PeriGen
Advanced Perinatal Systems

Artificial Intelligence: A Primer for Perinatal Clinicians

June 14, 2018
Emily F. Hamilton, MD CM
Philip Warrick, PhD

Table of Contents

Introduction

Hardly a day goes by without some headline extolling the accomplishments of artificial intelligence (AI). It is likely that the last phone call you received alerting you to unusual activity on your credit card was prompted by an AI-based fraud detection system. AI applications surround us, correcting poor grammar, proposing the next word in a text message, providing blind spot warnings while driving or suggesting our next purchase based on our recent web browsing. AI has been with us for decades. Why is there a sudden uptick in interest? It has gotten much better and more relevant. Good news—the best is yet to come.

While spell checking and shopping suggestions are generally received as helpful and innocuous, medical AI applications evoke more cautious responses. Will I be replaced? Will I be misled and make an error or be made to look bad? These concerns and recent scientific developments have prompted this primer on AI relevant to perinatal care. The following text will explain in lay terms what AI can and cannot do, and why nursing and medicine are AI-safe professions.

What is Human Intelligence?

Human intelligence is a collection of many capacities including the ability to use language, learn, reason, understand, create, plan, problem solve, separate truth from fiction, be self-aware and have emotional knowledge. It involves processing information received by the five senses. It is not purely rational as it is affected by interactions with others, goals, beliefs, intuition and biological instincts.

What is Artificial Intelligence?

Artificial intelligence (AI) refers to a collection of mathematical techniques used to accomplish one or more human cognitive functions.

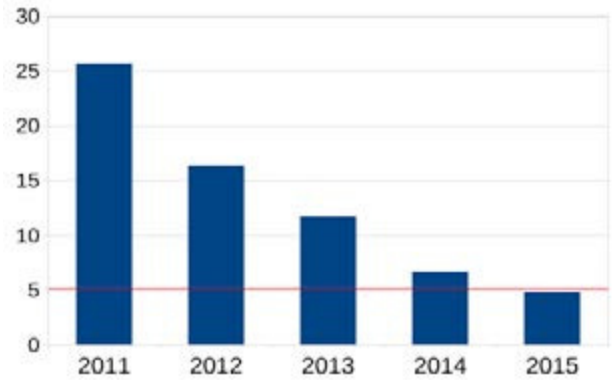
Artificial intelligence (AI) refers to a collection of mathematical techniques used to accomplish one or more human cognitive functions. Common problems for AI to solve include: visual pattern recognition, speech recognition, decision-making, understanding languages and operating robots. The two major approaches used in AI include logic and rule-based techniques and machine learning. Machine learning refers to techniques where internal algorithms find patterns in data and measure the strength of their association with the outcome of interest. With the presentation of more and more examples, the computer “learns” what associations are strongest and then applies that experience to the next set of data it receives.

Although AI techniques may excel at logic or solve a specific problem, they are not capable of real understanding or reasoning. Given the long list of cognitive processes we use to describe human intelligence, it is somewhat generous to apply the term “intelligence” to a computer program that can accomplish only one or two of these processes. That said, some artificial intelligence applications can perform specific tasks extremely well and highly consistently. For example, current AI-based applications can consistently outperform humans in games like Scrabble and Chess but do worse than humans at tasks like facial recognition or translation.[1]

One of the techniques used to compare human and artificial intelligence is the Turing test. In 1950, the acclaimed pioneer in computer science Alan Turing, proposed using a conversation to judge the capacity for a computer to converse in natural language and pass as a human. The Turing test is considered successful if the blinded human participant cannot reliably tell whether the conversation is with a machine or human. Earlier this year, Google demonstrated Google’s Duplex, an-AI based digital assistant, calling to make a restaurant reservation or haircut appointment. In both examples, the computer could converse about preferences, deal with nuanced human conversation and achieve its mission in a convincingly human fashion. The four-minute video clip of Google Duplex making an appointment is well worth watching. [2]

The performance of AI has improved rapidly in recent years. Each year, Stanford University holds a competition in machine vision to determine the ability of AI-based programs to identify objects in images. From its databank of 10 million labelled images, contestants receive the same set of images for training and then are tested on an independent set of 200,000 images. Figure 1 shows the error rate of the winner over time. The red line indicates human performance. Human performance was matched in 2015. [3]

Figure 1. Decreasing error rates in image recognition in the Stanford ImageNet competition



Why has AI performance improved so much?

The coexistence of four synergistic factors is contributing to rapid advances in AI.

1.Big data

Abundant and accurate data is prerequisite for training most AI systems. The daily production of internet data is estimated to be around 2.5 quintillion bytes (2.5 x1018). Given that a single letter is one byte and a typical novel is about 40,000 words, this data production is equivalent to about 12.5 trillion novels each day! In healthcare, computing devices embedded in technologies including wearable sensors, inpatient monitors and electronic medical records, can connect and exchange data producing an immense amount of detailed information.

2. Massive computing power

Training complex AI systems on huge datasets is computationally demanding. The invention and availability of the Graphical Processing Unit (GPU) provided a giant leap in computational capacity. A single GPU processor typically contains hundreds of cores and can run processes in parallel compared to a CPU which contains only a few cores and performs its calculations much slower. In 2011, Google's brain project used 2,000 CPUs to train an AI system to recognize the difference between cat and dog videos. A few years later, the same task was accomplished using a set of 12 GPUs. With GPUs, it is now technically and financially feasible for the public to acquire fast and extensive computing power.[4]

3. Vastly more powerful AI techniques

The techniques underpinning AI are also advancing in concert with hardware improvements. It is beyond the scope of this article to describe these advances in detail but key to underline that these techniques consider many more subtle aspects, remember past tendencies, and continuously learn from past success and failures.

4. Greater accessibility

Another significant advance in 2016 was the release of TensorFlowTM by Google. TensorFlow is an open source software platform that was originally developed by researchers and engineers working on the Google Brain Team within Google's Machine Intelligence research organization for the purposes of conducting machine learning and deep neural network research. In short, anyone can use this platform to develop AI applications, taking advantage of many traditional machine learning tools as well as an extensive state of the art deep learning module and have access to improvements as they become available. Within the first two years there were 11 million downloads of TensorFlow across the world. In 2018, Google released the Cloud Machine Learning service

that makes it possible to run TensorFlow on Google's extensive software and hardware infrastructure. [5]

AI Techniques and Terminology

Several computing techniques can perform human-like functions and hence could fall under the rubric of AI. Machine learning methods use software algorithms designed to find the most discriminating or predictive arrangement of factors for a known outcome. They do not rely upon predetermined assumptions about what is important, or linear relationships or what cutoff values are practical. They let the data speak for themselves. We will concentrate on two machine learning techniques—Decision Trees and Deep Learning Neural Networks—and provide some perinatal examples. Both techniques are well-defined mathematical procedures, and both resemble how we think as clinicians.

> Machine learning methods do not rely upon predetermined assumptions about what is important.

Decision Trees

As clinicians, we are familiar with management algorithms—a tree with a sequence of steps and branches. Such algorithms are usually designed by a panel of experts and based on published literature, personal experience and practicality. The branching criteria are often kept simple to be easy to remember and follow. Programming such an algorithm is not considered AI or machine learning as it is a transcription of human opinions.

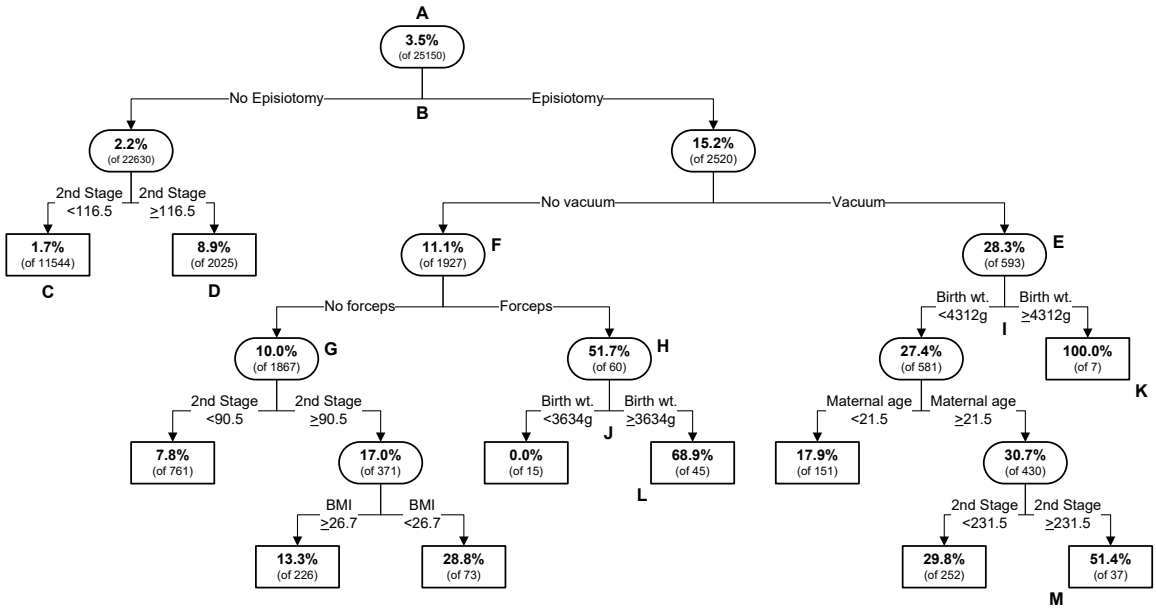
The technique of Classification and Regression Trees (CART) is a machine learning technique that will generate a tree that can be displayed in a familiar graphical format. From all variables under consideration, the procedure CART selects the single factor that best separated the group into those with and those without the outcome of interest to form the first branch point or node. Once the first node is formed, the same procedure finds the next most discriminating factor and forms the "child" node. At each junction CART identifies the optimal cutoff value for continuous variables. Splitting stops when the process determines that there is no further discriminating advantage with any of the remaining factors. The resulting structure defines clusters of risk factors that often bear resemblance to clinical reasoning. Furthermore, the probability of the outcome in question is calculated in each terminal leaf.

The tree structure provides clinicians with the opportunity to see the decision-making pathway and to match that against clinical experience. It is easier to have confidence in a process with visible steps that make clinical sense compared to one where the internal steps are hidden.

CART applications are developing in a wide range of clinical situations, such as the prediction of outcomes with obesity, diagnosis of neurological diseases, the prediction of cardiovascular outcomes the identification of subgroups with different risks in epidemiologic investigations.[6-10] Figure 2 displays a tree devised by CART that we created for estimating the probability of severe perineal laceration based on maternal size, parity, birthweight and duration of second stage and method of delivery.[11] A patient in leaf C has a 1.7% probability of a severe perineal laceration whereas a patient in leaf L has a 68.9% chance.

Figure 2. A Classification and Regression Tree for predicting likelihood of severe perineal laceration.

The branching points of this tree do reflect well-known risk factors for perineal laceration. If they did not, we would have sound reasons to be skeptical. It differs



from general clinical trees by providing precise and optimal cutoff points that are not rounded or simplified to be easy to remember. Also, it shows the estimated probability of severe laceration in each of the various clusters. The clusters, branching criteria and probability are all derived from actual data.

Neural Networks

Artificial neural networks are mathematical processes that bear some resemblance to biological neural pathways in the brain. The connections between specific neurons gain weight (statistical importance) based on the strength of their association with the final condition of interest.

The psychologist Donald Hebb observed that repeated activation of a specific neural pathway changed the actual physical connection between the affected neurons causing improved synaptic efficiency. This was one of the earliest descriptions of neurophysiological basis for learning with repeated practice. Artificial neural networks are mathematical processes that bear some resemblance to biological neural pathways in the brain. Artificial neural networks contain artificial neurons which are arranged in layers with connections between adjacent layers. An artificial neuron is a mathematical process that converts a combination of inputs to a number which is passed to the neurons in the next layer. The connections between specific neurons gain weight (statistical importance) based on the nature and strength of their association with the final condition of interest.

A simple example helps to describe the process. Suppose that a neural net's objective is to pick out the oranges in a collection of different kinds of fruit. It is a simple feed forward neural network with supervised learning. Feed forward refers to the fact that each layer affects only subsequent layers. Supervised learning means that the neural network is trained on a specific dataset where the neural network is told which fruits are oranges (and which are not).

During training, the neurons in the first layer receive inputs, namely the various features of the objects in the basket, say surface color, weight, shape, volume and diameter. Each neuron in the first layer processes the feature information in a different way. For example, one neuron might calculate the objects density and based on repeated observations where it knows what fruits are oranges, determine that if density > 1.25g/L the object is more likely to be an orange than another fruit. If the result of the density calculation is repeatedly and reliably associated with the orange it will obtain heavier weights. This means that in a second layer this result is more influential than another neurons output with a lesser weight.

Each first layer neuron sends its results to the neurons in the second layer. Second layer neurons make more complex decisions based on the messages from the first layer. With more artificial neurons per layer and more layers the

artificial neural network can consider a wider range of combinations and more complex relationships. The final output is a single probability that the object is an orange. Several different pathways through the neural network may lead to the conclusion “orange.” Each pathway is a different mathematical formula and its weight reflects how strong an association it has with “oranges.” Neural networks allow flexibility in the way a final decision is made. Different pathways may lead to the conclusion that the fruit is an orange. This capacity is attractive for healthcare issues because there is often considerable biological variation in the appearance of a specific disease and its symptoms.

Google’s TensorFlow playground lets you alter the size of a neural network and see how well it can solve a challenge.

Neural networks may be designed to have several final outputs such a probability that the object is one of several kinds of fruit. They may use different ways of learning, for example, learning as they gain experience in the real world, or by placing a greater penalty on certain kinds of errors. For neural networks concerned with processes, like the evolution of a disease, they can remember and use information on the patient’s recent and remote past.

Google has designed a TensorFlow playground that lets you alter many parameters such as the number of layers and neurons to see how well the neural network can solve a challenge.[12] By running it you will see how the neural network learns as it is presented with more and more training cases. Some examples are shown in Figure 3 and 4.

Figure 3. A sample problem from the TensorFlow playground.

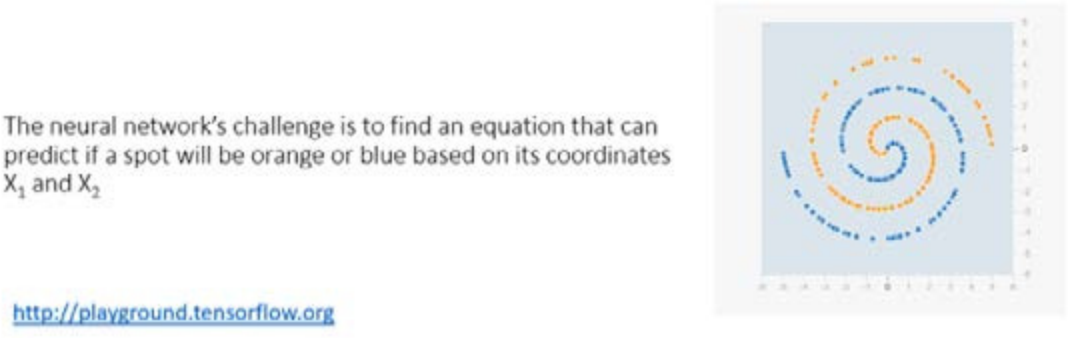


Figure 4 shows results with three different neural network arrangements with increasing complexity. The colors in the box indicate the neural network prediction regarding color. The intensity of the color shows how confident the prediction is. The weights of the connections are indicted by the line thickness.

The top very simple neural net can only find simple decision boundaries. The middle one has more inputs and can find more complex decisions boundaries but it is not very confident about most regions within the colored square. The bottom neural net with more neurons and more layers is able to find very complex decision boundaries and be far more confident about all decisions. Once the neural network is trained, it is tested on a dataset that it has never seen before. Testing loss reflects the error rate of the final neural network.

Figure 4. Three examples of neural networks from the TensorFlow playground.

Deep Learning Neural Networks

Increasing computational power of computers has facilitated the development of deep learning networks with many more layers, more neurons and different internal mechanisms for learning.

> Deep learning neural networks have many more layers and neurons and different internal learning mechanisms. With sufficient training data and these neural networks can distinguish more complex and nuanced relationships compared to simpler neural networks.

In 2016, Google Deep Mind researchers reported a landmark in deep learning research.[13] They created a program, AlphaGo, to play the Chinese board game Go. The number of possible moves in this game is vast, around 250^{150}. Thus, it is not feasible to search every possible combination of moves to find the best move. AlphaGo beat a world class professional master layer of Go 5 games to 0 and won over other Go programs in 99.8% of games. Their approach used a combination of two deep neural networks. The "Policy network" learned the general moves that experts played, the second "Value network" did a limited simulation of possible future moves based on the Policy Network's top decisions, in order to choose the best next move. Each neural network contained 13 layers, around 2,500 neurons and over 25,000 weights (connections). This achievement represented a landmark because AlphaGo's success was based on a combination of decisions related to strategy and a limited look ahead at possible consequences rather that a calculation of every possible sequence of moves to find the best choice.

In merely one year, the same team topped this achievement with AlphaGo Zero! [14] AlphaGo Zero learned just from playing Go. The initial neural network knew only the basic rules. It then played games against itself. As it played, the neural network was updated based on the experience and outcome of past games. After 40 days of training (and 4.9 million games!) it beat all previous versions of AlphaGo. (100 games to 0).

This result demonstrates the power of continuous learning from experience. It also raises two important points about the learning experience. Neural networks that learn from human judgement may inherit the limitations of that human judgement. If a neural work learns on a biased sample of data, it will develop the same bias.

In medical applications it is very important that learning be carried out on an accurate and truly representative dataset because an artificial neural network is not truly intelligent, it cannot reason, cannot distinguish truth from falsehood or generalize to a markedly different situation that it has not experienced during the training phase.

Deep Learning for EFM Pattern Recognition

During the past few years we have applied Google's TensorFlow, a deep learning neural network to the problem of fetal heart rate pattern recognition and seen considerable improvement in performance compared to what was possible with a series of much smaller neural networks available in the past.

Measuring the accuracy of computerized pattern recognition is a challenge because there is no "Gold Standard," that is, no formal set of labeled tracings in the industry or by national professional associations that can be used as a standard against which new analysis techniques can be compared. Clinician inconsistency in labelling EFM patterns is well known.[15-17]

In the absence of an available gold standard for testing purposes we used a multi-case/multi-reader technique to create a Standardized Test Set. The Standardized Test Set includes 235.5 hours of digital tracings from 91 babies with gestational ages ranging from 28 weeks to term.

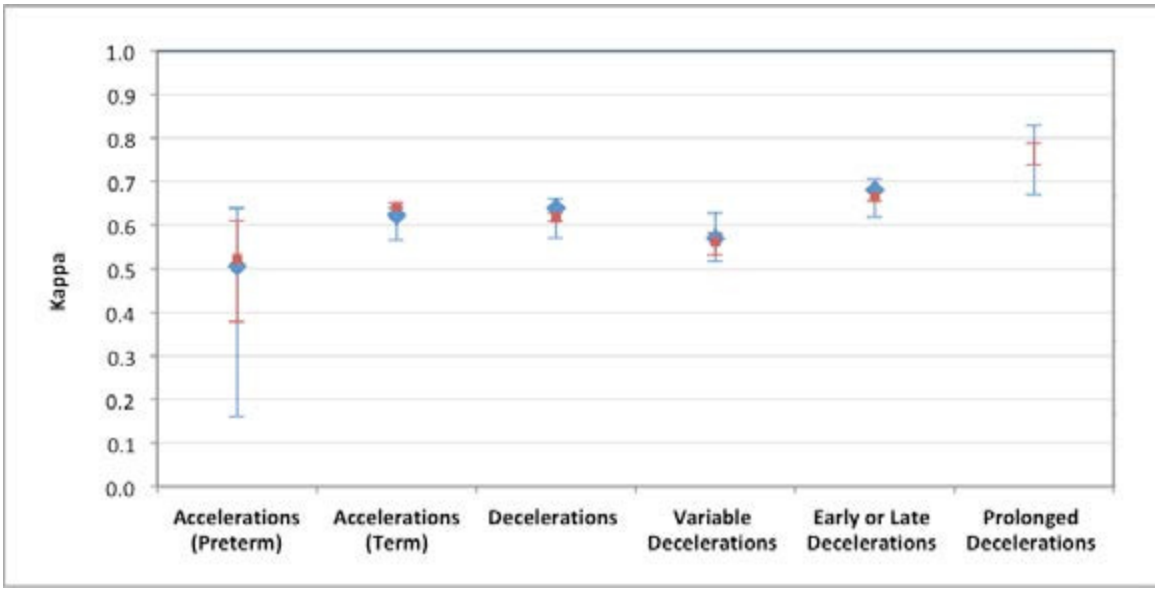
Each tracing was analyzed independently by six different clinicians using the standard NICHD and ACOG nomenclature and definitions. The clinician readers had a high level of expertise in electronic fetal monitoring and relevant clinical experience. The average experience in active labor and delivery care was 17.9 years and almost all had held senior teaching or academic positions. The group included representatives from nursing, midwifery, Obstetrics and Gynecology and Maternal Fetal Medicine.

Each acceleration or deceleration or baseline detected by a majority of the clinicians (at least 3 out of 5) became a Test Case. We measured the agreements and disagreements of each clinician reader with a set of majority opinions that did not involve that individual. The statistics on the agreements and disagreements for all clinicians constituted the Clinician Reference. We used the same Test Cases to assess the computerized methods.

In Figure 5, we show the performance using Kappa statistics of the computerized method (in red) based on a deep learning neural network and the Clinician Reference (in blue). Kappa statistics are useful because they consider both

agreements and disagreements beyond what is expected by chance. They are also used in most of the published reports regarding clinician agreement on EFM patterns. Since high agreement can come at the expense of disagreement and vice versa it is useful to have a statistic which considers both agreements and disagreements.[18] Perfect agreement would result in a Kappa statistic of 1.0 and random agreement would result in a Kappa statistic of 0.0.

Figure 5. Kappa statistics for the Clinician Reference (blue) and the Deep



Learning Neural Net (red)

A typical interpretation of the agreement level with Kappa scores is:
< 0.2 poor
0.2-0.39 fair
0.4-0.59 moderate
0.6-0.79 good
≥ 0.8 very good

These are remarkable results indicating that with deep learning it is possible to identify individual features in EFM tracings with a performance than on average is very similar to highly qualified clinicians working under ideal conditions.

Why is Healthcare an AI-Safe Profession?

There is little doubt that machine learning techniques can perform remarkable tasks; however, the idea of AI operating in clinical medicine does create angst. Skepticism is natural and prudent in a field where the consequences of error

are enormous. No one questions the superior capacity of a hand held $5 pocket calculator for mental arithmetic. So what causes resistance to AI in medicine?

> Healthcare is an AI-safe profession because it is complex, requires establishing human relationships and understanding and abstract reasoning. Computers don't form relationships, are not empathetic and cannot truly understand or reason.

In part, the term "intelligence" rankles. From philosophers to biologists, the capacity of the mind has been a defining characteristic of humanity. Rene Descartes wrote "I think therefore I am." Carl Linnaeus labelled the human species Homo Sapiens—man with wisdom or intelligence. Healthcare professionals spend many years acquiring knowledge, perfecting their clinical skills and passing certification examinations. Clinical acumen is central to our worth as clinicians and any challenge to our medical intelligence is not to be taken lightly. It is true AI techniques can excel at certain kinds of tasks but that does not equal human intelligence, and it certainly doesn't equal clinical judgement.

> Healthcare can be enhanced by AI because it provides consistent analysis that is objective, data driven and can counter human lapses related to wishful thinking, tunnel vision, boredom, inexperience and fatigue.

In part, the angst is promoted by sci-fi hype where human-like robots do replace humans. This scenario is far from reality in medicine. The medical tasks that AI performs currently are very circumscribed. Robots lift patients or steady instruments, image readers find tumors on x-rays or cytology smears, pattern recognition software labels features on EFM or EKG strips and probabilistic models estimate prognosis. These tasks form an important part of medical reasoning but only a part.

The tasks of clinicians are far more comprehensive. Even obtaining a complete and accurate history can be a challenge. Clinicians must illicit the medical history from patients and possibly their families, often considering nonverbal cues, rephrasing questions, reconciling inconsistences and resolving misunderstanding. Physical exam data and lab data need to be factored in including the possibility of lab error and missing data. Decisions and recommendations are based on these findings coupled with formal medical knowledge regarding physiology, pathology and therapies. Decisions need to consider patient preferences which in turn often require patient education.

Patients develop relationships with clinicians based on many factors including their confidence that the professional is competent and can be trusted to serve their best interest. In short, the relationship between healthcare professional and their patients is very multidimensional and constantly evolving. It is not purely rational. It is influenced by pain, fear, and intensely protective maternal instincts. Computers don't form relationships, are not empathetic and cannot truly understand or reason. However, AI can complement and make that professional more efficient, provide consistency and reduce errors. Wishful thinking, tunnel vision, recent vivid experiences, boring repetitive tasks, undue emotional overlay all contribute to human errors and many of these tendencies can be mitigated by application of machine learning techniques. Creativity, intuition, empathy, deep understanding and high-level reasoning rest squarely with clinicians.

Gary Kasparov the longstanding world champion chess player is a strong proponent of the man-machine combination. In 1997, he was defeated by IBM's chess playing program, Big Blue. After this defeat he began to explore the idea of human and computer, not human versus computer. Imagine the combination of his experience and computer memory, his intuition and computer calculation of probability, his passion and computer objectivity. A few years later the merits of his ideas were demonstrated in a free style competition where two amateur chess players assisted by three PCs won over master chess players. They also won over a strong chess playing program.

Notwithstanding the remarkable achievements in the science of machine learning, medical AI applications are relatively embryonic. In medicine, we cannot manufacture data for training as it is possible with games like Chess or Go. The creation of large datasets is now feasible due to the widespread use of electronic medical records, electronic monitoring data and digitized images in radiology and pathology. In turn, these massive datasets are providing crucial substrates for machine learning techniques.

The Fourth Industrial Revolution

Historians have described three industrial revolutions beginning with the steam engine, then the age of electricity and mass production followed by the digital age with computers and the internet. According to the World Economic Forum, an independent Swiss non-profit organization committed to global improvements, we have entered the fourth industrial revolution.[9] This is a period where technologies (AI, robotics, nanotechnology, quantum computing and the Internet of Things) are blurring the lines between the physical and biological spaces. This is a most exhilarating time for perinatal medicine and AI and the best is yet to come.

References

1. Progress in artificial intelligence. Wikipedia. https://en.wikipedia.org/wiki/Progress_in_artificial_intelligence. Accessed November 4, 2018.

2. Google duplex demonstration. ExtremeTech. https://www.extremetech.com/computing/269030-did-google-duplexs-ai-demonstration-just-pass-the-turing-test. Accessed November 4, 2018.

3. Stanford ImageNet Rompetition results. ImageNet and Wikipedia. http://www.image-net.org/challenges/LSVRC/ and https://en.wikipedia.org/wiki/Progress_in_artificial_intelligence#/media/File:Classification_of_images_progress_human.png. Accessed November 4, 2018.

4. Accelerating AI with GPUs: a new computing model. Nvidia. https://blogs.nvidia.com/blog/2016/01/12/accelerating-ai-artificial-intelligence-gpus/. Accessed November 4, 2018.

5. Google TensorFlow Summit—2018 keynote address. YouTube. https://www.youtube.com/watch?v=kSa3UObNS6o. Accessed November 4, 2018.

6. Lo BW, Fukuda H, Angle M, et al. Aneurysmal subarachnoid hemorrhage prognostic decision-making algorithm using classification and regression tree analysis. Surg Neurol Int. 07-Jul-2016;7:73.

7. Lee YC, Lee WJ, Lin YC, et al. Obesity and the decision tree: predictors of sustained weight loss after bariatric surgery. Hepatogastroenterology. 2009;56(96):1745–1749.

8. Salas-Gonzalez D, Górriz JM, Ramírez J, et al. Computer-aided diagnosis of Alzheimer's disease using support vector machines and classification trees. Phys Med Biol. 2010;55(10):2807–2817.

9. Ioannidis JP. Prediction of cardiovascular disease outcomes and established cardiovascular risk factors by genome-wide association markers. Circ Cardiovasc Genet. 2009;2(1):7–15.

10. Goldman L, Weinberg M, Weisberg M, et al. A computer-derived protocol to aid in the diagnosis of emergency room patients with acute chest pain. N Engl J Med. 1982;307(10):588–596.

11. Hamilton EF, Smith S, Yang L, Warrick P, Ciampi A. Third- and fourth-degree perineal lacerations: defining high-risk clinical clusters. Am J Obstet Gynecol. 2011;204(4):309.e1–e6.

12. Smilkov D and Carter S. Tinker with a neural network right here in your browser. Google TensorFlow Playground. https://playground.tensorflow.org. Accessed Nov. 4, 2018.

13. Silver D, Huang A, Maddison CJ, et al. Mastering the game of Go with deep neural networks and tree search. Nature. 2016;529(7587):484–489.

14. Silver D, Schrittwieser J, Simonyan K, et al. Mastering the game of Go without human knowledge. Nature. 2017;550(7676):354–359.

15. Chauhan SP, Klauser CK, Woodring TC, Sanderson M, Magann EF, Morrison JC. Intrapartum nonreassuring fetal heart rate tracing and prediction of adverse outcomes: interobserver variability. Am J Obstet Gynecol. 2008;199(6):623.e1–e5.

16. Blackwell SC, Grobman WA, Antoniewicz L, Hutchinson M, Gyamfi Bannerman C. Interobserver and intraobserver reliability of the NICHD 3-tier fetal heart rate interpretation system. Am J Obstet Gynecol. 2011;205(4):378.e1–e5.

17. Bernardes J, Costa-Pereira A, Ayres-de-Campos, Van Geijn HP, Pereira-Leite L. Evaluation of interobserver agreement of cardiotocograms. Int J Gynaecol Obstet. 1997;57:33–37.

18. Graphpad Statistical Software. QuickCalcs. https://www.graphpad.com/quickcalcs/kappa1.cfm. Accessed November 4, 2018.