vast gap between human and machine intelligence remains, especially with respect to efficient and generalizable learning. Their statement that "Graph networks are designed to promote building complex architectures using customizable graph-to-graph building blocks, and their relational inductive biases promote combinatorial generalization and improved sample efficiency over other standard machine learning building blocks" may well-show into the right direction for R&D in the field of CARS relating to complex clinical decision making supported by machine intelligence.

Some expert systems based on graph networks have already been developed in the specific context of assisting medical diagnostic and therapeutic procedures in radiology and surgery. Machine learning, deep learning (DL), and clinical decision support systems are typical examples of MI in sessions of past CARS congresses, see Fig. 1 [11, 12].

Within this specific medical focus, MI is providing new methodological, technical and clinical capabilities using advanced mathematical models and innovative information technology tools.

Examples of questions relating to MI that deserve attention

Even though a review of the ongoing research in the areas outlined above is beyond the scope of this Editorial, it may be appropriate to point out a few major research questions and possible directions the answers may provide.

This IJCARS Special Issue on CARS addresses, in principle, six critical (CARS) questions relating to the substance, relevance, applications, impact, and implications of mathematical methods and algorithms of MI in the domain of clinical applications:
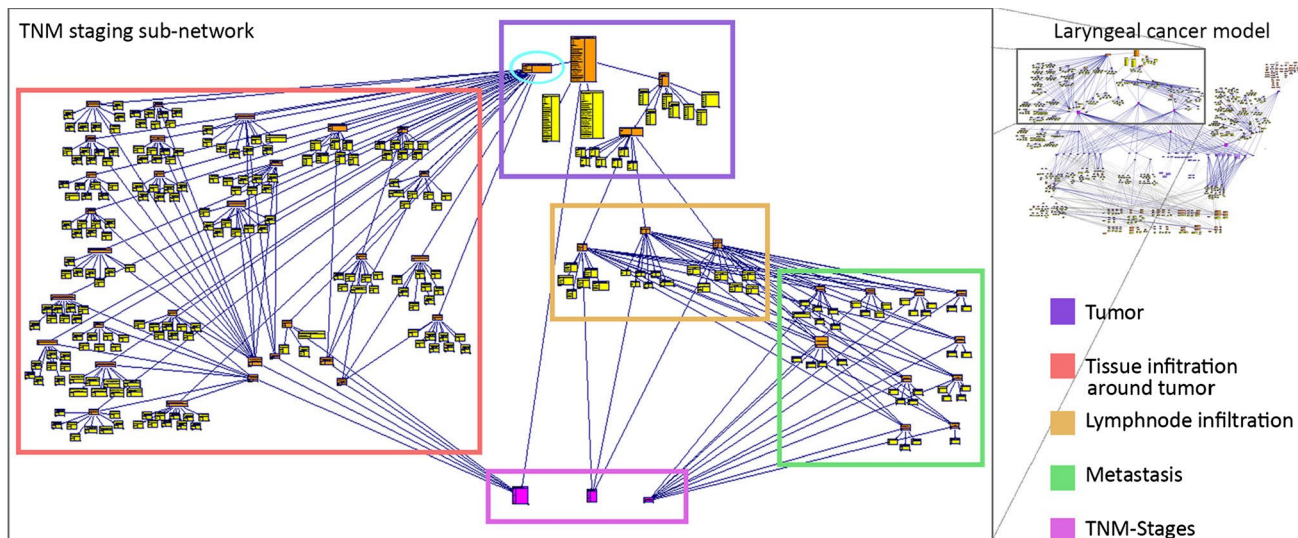
Fig. 1 Bayesian network model for treatment decision support of laryngeal cancer (above right). Variables (which can be more than 1000 for the given example) are manually arranged according to clinically related topics and highlighted by colored rectangles for illustrative purpose only. The enlarged part shows the TNM staging Bayesian sub-network

(1) What qualifies a mathematical method or an information technology tool to be considered as machine, artificial, or computational intelligence (or any other synonym or near-synonym) for radiology or surgery, e.g., from the field of image recognition, NLP, complex clinical decision making, treatment personalization and optimization, intelligent robotics, and instrumentation (sensors and actors)?

(2) Which mathematical methods or information technology tools are of particular relevance for applying MI in radiology and surgery, e.g., applicability of DL-structured neural networks, graphical models such as Bayesian networks, uncertainty quantification (UQ), support vector machines, genetic algorithms, generative adversarial networks (GANs)?

(3) How can these mathematical methods or information technology tools for MI be applied to improve clinical workflow and/or patient outcomes, e.g., role of human machine communication/interaction, supporting situational awareness, use of architectures, such as medical information and model management systems (MIMMS) with DL engines and utility-based and other intelligent software agents?

(4) When can results and impact of MI be expected for improved clinical workflow and patient outcomes, e.g., effective adoption of MI with incremental, substantial, or potentially transformational impact?

(5) What are the potential economic, decision theoretic, social, and ethical implications of MI, in radiology and surgery specifically, and in health care generally, e.g., reviewing of some of J. Weizenbaum's concerns in the context of CARS? What is the role of data driven or evidence-based decision making as compared to, or complemented by a model-based medical evidence in light of uncertainty, bias, intuition, confounding, and unknown variables?

(6) How will the long-term development of physician's cognition, decision making, actuating, and intuition capabilities be affected by synergistic and intelligent human–machine systems employed in radiology and surgery? Will physician's classic strength in multivariate thinking eventually be replaced by algorithmic thinking?

The potential answers to these questions are likely to be of a very divergent nature. With this IJCARS Special Issue on CARS 2019, an attempt is being made to address, in an exemplary manner, a few selected research topics in order to gain some insights into the realm of what can be considered to be MI in medicine.

The following provides a brief synopsis of six papers, three each from radiology and surgery, on applying intelligent methods and tools in the light of some of the questions asked as outlined above.

Examples of MI in radiology

A deep learning framework for efficient analysis of breast volume and fibroglandular tissue using MR data with strong artifacts

T. Ivanovska, T. G. Jentschke, A. Daboul, K. Hegenscheid, H. Völzke, F. Wörgötter

Georg-August-University Göttingen; University Medicine Greifswald; Unfallkrankenhaus Berlin, Germany The main purpose of the work presented in this paper is to develop, apply, and evaluate an efficient approach for breast density estimation in magnetic resonance imaging (MRI) data with strong intensity inhomogeneities. To evaluate the breast density, i.e., to measure the complete breast volume and the parenchyma, it usually requires to investigate the risk in a certain population group. This group is usually represented by women without strong pathological findings, i.e., no tumors are present in the data.

The given framework consists of five steps: correction of artifacts, data augmentation, breast volume segmentation, breast volume masking with nipple removal, and fibroglandular tissue segmentation. Specifically for breast volume segmentation, nipple extraction, and fibroglandular tissue segmentation steps, a well-known deep learning architecture has been employed. Following the well-known N4ITK Bias Correction Algorithm to remove the intensity inhomogeneity from the breast datasets, the segmentation steps utilize a two class 2D U-Net deep learning architecture.

The presented method reaches an average Dice similarity coefficient (DSC) of 0.925. Partially, the improvement in the DSC for parenchymal tissue segmentation, as compared to previous results by the same group with a $DSC = 0.83$ and other classical state-of-the-art approaches, is due to the more accurate total breast volume segmentation.

It must be emphasized, however, that the described framework is not a CAD system for tumor and lesion detection. Nevertheless, the proposed solution has the potential to improve the clinical workflow for screening purposes for breast cancer from results obtained after applying the method to big epidemiological data possibly with thousands of participants.

Computer-aided diagnosis of gastrointestinal stromal tumors: a radiomics method on endoscopic ultrasound image

X. Li, F. Jiang, Y. Guo, Z. Jin, Y. Wang

Fudan University, Changhai Hospital, China A manifold of algorithms for radiomics-based CAD classification systems have been proposed and applied for the diagnosis of tumors in various organs such as the lung, breast, thyroid and brain. Very little attention has so far been given to identify gastrointestinal stromal tumors (GISTs) on gastro endoscopic ultrasound (G-EUS) images.

The main purpose of the work presented in this paper is to automatically extract quantitative features from G-EUS images and to develop, apply, and evaluate a radiomics-based CAD classification system to improve the preoperative diagnostic accuracy of the rare higher risk group (HRG) from those of the lower risk group (LRG).

Radiomics-based risk assessment requires mechanisms for data sharing and requires the availability of data across many patient and tumor types. In the present study, this has been achieved by collecting G-EUS images from 19 hospitals of four different risk level GISTs. The dataset included 168 case HRG GISTs and 747 case LGR GISTs.

Prostate cancer detection using residual networks

H. Xu, J. S. H. Baxter, O. Akin, D. Cantor-Rivera

Ezra AI, Toronto, Canada, University of Rennes, France, Memorial Sloan Kettering Cancer Center, New York, NY, USA One of the specific aims of the ACR (American College of Radiologists) PI-RADS™ v2 (Prostate Imaging-Reporting and Data System) is to enhance interdisciplinary communications of radiologists with referring clinicians such as urologists, pathologists, and others. Taking into account that the prostate has a complex 3D anatomy with respect to the distribution of fibromuscular stroma and glandular tissue, the segmentation model used in PI-RADS™ v2 employs thirty-nine sectors/regions: thirty-six for the prostate, two for the seminal vesicles, and one for the external urethral sphincter. In PI-RADS™ v2, it is postulated that "Computer-aided evaluation (CAE) technology may improve workflow (display, analysis, interpretation, reporting, and communication), provide quantitative pharmacodynamic data, and enhance lesion detection and discrimination performance for some radiologists, especially those with less experience interpreting mp-MRI exams."

With reference to the CARS question #2 above, the paper by Helen Xu et al. addresses possible mathematical methods and information technology tools which are of particular relevance for applying AI algorithms in radiology,

here specifically to identify suspicious lesions on prostate mp-MRI, but initially on a subset of the thirty-six sectors/regions for the prostate. It appears that residual neural networks (ResNets) comprise a class of particularly well-suited modeling methods which can be considered to train deeper neural networks easier and faster than other architectures such as adversarial networks, end-to-end deep neural networks, or multimodal convolutional neural networks.

For the training of their ResNet, three radiologists evaluated axial T2-weighted (T2 W), apparent diffusion coefficient (ADC) map, and high b-value (BVAL) diffusion-weighted images, and segmented lesions that were PI-RADS v2 assessment category 3 or greater, with category 1 being most likely to be benign and 5 being highly suspicious of malignancy. Dynamic Contrast-Enhanced (DCE) MRIs, however, were not included in the training set. As stated in the ACR PI-RADS™, "most published data show that the added value of DCE over and above the combination of T2W and DWI is modest."

The segmentations generated from mp-MR images of 346 subjects by the most expert radiologists were used as ground truth for network training and validation. Segmentations produced by two other radiologists were used to establish a baseline comparison of the network performance. After a successful training process for the network, the receiver operating curve (ROC) analysis demonstrated an area under the curve (AUC) of 97% for the ResNet detected lesions.

An interesting observation made in this paper was with respect to a quantitative analysis, for example comparing lesions outlined by the network against three radiologists, which showed a higher agreement with the segmentations by the most junior radiologist!

With reference to the CARS question #2 and #3 above, the paper by Helen Xu et al. addresses possible mathematical methods and information technology tools which are of particular relevance for applying AI algorithms in radiology.

Examples of MI in Surgery

Tissue classification of oncologic esophageal resectates based on hyperspectral data

M. Maktabi, H. Köhler, M. Ivanova, B. Jansen-Winkeln, J. Takoh, S. Niebisch, S. M. Rabe, T. Neumuth, I. Gockel, C. Chalopin

Innovation Center Computer Assisted Surgery (ICCAS); University Hospital Leipzig, Germany This preliminary study represents a promising application of both, machine learning and spectral analysis in the field of tissue classification, here specifically to differentiate malignant from healthy

tissue based on hyperspectral image (HSI) recordings of esophagus and stomach resectates.

The HSI-camera provides hypercubes with a high spectral resolution of 5 nm in the visible and near infrared range from 500 to 1000 nm, including 100 spectral values. The spatial resolution of the images is 640×480 pixels (x-, y-axes) with a spatial resolution of 0.1 mm/pixel. For the classification of the spectra, four different standard classification approaches, i.e., k-nearest neighbors (k-NN), Random Forest (RF), Support Vector Machines (SVM), and Multilayer Perceptron classifier (MLP), were used.

The HSI dataset was relatively small and divided into training validation and test sets, with annotated HSI data of nine patients and the spectra of two patients for the training validation and test set, respectively.

Even though the detection performance for the experiments reported is not so high and considering that the HSI modality is still relatively new, the study shows promising results for the future use of HSI for detection of esophagus cancers, to visualize the tumor margins of resected tissue and eventually the removal of in vivo cancer tissue.

Automatic annotation of surgical activities using virtual reality environments

A. Huaulmé, F. Despinoy, S. A. H. Perez, K. Harada, M. Mitsuishi, P. Jannin

University of Rennes, France; University of Tokyo, Japan To automate the recognition of essential parts of the surgical workflow and their evolution over time in the specific context of surgical task, phase, gesture, or surgical activity recognition in general, is an important functionality for the design of human–machine collaborative systems in the operating room (OR).

With reference to the CARS question #2 above and in particular #3 referring to improve clinical workflow, the paper by Arnaud Huaulmé et al. addresses possible information technology tools which are of particular relevance for applying intelligent software in surgery, here specifically for surgical workflow analysis in the OR. It presents the work carried out by research groups in Tokyo and Rennes during the past few years on the developed and application of machine learning methods in order to achieve intelligent assistance for automatic annotation for surgical process models (SPMs). The aim is to reduce dependence on human intervention in the annotation process.

Starting with information derived from virtual–reality environments provided by surgical task simulators, some rules for transcription and additional contextual information, the proposed system ASURA (Automatic SimUlatoR Annotator) interprets this information in order to provide annotated surgical activities, steps, and finally surgical phases.

ASURA is extensively validated in the context of a peg-transfer task performed on a VR simulator by providing validation metrics relating to time considerations between manual and automatic annotations as well as intra- and inter-observer variability concerning timing and accuracy of manual annotations for surgical process modeling.

It appears that the proposed ASURA system may be a well-suited architecture for achieving automatic surgical process modeling, when applied to capture the dynamics of surgical activities within a reasonable level of complexity. How to predict and differentiate between user intentions in surgical workflows, however, remains to be an interesting question to be addressed in the future. This applies also to situation or context awareness of SPMs.

Toward versatile cooperative surgical robotics: a review and future challenges

P. Schleer, S. Drobinsky, M. de la Fuente, K. Radermacher

Helmholtz Institute for Biomedical Engineering, Aachen, Germany This group at the RWTH in Aachen has been instrumental in the design of dynamic networks for medical devices and IT systems in the OR and corresponding standards, which may also be applied to cooperative robotic system. The aim of the paper is to provide a review on various surgical disciplines supporting different surgical task sequences and differing ways of human–machine cooperation or degrees of automation. This is followed by an overview of cooperative robots in surgery.

Human–machine interaction is being analyzed for different classes of synergistic robotic systems, specifically handheld, hands-on, and tele-manipulated devices. Essential functional characteristics are described in order to identify generic cooperative robotic device profiles (CRDP), features and use cases which are summarized in a classification scheme. Distinct CRPDs are needed to enhance versatility, improve benefit-to-cost ratio and, thereby, market spread of surgical robotics.

In combination with an open communication standard for the operating theater, a very critical part in this scenario is the possibility of arbitration (mentioned 5 times!) between human and machine. With reference to the six critical CARS questions as outlined at the beginning of this editorial, specifically question 3 and 6, arbitration characteristics will be an essential part of intelligent human–machine systems not only for applications in surgery but eventually also in radiology.

Concluding remarks

In summary, the above CARS papers do not necessarily give a representative view of what Machine Intelligence is all about, but they indicate the diversity of applications of MI

related methods and tools in radiology and surgery. Papers presented at CARS generally, and the observations made regarding the roots of AI allow a tentative definition of MI as related to CARS [13].

Something to be assigned "intelligent," in the context of this Editorial implies "a system which has an adequate representation of the present situation (situational model) and an executable plan (process model) to proceed from the present situation to the best possible next situation."

To proceed to the best possible next situation, the system needs to have a model of the desired future situation and a model of the workflow, bringing into this definition also the issue of cause and effect. In any case, the modeling aspect (which implies the capability of cognition) and the algorithmic component to move from one situation to the next situation (which may imply complex decision making), could be considered to be the core components of intelligence, whether for human, machine, or animal intelligence.

The "machine learning" aspect is of course part of it, and if this is accepted to be of importance, we do not need to define "Artificial Intelligence," but "Machine Intelligence" instead. When a machine is capable of learning, the result is machine intelligence and not an artificial intelligence!

Compliance with ethical standards

References

1. Turing AM (1950) Computing machinery and intelligence: a quarterly review of psychology and philosophy. Mind 59(236):433–460. https://doi.org/10.1093/mind/LIX.236.433

2. Wilkes M (1992) Artificial intelligence as the year 2000 approaches. Commun Assoc Comput Mach 35(8):17–20

3. Lee JAN (1995) Computer pioneers. IEEE Computer Society Press, Los Alamitos

4. Weizenbaum J (1966) ELIZA: a computer program for the study of natural language communication between man and machine. Commun Assoc Comput Mach 9:36–45

5. Collins NL, Michie D (1971) Machine intelligence, vol 1 + 2. Edinburgh University Press, Edinburgh

6. Pearl J (1986) Fusion, propagation, and structuring in belief networks. Artif Intell 29(3):241–288

7. Pearl J, MacKenzie D (2018) The book of why, the new science of cause and effect. Basic Books, New York

8. Begoli E, Bhattacharya T, Kusnezov D (2019) The need for uncertainty quantification in machine-assisted medical decision making. Nat Mach Intell 1:20–23

9. Müller VC, Bostrom N (2014) Future progress in artificial intelligence: a survey of expert opinion. In: Müller VC (ed) Fundamental issues of artificial intelligence. Springer, Berlin

10. Battaglia PW, Hamrick JB, Bapst V, Sanchez-Gonzalez VA, Zambaldi V, Malinowski M, et al (2018) Relational inductive biases, deep learning, and graph networks. arXiv:1806.01261v3 [cs.LG]

11. Cypko MA (2017) Therapy decision support system using Bayesian networks for multidisciplinary treatment decisions. Ph.D. Thesis, Fakultät für Mathematik und Informatik der Universität Leipzig

12. Cypko MA, Stoehr M, Kozniewski M, Druzdzel MJ, Dietz A, Berliner L, Lemke HU (2017) Validation workflow for a clinical Bayesian network model in multidisciplinary decision making in head and neck oncology treatment. Int J CARS 12:1959–1970

13. Lemke HU (2013) Machine intelligence in the OR: optimizing the effectiveness of surgeons with IT tools. International Hospital, Guwahati