Chapter 1
Introducing AI in Medicine

Trevor A. Cohen, Vimla L. Patel, and Edward H. Shortliffe

After reading this chapter, you should know the answers to these questions:

- How does one define artificial intelligence (AI)? What are some ways in which AI has been applied to the practice of medicine and to health care more broadly?
- How does one define cognitive informatics (CI)? How can the CI perspective inform the development, evaluation and implementation of AI-based tools to support clinical decision making?
- What are some factors that have driven the current wave of interest in AI methods?
- How can one compare and contrast knowledge-based systems with machine learning models? What are some of the relative advantages and disadvantages of these approaches?
- Considering the current state of progress, where is research and development most urgently needed in the field and why?

T. A. Cohen (✉)
University of Washington, Seattle, WA, USA
e-mail: cohenta@uw.edu

V. L. Patel
New York Academy of Medicine, New York, NY, USA

E. H. Shortliffe
Columbia University, New York, NY, USA

© The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

T. A. Cohen et al. (eds.), *Intelligent Systems in Medicine and Health*, Cognitive Informatics in Biomedicine and Healthcare,
https://doi.org/10.1007/978-3-031-09108-7_1

The Rise of AIM

Knowledge-Based Systems

The term “artificial intelligence” (AI) can first be found in a proposal for a conference that took place at Dartmouth College in 1956, which was written by John McCarthy and his colleagues [1]. The research to be conducted in this two-month conference was built upon the “conjecture that every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it.” This conference is considered a seminal event in AI, and was followed by a steady growth of interest in the field that is reflected by the frequency with which the term ‘artificial intelligence’ appeared in books of this era (Fig. 1.1). There was a first peak of activity in the mid-1980s that followed a period of rapid progress in the development of knowledge-based **expert systems**, systems that were developed by eliciting knowledge from human experts and rendering this content in computer-interpretable form. Diagnostic reasoning in medicine was one of the first focus areas for the development of such systems, providing proof that AI methods could approach human performance in tasks demanding a command of a rich base of knowledge [3]. This shows that medical decision making has long been considered a paradigmatic example of intelligent human behavior, and has been a focus of—and has had an influence on—AI research for decades.

The historical trend in term usage in Fig. 1.1 also reveals a dip in enthusiasm and in support for AI endeavors following the peak in the 1980s (one of the so-called ‘AI Winters’), for reasons that are discussed in Chap. 2. For the purpose of this introduction, we focus on the events of recent years, which have seen rapid growth in interest in AIM applications driven by media attention to AI in general (evident to the right of Fig. 1.1), coupled with high profile medical demonstrations of diagnostic

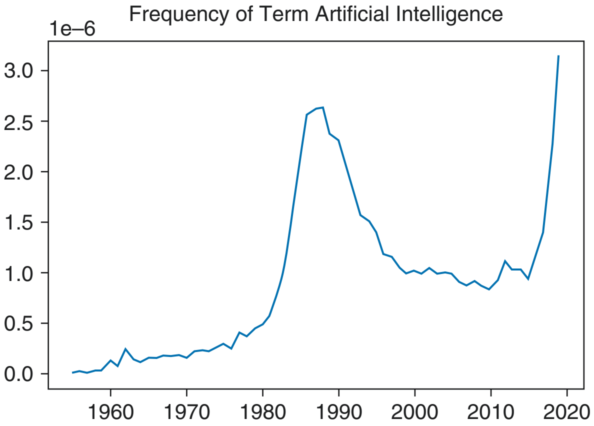

Fig. 1.1 Frequency with which the term ‘artificial intelligence’ appears in books published between 1950 and 2019 and digitized by Google (data obtained from the Google Books n-gram viewer website [2]). 1e-6 indicates the order of frequency of occurrence of the term (e.g. approximately 2.5 occurrences per million bigrams at the peak in the late eighties)

accuracy, particularly in image recognition. This growth is part of a larger picture in which the capabilities of **artificial neural networks**—originally conceived as models of human information processing and learning [4, 5]—have been enhanced through a convergence of the availability of large data sets for training, refinements in training approaches, and increases in computational power.

Neural Networks and Deep Learning

Loosely inspired by the interconnections between neurons in the human brain, artificial neural networks consist of interconnected functional units named neurons, each producing an output signal determined by their input data, weights assigned to incoming connections, and an **activation function** that transforms cumulative incoming signals into an output that is passed on to a next layer of the network. The weights of a neural network serve as parameters that can be altered during training of a model, so that the output of the neural network better approximates a desired result, such as assigning a high probability to the correct diagnostic label for a radiological image. When used in this way, neural networks exemplify the paradigm of **supervised machine learning**, in which models learn from labels (such as diagnoses) assigned to training data. This approach is very different in nature from the deliberate engineering of human knowledge that supported the expert systems in the first wave of AIM (see Chap. 2 and, for detailed accounts of knowledge modeling and machine learning methods, see Chaps. 4 and 6 respectively).

While machine learning models can learn to make impressively accurate predictions, especially when large data sets are available for training, systems leveraging explicitly modeled human knowledge—systems intended to reason *as humans do*—are much better positioned to explain themselves (for an example, see Box 1.1) than systems that have been developed to optimize accuracy without considering human cognition. Explanation has long been recognized as a desirable property of AI systems for automated diagnosis, and as a prerequisite for their acceptance by clinicians [6] (and see Chap. 8). However, the general trend in machine learning has been that accuracy comes at the cost of interpretability, to the point at which restoring some semblance of interpretability to the predictions made by contemporary machine learning models has emerged as a field of research in its own right—explainable AI—with support from the Defense Advanced Research Projects Agency (DARPA),[1] the same agency that initiated the research program on network protocols that ultimately led to a consumer-accessible internet.

[1] See https://www.darpa.mil/program/explainable-artificial-intelligence (accessed August 18, 2022) for details.

Box 1.1 An explanation provided by the MYCIN system in response to a user entering "WHY": From Shortliffe et al. 1974 [7]

- **WHY
- [1.0] It is important to find out whether there is therapeutically significant disease associated with this occurrence of ORGANISM-1.
- It has already been established that:
- [1.1] the site of the culture is not one of those which are normally sterile, and
- [1.2] the method of collection is sterile
- Therefore, if:
- [1.3] the organism has been observed in significant numbers
- Then: there is strongly suggestive evidence (.9) that there is therapeutically significant disease associated with this occurrence of the organism
- [Also: there is strongly suggestive evidence (.8) that the organism is not a contaminant]

This trend toward accurate but opaque predictions has accelerated with the advent of **deep learning** models—neural networks that have multiple intervening layers of neurons between input data and output predictions. While deep neural network architectures are not new phenomena (see for example the important paper by Hinton et al. [8]), their performance when trained on large data sets has produced dramatic improvements in results attained across fundamental tasks such as speech recognition, question answering and image recognition**.**

Figure 1.2 shows the extent of recent improvements for three key benchmarks: the Stanford Question Answering Dataset (SQUAD [9])—over 100,000 comprehension questions related to short articles; ImageNet—over 14 million images each assigned one of two hundred possible class labels [10]; and LibriSpeech—over 1000 hours of speech with matching text from audiobooks [11]. Of note, with both SQUAD and ImageNet, human performance on the tasks concerned has been estimated, and superseded by deep learning models.

Conceptually, the advantages of deep learning models over previous machine learning approaches have been attributed to their capacity for **representation learning** [12]. With prior machine learning approaches, performance generally advanced through engineering ways to represent incoming data (such as pixels of an image representing a handwritten digit) that led to better downstream machine learning performance (representations such as a count of the number of loops in a handwritten digit [13]). With deep learning models, the lower layers of a network can learn to represent incoming data in ways that facilitate task performance automatically.[2] Of particular importance for domains such as medicine, where large labeled data

[2] While deep learning models excel at learning representations that lead to better predictive modeling performance, representation learning is broader than deep learning and includes a number of previously established methods. For a review of developments up to 2013, see [14].

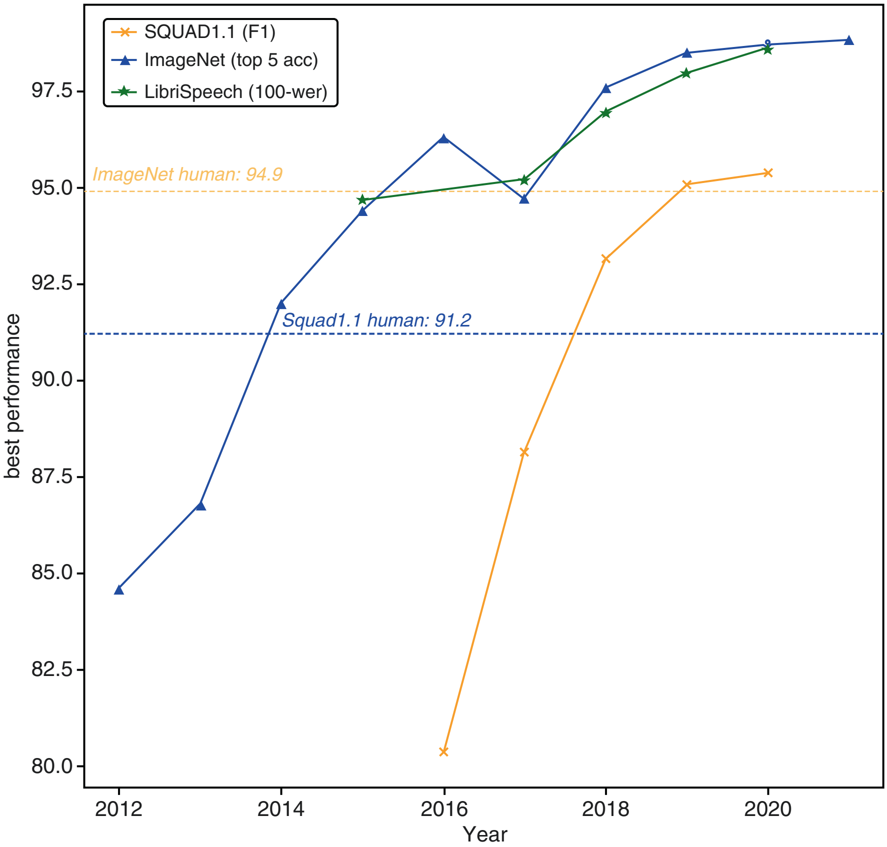

Fig. 1.2 Best documented performance, by year, on three key benchmarks (data from the 2021 AI Index Report [15, 16]). (1) SQUAD1.1 = Stanford Question Answering Dataset (version 1.1). Performance metric is "F1" (the balanced f-measure; see Chap. 6); (2) ImageNet - performance metric is "top 5 acc" (the percent of images in which the correct label, among 200 possibilities, appeared in the top 5 predictions); (3) LibriSpeech - performance metrics is "100-wer" (a transformation of the word error rate, with 100 indicating every word in a recording was recognized correctly). Dashed lines indicate documented human performance on the task concerned, which has been superseded by AI in both cases

sets are relatively difficult to obtain, the ability to extract useful representations for one task can often be learned from training on another related one. This ability to apply information learned from one task or data set to another is known as **transfer learning**, and is perhaps best exemplified by what has become a standard approach to classifying medical images (see Chap. 12): adding a classification layer to a deep neural network that has been pretrained on the task of recognizing non-medical images in ImageNet [17]. Similarly, fine-tuning of models such as Google's **BERT** and Open-AI's GPT series, which were originally trained to predict held-out words in large amounts of text from a range of digital data sources, has advanced performance across a broad range of **natural language processing (NLP)** tasks [18, 19].

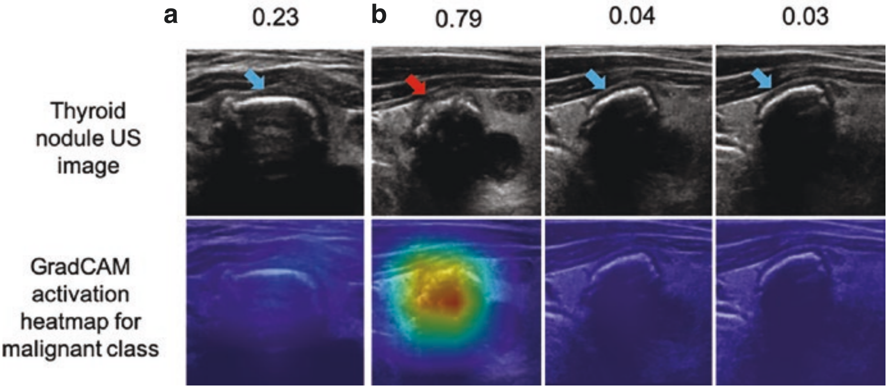

Fig. 1.3 Recognition of a subtle diagnostic cue by a deep neural network trained to detect thyroid cancer in different ultrasound images of the same nodule. Each image (top row) is annotated with the probability of malignancy according to the model, and is paired with a visualization of the pixels attended to by the deep learning model when making a prediction for whether an image is in the "malignant class", developed using the GradCam method [20]. Only the second image from the left exhibits the diagnostic feature of interrupted eggshell calcification, in which the rim of the opaque "shell" of calcification (blue arrows in the top row) is disrupted (red arrow). The GradCam visualization reveals the model has learned to attend to this subtle diagnostic feature. Image courtesy of Dr. Nikita Pozdeyev

Machine Learning and Medical Practice

Of course, outperforming humans on the repetitive and mundane task of selecting among hundreds of possible labels for a given image, or surpassing their accuracy in answering multiple choice questions about particular passages, does not necessarily provide an indication that deep neural networks could meet the requirements for flexibility, prioritization and adaptive decision making under uncertainty needed to *replace* medical practitioners in a busy clinical environment (audiobooks are also far less challenging to transcribe than recordings captured in a naturalistic environment—see Chap. 9 for a related discussion of automated medical transcription).

Nonetheless, the ability to recognize diagnostically important features is a fundamental task in interpreting medical images (as illustrated in Fig. 1.3—see also Chap. 12). A system capable of answering clinical questions accurately on the basis of written notes would make the information that these notes contain amenable to downstream computational processing for decision support or observational research (methods to achieve such ends are discussed in detail in Chap. 7). Furthermore, similar advances in performance have been achieved by predictive models in medicine, due in part to the large volume of digitized medical data that has accompanied the adoption of **electronic health record** (EHR) systems,[3] and the widespread use of digital platforms for image storage and retrieval (see Chap. 3) [22].

[3] In the United States this increase in adoption is attributable to the incentivization structures provided by the Health Information Technology for Economic and Clinical Health (HITECH) act of 2009 [21].

For example, a 2016 paper in the *Journal of the American Medical Association* describes an impressively accurate deep learning system for the diagnosis of diabetes-related eye disease in images of the retina [23]. Similarly, a widely-cited 2017 paper in *Nature* describes the application of deep learning to detect skin cancer [24], with the resulting system performing as well as 21 board-certified dermatologists in identifying two types of neoplastic skin lesions. These systems leveraged recent advances in AI, including deep neural network architectures and approaches to train them efficiently, as well as large sets of labeled data that were used to train the networks—*over 125,000 images* in each study. The dermatology system benefitted further from pre-training on over *1.25 million non-medical images* labeled with 1000 object categories. Beyond imaging, deep learning models trained on EHR data have learned to predict in-hospital mortality, unplanned readmission, prolonged length of stay, and final discharge diagnosis—in many cases outperforming traditional predictive models that are still widely used in clinical practice [25]. In this case, models were trained on data from over 200,000 hospitalized adult patients from two academic medical centers, considering *over 40 billion sequential data points* collectively.

These advances have attracted a great deal of press attention, with frequent articles in prominent media outlets considering the potential of AI to enhance—or disrupt—the practice of medicine [26–28]. As we have discussed in the preface to this volume, neither AI systems with physician-level performance nor media attention to such systems are without precedent, even in the days before advances in computational power and methodology mediated the current explosive interest in machine learning. However, the convergence of an unprecedented availability of clinical data with the maturation of machine learning models (and the computational resources to train them at scale) has allowed the rapid development of AI-based predictive models in medicine. Many demonstrate impressive results beyond those we have briefly described here. Furthermore, the proven commercial viability and public acceptance of such models in other areas have offset some of the skepticism with which AI models were greeted initially. Having seen the effectiveness with which machine learning models leverage data to deliver our entertainment and shopping recommendations on a daily basis, why would we not wish such systems to assist our clinicians in their medical practice? A strong indicator of the commercial potential of AI-based systems in medicine is the emergence of regulatory frameworks for their application in practice (see also Chap. 18) [29], with a number of AI systems already approved for medical use in the United States (Fig. 1.4) and Europe [30].

The Scope of AIM

A fundamental question in the study (and regulation) of AIM systems concerns the definition of the term “Artificial Intelligence”. Given the breadth of approaches that have been categorized as related to AI, it is perhaps not surprising that there is no universally-accepted definition of this term, and that the extent to which contemporary deep learning approaches constitute AI is still vigorously debated [32, 33]. A representative sample of AI definitions is provided in Box 1.2. While there are

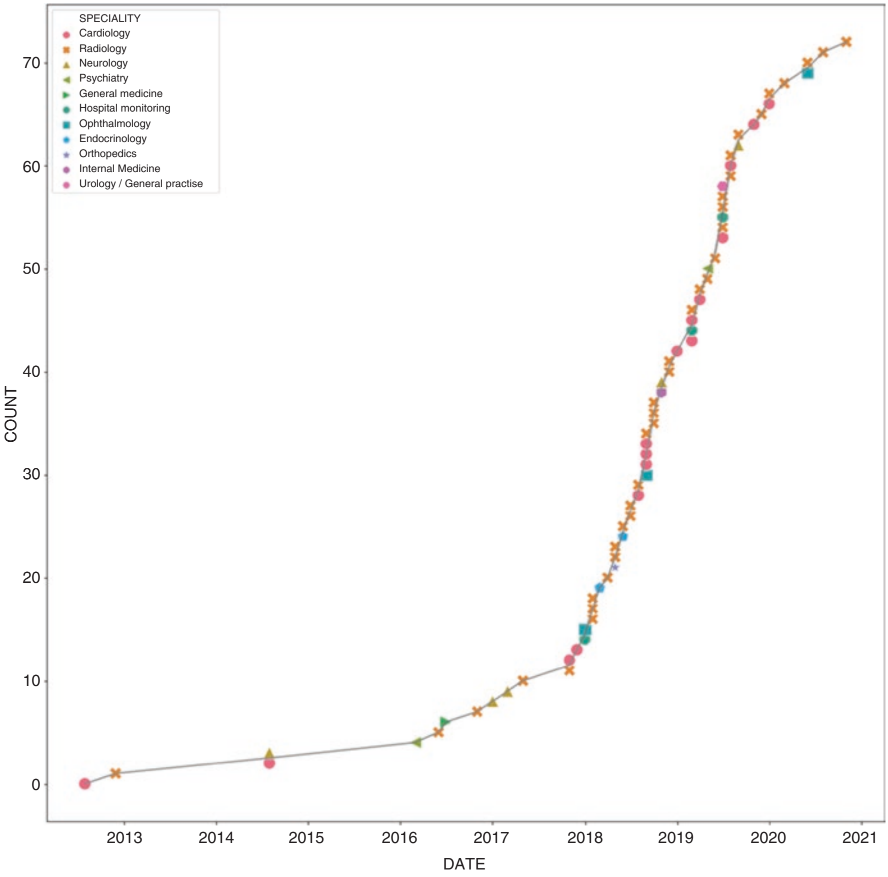

Fig. 1.4 FDA approvals for AI-related products by specialty (data drawn from medicalfuturist.com [30, 31]) with radiology systems (X) the most common category

Box 1.2 Sample Definitions of Artificial Intelligence

- *"The study of complex information processing problems that often have their roots in some aspect of biological information processing" (Marr, 1977)* [34]
- *"…the study of ideas that enable computers to do the things that make human beings seem intelligent: the ability to reason symbolically, the ability to acquire and apply knowledge, and the ability to manipulate and communicate ideas" (Winston, 1977)* [35]
- *"….the part of computer science concerned with designing intelligent computer systems, that is, systems that exhibit the characteristics we associate with intelligence in human behavior – understanding, language, learning, reasoning, solving problems and so on" (Barr* et al.*, vol 1, 1981, p. 3)* [36]

- *"The branch of computer science that is concerned with the automation of intelligent behavior" (Luger and Stubblefield, 1993)* [37]
- *"It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable". (McCarthy, 2007)* [38]

clearly common threads that run among them, notably the emphasis on intelligence (loosely defined by Barr as exhibiting the characteristics we associate with intelligence in human behavior, or by Winston as emphasizing the use of knowledge and an ability to communicate ideas), the definitions also reflect a departure from the cognitive motivations of AI at its inception—performance of tasks as humans do—to the more pragmatic motivations of the performance-oriented systems that are commonly termed AI today. Note that McCarthy in particular asserts explicitly that biological constraints need not apply. Of course, motivations for understanding how machines might solve a problem presumed to require human intelligence are not exclusively pragmatic, as this topic is also of considerable academic interest.

As one might anticipate given the fluidity of definitions of AI in general, the notion of what qualifies as AI in medicine is also in flux. At the inception of the field, the focus was on systems that could reason, leveraging encoded knowledge (including probabilistic estimates or uncertainty) derived from clinical experts. Such formalization of knowledge to render it computable also underlies the clinical decision support rules embedded in contemporary EHR systems. However, few would argue that the individual rules firing alerts in such systems constitute AI, even when considered collectively (see the discussion of warnings and alerts in Chap. 17). It seems, therefore, that the perceived difficulty of the tasks accomplished by a system determine whether it is thought to have exhibited intelligent behavior. Today, machine learning approaches (including deep neural networks) are strongly associated with the term AI. These systems are not designed to reason, but instead learn to recognize patterns, such as diagnostic features of radiology images, leading to performance on constrained tasks that is comparable to that of highly trained physicians. As such it is easy to argue that they exhibit intelligent human behavior, at least in the context of a task for which large amounts of labeled training data are readily available. Furthermore, such models can make predictions that are beyond the capabilities of human experts at times, such as prediction of cardiovascular risk factor status from retinal fundus photographs [39], or prediction of 3-D protein structure from an amino acid sequence [40]. Perhaps as a consequence of the lack of funding for research associated with the term AI during periods in which it was out of favor (see Chap. 2), a great deal of machine learning work in the field was not framed as AI research, but would be perceived this way in retrospect. Analogous to the case with rule-based models, this raises the question of how sophisticated a machine learning model is required to qualify as AI. For example, would a system based on a logistic regression model trained on a handful of features, with less than ten trainable parameters constitute AI? Perhaps, as with rules, the main question concerns the nature of the task that the model is able to accomplish, with a benchmark for AIM being the automated accomplishment of tasks that would be challenging for a highly trained human.

From Accurate Predictions to Clinically Useful AIM

However, irrespective of whether the engineers of AIM systems attempt to emulate human-like problem-solving processes, the ultimate goal of such efforts is often to support decision making by human clinicians at the point of care. The role of AIM in improving the quality, efficiency and safety of clinical practice exists within a larger system that includes human decision makers [41]. As such, both the remarkable capabilities and recognized constraints of human information processing must also be considered when designing and deploying AI-based systems, even if the systems concerned do not explicitly attempt to emulate human information processing methods. The consideration of the broader context in which AI-based systems must operate to influence patient care reveals a number of challenges that must be overcome in order to bridge the gulf between systems that perform well in the context of a constrained reference set, and systems that provide clinical utility at the point of care. Many of these challenges have been recognized since the inception of the field. In a 1975 paper, Shortliffe and Davis identified a series of seven considerations for expert system evaluation that suggest a path from conception of a system to clinical utility (Table 1.1; see also Chap. 17).

Of note, most of the work on accurate automated medical image interpretation we have discussed addresses only the second consideration in Table 1.1, and improving the ability of machine learning models to approach (or even surpass) the accuracy of expert clinicians has remained the focus of much recent work [43]. However, such models must be embedded in systems that are both usable and acceptable to clinicians if they are to exert an effect on management to improve outcomes for patients or to advance institutional or societal priorities such as cost-effectiveness. Furthermore, the design of AI systems should be motivated by the needs of clinicians, which are best understood in the context of the processes and environmental constraints in which they work [41].

Table 1.1 Overview: seven considerations for system evaluation [42]

	Demonstration		Impact
1	Need	5	Management
2	Expert-level performance	6	Patient outcome
3	Usability	7	Cost-effectiveness
4	Acceptance by clinicians		

The Cognitive Informatics Perspective

Why CI?

It is our view that the discipline of *cognitive informatics* (CI) [44–46], which brings the perspective of the cognitive sciences to the study of medical decision making by human beings and machines, is uniquely positioned to address these challenges. Through its roots in the study of medical reasoning [47–49], CI provides a sound scientific basis from which to consider the relationship between current technologies and human intelligence. CI has extended its area of inquiry to include both human-computer interaction and the study of the effects of technology on the flow of work and information in clinical settings [50–53]. Accordingly CI is well-positioned to inform the integration of AIM systems into clinical practice, and more broadly to inform the design of AI systems that complement the cognitive capabilities of human decision makers, in alignment with seminal ideas concerning the potential of cooperative human-machine systems [54].

The Complementarity of Human and Machine Intelligence

As is discussed in Chap. 5, evaluations in the context of image processing tasks have demonstrated that the performance of human beings and machines working in concert can result in better diagnostic accuracy than either machines or human beings alone [55–57]. In some ways this is not surprising, given the different strategies human experts and machines employ to achieve diagnostic accuracy. Cognitive studies of radiologists have shown that experts in this domain integrate their knowledge of anatomical structures and their projections onto two-dimensional images, with their knowledge of general physiology and specific disease processes. This allows radiologists to generate initial hypotheses that narrow the focus of their search for a definitive diagnosis [47]. In contrast, contemporary neural network models learn to identify radiological abnormalities by training two-dimensional "feature detectors" to recognize regions that are useful in distinguishing between diagnostic categories in the training data (as illustrated previously, in Fig. 1.3), irrespective of where within an image these regions may occur [58]. Differences in the processes through which neural networks and human experts interpret images can also be detected empirically. Recent work has shown that human beings and machines focus on different features when interpreting histology slides [59].

Acknowledgment of these differences leads naturally to the conclusion that a human/AI collaborative team has the potential to make better decisions than those that would emerge from a fully automated or exclusively manual process (see, for example, the discussion of QMR in Chap. 2). However, many open questions remain regarding how best to realize this potential. A promising proposal concerns deliberately designing AI systems to compensate for known "blind spots" in clinical decision making [60], such as biases in diagnostic reasoning that have been identified through cognitive research [61], or distracted attention in busy clinical settings [62]. Alternatively, one might envision developing ways to distribute labor across a human/AI collaborative system to maximize the expected utility of this system, taking into account both the accuracy of model predictions and the time required for a human actor to reassess them. Recent work has developed an approach to optimizing collaborative systems in this way, resulting in some experiments in systems that increase high-confidence predictions (i.e. predictions to which the model assigns extremely high or low probability) at the expense of its accuracy in edge cases (i.e. predictions close to the model's decision boundary), where human input could resolve model uncertainty [63].

Mediating Safe and Effective Human Use of AI-Based Tools

CI methods are already well established as means to evaluate the **usability** of decision support tools [45, 46]. Findings from this line of research have led to recommendations that the usability of clinical systems should be prioritized as a means to enhance their acceptability and safety [64]. In contrast to system-centric methods of usability evaluation, such as heuristic evaluations by usability experts [65], CI approaches attempt to understand the thought process of a user, which is particularly important in knowledge-rich domains, such as medicine, where both knowledge of the system being used and of the domain are required to perform tasks correctly [66]. This can be accomplished through analysis of a **think-aloud protocol**, collected by prompting users to verbalize their thoughts during the process of completing representative tasks [67]. This approach is similarly well-suited to the study of clinician interactions with AI-based systems, where users must make clinical decisions on the basis of their estimation of the veracity of system output.

Critical questions concerning the nature of these interactions remain to be answered. One such question concerns how best to represent model predictions. For example, recent work in dermatology diagnosis found that advantages in performance for a human-computer collective were contingent upon the granularity (probabilities of all of the diseases in the differential diagnosis vs. a single global risk of malignancy) and cognitive demand of the representation used to convey predictions to physicians [57]. Analysis of verbal protocols collected during interactions with interfaces, using alternative representations of the same predictions, could inform our understanding of *why* this is the case by revealing the reasoning dermatologists use when deciding whether to accept a particular recommendation. Another important question concerns the role of explanations provided by a system in influencing human decision making. Intriguingly, recent research has shown that revealing the influence of input features (here, words in a passage of text) on model predictions increases the likelihood that users will accept the validity of these predictions, irrespective of whether they are accurate [68]. This suggests that displaying feature salience may not be adequate to support the fault detection procedures that are a prerequisite to safe and resilient human/AI collaborative systems. CI methods are well-suited to identify the thought processes through which faulty AI decisions are (or are not) identified when considering explanations, to inform the development of effective systems in which process are *both* highly automated *and* subject to human control. This should arguably be the case for systems making critical medical decisions, where mistakes have irreversible consequences [69].

Concluding Remarks

In this chapter, we have provided an introduction to AIM, with a focus on recent developments. In doing so, we have highlighted some key challenges that AI models must meet if they are to achieve the goal of improving the efficiency, safety and quality of health care. We have argued that the field of CI is well-suited to address these challenges, by providing greater insight into the role of the human component of human/AI collaborative systems, to inform their design and evaluation. Consideration of the cognitive processes through which human beings evaluate, interpret and act upon the recommendations made by AI systems is fundamental to the development of solutions that enhance the capabilities of clinicians and researchers in the biomedical domain. Accordingly, one of our goals in developing this volume has been to provide a resource to support the multidisciplinary training required to design and implement AI methods with the potential to enhance the practice of medicine as well as life science research in human biology.

Questions for Discussion

- What is an example of a recent technological advancement in AIM, and what are its implications for clinical practice?
- Provide your own definition of AIM that reflects the discussion in this chapter (i.e., do not simply pick one from Box 1.2). Do any aspects of the field of which you are aware fall outside the scope of this definition?
- What are the main application areas and techniques for AIM?
- AI in medicine has a long history, and AIM technologies have been proposed as a potential disruptor of the healthcare industry before. What current contextual factors might increase or limit the potential for broad adoption?

Further Reading

Chang, AC. Intelligence-Based Medicine: Artificial Intelligence and Human Cognition in Clinical Medicine and Healthcare. Academic Press (Elsevier); July 8th 2020.

- This book provides a survey of AI methods from clinical and data science perspectives, with an emphasis on their implementation in, and impact upon, medicine and its subspecialties.

Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. Briefings in Bioinformatics. 2018 Nov 27;19 (6):1236–1246.

- This paper provides an overview and of deep learning applications in healthcare up to 2018, and introduces a number of issues that are addressed in the current volume.

Patel VL, Kannampallil TG. Cognitive informatics in biomedicine and healthcare. Journal of biomedical informatics. 2015 Feb 1;53:3–14.

- This paper provides a definition and overview of the field of cognitive informatics, with a focus on biomedical applications.

Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nature Medicine. Nature Publishing Group; 2019 Jan;25 (1):44–56.

- This paper provides an overview of AI applications in healthcare, including a thoughtful account of challenges that distinguish this domain from others in which AI applications have established their value.

Zhang D, Mishra S, Brynjolfsson E, Etchemendy J, Ganguli D, Grosz B, Lyons T, Manyika J, Niebles JC, Sellitto M, Shoham Y, Clark J, Perrault R. The AI Index 2021 Annual Report. arXiv:210306312 [cs] [Internet]. 2021 Mar 8 [cited 2021 Apr 24]; Available from: http://arxiv.org/abs/2103.06312

- Stanford's AI Index Report provides an overview of national and global AI trends in research and industry.

References

1. McCarthy J, Minsky ML, Rochester N, Shannon CE. A proposal for the Dartmouth summer research project on artificial intelligence, August 31, 1955. AIMag. 2006;27(4):12.
2. Google Books Ngram Viewer [Internet]. [cited 2021 June 25]. Available from: https://books.google.com/ngrams.
3. Yu VL, Buchanan BG, Shortliffe EH, Wraith SM, Davis R, Scott AC, Cohen SN. Evaluating the performance of a computer-based consultant. Comput Programs Biomed. 1979;9(1):95–102.
4. Rosenblatt F. The perceptron: a probabilistic model for information storage and organization in the brain. Psychol Rev. 1958;65(6):386.
5. McClelland JL, Rumelhart DE, Group PR. Parallel distributed processing. Boston, MA: MIT Press; 1986. p. 1.
6. Swartout WR. Explaining and justifying expert consulting programs. Computer-assisted medical decision making. Springer; 1985. p. 254–71.
7. Shortliffe EH, Davis R, Axline SG, Buchanan BG, Green CC, Cohen SN. Computer-based consultations in clinical therapeutics: explanation and rule acquisition capabilities of the MYCIN system. Comput Biomed Res. 1975;8(4):303–20.
8. Hinton GE, Osindero S, Teh Y-W. A fast learning algorithm for deep belief nets. Neural Comput. 2006;18(7):1527–54.

9. Rajpurkar P, Zhang J, Lopyrev K, Liang P. Squad: 100,000+ questions for machine comprehension of text. In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
10. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L. Imagenet: a large-scale hierarchical image database. 2009 IEEE conference on computer vision and pattern recognition. IEEE; 2009. p. 248–55.
11. Panayotov V, Chen G, Povey D, Khudanpur S. Librispeech: an ASR corpus based on public domain audio books. 2015 IEEE international conference on acoustics, speech and signal processing (ICASSP). 2015. p. 5206–5210.
12. LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–44.
13. Kumar G, Bhatia PK. A detailed review of feature extraction in image processing systems. 2014 fourth international conference on advanced computing communication technologies. 2014. p. 5–12.
14. Bengio Y, Courville A, Vincent P. Representation learning: a review and new perspectives. IEEE Trans Pattern Anal Mach Intell. 2013;35(8):1798–828.
15. Zhang D, Mishra S, Brynjolfsson E, Etchemendy J, Ganguli D, Grosz B, Lyons T, Manyika J, Niebles JC, Sellitto M, Shoham Y, Clark J, Perrault R. The AI Index 2021 annual report. arXiv:210306312 [cs] [Internet]. 2021 Mar 8 [cited 2021 Apr 24]. Available from: http://arxiv.org/abs/2103.06312.
16. AI Index 2021 [Internet]. Stanford HAI. [cited 2021 June 25]. Available from: https://hai.stanford.edu/research/ai-index-2021.
17. Shin H-C, Roth HR, Gao M, Lu L, Xu Z, Nogues I, Yao J, Mollura D, Summers RM. Deep convolutional neural networks for computer-aided detection: CNN architectures, dataset characteristics and transfer learning. IEEE Trans Med Imaging. 2016;35(5):1285–98.
18. Devlin J, Chang M-W, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 conference of the North American Chapter of the Association for computational linguistics: human language technologies, Vol. 1 (Long and Short Papers). 2019. p. 4171–4186.
19. Radford A, Wu J, Child R, Luan D, Amodei D, Sutskever I. Language models are unsupervised multitask learners. OpenAI Blog. 2019;1(8):9.
20. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: Visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE international conference on computer vision. 2017. p. 618–626.
21. Adler-Milstein J, Jha AK. HITECH act drove large gains in hospital electronic health record adoption. Health Aff. 2017;36(8):1416–22.
22. Bauman RA, Gell G, Dwyer SJ. Large picture archiving and communication systems of the world--part 1. J Digit Imaging. 1996;9(3):99–103.
23. Gulshan V, Peng L, Coram M, Stumpe MC, Wu D, Narayanaswamy A, Venugopalan S, Widner K, Madams T, Cuadros J. Development and validation of a deep learning algorithm for detection of diabetic retinopathy in retinal fundus photographs. JAMA. 2016;316(22):2402–10.
24. Esteva A, Kuprel B, Novoa RA, Ko J, Swetter SM, Blau HM, Thrun S. Dermatologist-level classification of skin cancer with deep neural networks. Nature. 2017;542(7639):115–8.
25. Rajkomar A, Oren E, Chen K, Dai AM, Hajaj N, Hardt M, Liu PJ, Liu X, Marcus J, Sun M, Sundberg P, Yee H, Zhang K, Zhang Y, Flores G, Duggan GE, Irvine J, Le Q, Litsch K, Mossin A, Tansuwan J, Wang D, Wexler J, Wilson J, Ludwig D, Volchenboum SL, Chou K, Pearson M, Madabushi S, Shah NH, Butte AJ, Howell MD, Cui C, Corrado GS, Dean J. Scalable and accurate deep learning with electronic health records. NPJ Digit Med. 2018;1(1):1–10.
26. Mukherjee S. A.I. versus M.D. [Internet]. The New Yorker. [cited 2021 Apr 15]. https://www.newyorker.com/magazine/2017/04/03/ai-versus-md.
27. Metz C. A.I. shows promise assisting physicians. The New York Times [Internet]. 2019 Feb 11 [cited 2021 Apr 15]. https://www.nytimes.com/2019/02/11/health/artificial-intelligence-medical-diagnosis.html.

28. O'Connor A. How artificial intelligence could transform medicine. The New York Times [Internet]. 2019 Mar 11 [cited 2021 Apr 15]. https://www.nytimes.com/2019/03/11/well/live/how-artificial-intelligence-could-transform-medicine.html.
29. Health C for D and R. Artificial intelligence and machine learning in software as a medical device. FDA [Internet]. FDA; 2021 Jan 11 [cited 2021 Apr 19]. https://www.fda.gov/medical-devices/software-medical-device-samd/artificial-intelligence-and-machine-learning-software-medical-device.
30. Benjamens S, Dhunnoo P, Meskó B. The state of artificial intelligence-based FDA-approved medical devices and algorithms: an online database. NPJ Digit Med. 2020;3(1):1–8.
31. The Medical Futurist [Internet]. The Medical Futurist. [cited 2021 Apr 19]. Available from: https://medicalfuturist.com/fda-approved-ai-based-algorithms.
32. Marcus G. Deep learning: a critical appraisal. arXiv preprint arXiv:180100631. 2018.
33. Zador AM. A critique of pure learning and what artificial neural networks can learn from animal brains. Nat Commun. 2019;10(1):1–7.
34. Marr D. Artificial intelligence—a personal view. Artif Intell. 1977;9(1):37–48.
35. Winston PH. Artificial Intelligence. Reading, MA: Addison-Wesley; 1977.
36. Barr A, Feigenbaum EA. The handbook of artificial intelligence (Vol. 1). Los Altos, CA: William Kaufman; 1981.
37. Luger GF, Stubblefield WA. Artificial intelligence (2nd ed.): structures and strategies for complex problem-solving. USA: Benjamin-Cummings Publishing Co., Inc.; 1993.
38. McCarthy J. What is artificial intelligence? [Internet]. What is artificial intelligence? 2007 [cited 2021 Apr 20]. http://www-formal.stanford.edu/jmc/whatisai/whatisai.html.
39. Poplin R, Varadarajan AV, Blumer K, Liu Y, McConnell MV, Corrado GS, Peng L, Webster DR. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. Nat Biomed Eng. 2018;2(3):158–64.
40. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;15:1–11.
41. Berg M. Patient care information systems and health care work: a sociotechnical approach. Int J Med Inform. 1999;55:87–101.
42. Shortliffe T, Davis R. Some considerations for the implementation of knowledge-based expert systems. SIGART Bull. 1975;55:9–12.
43. Topol EJ. High-performance medicine: the convergence of human and artificial intelligence. Nat Med. 2019;25(1):44–56.
44. Wang Y. The theoretical framework of cognitive informatics. Int J Cogn Inform Nat Intell. 2007;1(1):1–27.
45. Patel VL, Kaufman DR. Cognitive science and biomedical informatics. In: Shortliffe EH, Cimino JJ, editors. Biomedical informatics: computer applications in health care and biomedicine. 5th ed. New York: Springer; 2021. p. 133–85.
46. Patel VL, Kannampallil TG. Cognitive informatics in biomedicine and healthcare. J Biomed Inform. 2015;53:3–14.
47. Lesgold A, Rubinson H, Feltovich P, Glaser R, Klopfer D, Wang Y. Expertise in a complex skill: diagnosing x-ray pictures. In: Chi MTH, Glaser R, Farr MJ, editors. The nature of expertise. Hillsdale, NJ: Lawrence Erlbaum; 1988. p. 311–42.
48. Elstein AS, Shulman LS, Sprafka SA. Medical problem solving: an analysis of clinical reasoning. Cambridge, MA: Harvard University Press; 1978.
49. Patel VL, Arocha JF, Kaufman DR. Diagnostic reasoning and medical expertise. Psychol Learn Motiv. 1994;31:187–252.
50. Kushniruk AW, Patel VL, Cimino JJ. Usability testing in medical informatics: cognitive approaches to evaluation of information systems and user interfaces. Proceedings/AMIA annual fall symposium. 1997. p. 218–222.

51. Kushniruk AW, Patel VL. Cognitive and usability engineering methods for the evaluation of clinical information systems. J Biomed Inform. 2004;37:56–76.
52. Malhotra S, Jordan D, Shortliffe E, Patel VL. Workflow modeling in critical care: piecing together your own puzzle. J Biomed Inform. 2007;40:81–92.
53. Cohen T, Blatter B, Almeida C, Shortliffe E, Patel V. A cognitive blueprint of collaboration in context: distributed cognition in the psychiatric emergency department. Artif Intell Med. 2006;37:73–83.
54. Licklider JC. Man-computer symbiosis. IRE transactions on human factors in electronics. IEEE. 1960;1:4–11.
55. Patel BN, Rosenberg L, Willcox G, Baltaxe D, Lyons M, Irvin J, Rajpurkar P, Amrhein T, Gupta R, Halabi S, Langlotz C, Lo E, Mammarappallil J, Mariano AJ, Riley G, Seekins J, Shen L, Zucker E, Lungren MP. Human–machine partnership with artificial intelligence for chest radiograph diagnosis. NPJ Digit Med. 2019;2(1):1–10.
56. Hekler A, Utikal JS, Enk AH, Hauschild A, Weichenthal M, Maron RC, Berking C, Haferkamp S, Klode J, Schadendorf D, Schilling B, Holland-Letz T, Izar B, von Kalle C, Fröhling S, Brinker TJ, Schmitt L, Peitsch WK, Hoffmann F, Becker JC, Drusio C, Jansen P, Klode J, Lodde G, Sammet S, Schadendorf D, Sondermann W, Ugurel S, Zader J, Enk A, Salzmann M, Schäfer S, Schäkel K, Winkler J, Wölbing P, Asper H, Bohne A-S, Brown V, Burba B, Deffaa S, Dietrich C, Dietrich M, Drerup KA, Egberts F, Erkens A-S, Greven S, Harde V, Jost M, Kaeding M, Kosova K, Lischner S, Maagk M, Messinger AL, Metzner M, Motamedi R, Rosenthal A-C, Seidl U, Stemmermann J, Torz K, Velez JG, Haiduk J, Alter M, Bär C, Bergenthal P, Gerlach A, Holtorf C, Karoglan A, Kindermann S, Kraas L, Felcht M, Gaiser MR, Klemke C-D, Kurzen H, Leibing T, Müller V, Reinhard RR, Utikal J, Winter F, Berking C, Eicher L, Hartmann D, Heppt M, Kilian K, Krammer S, Lill D, Niesert A-C, Oppel E, Sattler E, Senner S, Wallmichrath J, Wolff H, Gesierich A, Giner T, Glutsch V, Kerstan A, Presser D, Schrüfer P, Schummer P, Stolze I, Weber J, Drexler K, Haferkamp S, Mickler M, Stauner CT, Thiem A. Superior skin cancer classification by the combination of human and artificial intelligence. Eur J Cancer. 2019;120:114–21.
57. Tschandl P, Rinner C, Apalla Z, Argenziano G, Codella N, Halpern A, Janda M, Lallas A, Longo C, Malvehy J, Paoli J, Puig S, Rosendahl C, Soyer HP, Zalaudek I, Kittler H. Human–computer collaboration for skin cancer recognition. Nat Med. 2020;26(8):1229–34.
58. Soffer S, Ben-Cohen A, Shimon O, Amitai MM, Greenspan H, Klang E. Convolutional neural networks for radiologic images: a Radiologist's guide. Radiology. 2019;290(3): 590–606.
59. Kimeswenger S, Tschandl P, Noack P, Hofmarcher M, Rumetshofer E, Kindermann H, Silye R, Hochreiter S, Kaltenbrunner M, Guenova E, Klambauer G, Hoetzenecker W. Artificial neural networks and pathologists recognize basal cell carcinomas based on different histological patterns. Mod Pathol. 2020;13:1–9.
60. Horvitz E. One hundred year study on artificial intelligence: reflections and framing. Microsoft com. 2014
61. Chapman GB, Elstein AS. Cognitive processes and biases in medical decision-making. In: Chapman GB, Sonnenberg FS, editors. Decision-making in health care: theory, psychology, and applications. Cambridge: Cambridge University Press; 2000. p. 183–210.
62. Franklin A, Liu Y, Li Z, Nguyen V, Johnson TR, Robinson D, Okafor N, King B, Patel VL, Zhang J. Opportunistic decision making and complexity in emergency care. J Biomed Inform. 2011;44(3):469–76.
63. Bansal G, Nushi B, Kamar E, Horvitz E, Weld DS. Is the Most accurate AI the best teammate? Optimizing AI for teamwork. Proc AAAI Conf Artif Intell. 2021;35(13):11405–14.
64. Middleton B, Bloomrosen M, Dente MA, Hashmat B, Koppel R, Overhage JM, Payne TH, Rosenbloom ST, Weaver C, Zhang J. Enhancing patient safety and quality of care by improving the usability of electronic health record systems: recommendations from AMIA. J Am Med Inform Assoc. 2013;20(e1):e2–8.
65. Nielsen J, Molich R. Heuristic evaluation of user interfaces. Proceedings of the SIGCHI conference on human factors in computing systems. 1990. p. 249–256.

66. Horsky J, Kaufman DR, Oppenheim MI, Patel VL. A framework for analyzing the cognitive complexity of computer-assisted clinical ordering. J Biomed Inform. 2003;36:4–22.
67. Ericsson KA, Simon HA. Protocol analysis: verbal reports as data. Cambridge, MA: MIT Press; 1993.
68. Bansal G, Wu T, Zhou J, Fok R, Nushi B, Kamar E, Ribeiro MT, Weld D. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of the 2021 CHI conference on human factors in computing systems. New York, NY: Association for Computing Machinery; 2021. p. 1–16. https://doi.org/10.1145/3411764.3445717.
69. Shneiderman B. Human-centered artificial intelligence: reliable, safe & trustworthy. Int J Hum Comput Interact. 2020;36(6):495–504.