Sometimes the digital Doctor should admit "I don't know"

Yoav Freund[1] **and Hau-Tieng Wu**[2]

[1]UCSD, department, city, postcode, country; [2]Duke, department, city, postcode, country

Digital technology is causing a sea-change in Medicine. In particular the meteoric rise of AI in general and deep learning in particular raises the possibility that doctors will be replaced computers (1). The father of deep learning, Geoff Hinton, said in 2017: "It's just completely obvious that that in ten years deep learning is going to do better than Radiologists ... They should stop training radiologists now".

Sebastian thrun (1, 4) , another authority in deep learning gives a more nuanced perspective: argues that "... deep learning devices will not replace dermatologists and radiologists. They will augment professionals, offering the expertise and assistance". Artificial Intelligence (AI) and Intelligence Augmentation (IA) have been competing ideologies for decades. (see inset) Currently, AI is much more in Vogue IA. We present an argument that, in high risk domains, and in particular in medicine, IA is the better approach.

Artificial Intelligence and Intelligence Augmentation

The driving question of AI is: "are machines capable of behaving in a way that is indistinguishable by humans". Achieving this goal implies that humans can be replaced by machines.

Using computers to augment humans rather replace them is both tantalizing and utterly mundane. On the heady side, consider cyborgs whose anatomy is part human, part artificial and can with equal ease solve complex equations or write poetry. On the mundane side, ubiquitous technologies such as the smart phone abd google search are ways in which our capabilities are augmented by computers.

The idea of using computers to augment or amplify human intelligence has a very long history. The acronyms AI (Artificial intelligence) and IA (Intelligence Amplification or Intelligence Augmentation) have both become popular in the early 1960's(2, 3). These days, the acronym AI is popular, while the acronym IA is not. However, Sebastian Thrun's statement indicates that the idea of Intelligence augmentation is still on some people's mind.

Our approach is based on the observation that the level of attention paid to a patient varies greatly. At the high end, a patient in surgery or in the Intensive Care Unit (ICU) has the full attention of several doctors and nurses. At the low end, an elderly person who might be suffring from early onset dementia might be visited by a nurse or home aid once a day or less. (please correct!)

Rather than replacing doctors or nurses, we suggest that IA can free the staff from simple and repetitive tasks and allow them to devote their time to more complex decisions and to prsonal interaction with the patient.

Central to our approach is a quantification of prediction confidence. Such quantification is needed so that a patient monitor can sound an alarm when a patient is having a heart attack, but create only few false alarms. Similarly, it is needed when a diagnostics assistant can eliminate some diagnostic possibilities but not all of them.

We equate low prediction confidence with saying "I don't know". The interaction between the IA agent and the doctor is based on this ability. When diagnosis is done by elimination, saying "I don't know" the agent can narrow the set of possible diagnosis without reducing it to a single diagnosis. It is then the doctor to decide how to proceed, whether to perform more tests, or whether to choose a treatment.

The rest of the article is organized as follows. In Section 1

1. Ground truth, train and test error

In a highly cited paper in the journal Science (4) provides evidence supporting the claim that computers can diagnose skin cancer as well or better than board certified dermatologists (see insert).

Not surprisingly, each dermatologist gave a different diagnosis. As the goal was to to compare the performance of the DNN to that of the dermatologists they needed a independent source of ground truth. To that end they used the diagnosis of a biopsy as ground truth. It is arguable that this label is more accurate than the one given by the dermatologist, even though it too depends on the human judgement of the pathologist.

Leaving aside the question of the reliability of the pathologists. There is another, more fundamental problem. The data in the experiment was retrospective, i.e. it was collected from the records of past patients. Normally, patients get biopsied only if the dermatologist thinks there is a significant chance of **malignancy**, the set of biopsied examples is therefor biased towards **malignancy**. It is likely that using a classifier trained in this way on an unfiltered stream of patients will increase the number of patients unnecessarily getting a biopsy.

Generating an unbiased sample would require a controlled experiment in which a sample of patients is chosen and each of these patients gets diagnosed using both a biopsy and an image of the skin.

Yoav : Do you have other examples of using NN for medical diagnostics?

Moreover, as we elaborate on in the next section, ground truth is rarely available, all we can go on are the diagnostics of human diagnosticians.

Supervised Learning and ground truth
Roughly speaking, machine learning (ML) can be divided into unsupervised learning and supervised learning. In both, the task of the learning algorith is transform a set of examples into a model. In unsupervised learning the examples are undifferentiated raw measurements. In supervised learning, which is the focus of this article, each example consists of an input and a label. Typically, the labels are provided by a human expert. These labels define the ground truth and the goal of the learning algorithm is to make predictions that diverge as little as possible from the ground truth.

Skin cancer diagnosis using Deep Neural Networks
One of the papers that provided evidence that deep neural networks might be able to outperform humans is the work of Esteva et al (4). They trained a Deep neural network to classify images of skin into three categories: benign, malignant and non-cancerous. The network was then tested, along with twenty five dermatologists on images which were labeled by a pathologist analysis of the biopsy. The neural network outperformed the human dermatologist. This is, without a doubt, an impressive finding. However, it is based on a retrospective analysis, in other words, an analysis of historical data. To predict the performance of the DNN when used in a dermatology practice we need to how a dermatologist, or any other diagnosticians, arrives at their final diagnostics.

2. Uncertainty in medicine

Yoav :

Deep learning is based on supervised learning, sometimes called input-output learning (5, 6). Under this fomulation the only quantity of interest is the error on a held-out or test set.

Medical diagnostics is very different from supervised learning. It is a process of eliminating possibilities until the set of possibilities is sufficiently narrow so that treatment can be chosen. The ground truth, as in the "correct diagnostics" is almost never revealed, even in hindsite.

When a new patient arrives at a clinique, all diagnostics are possible. After a physical exam and talking with the patient, many possibilities are eliminated. In simple cases, this is enough for the doctor to confidently choose a treatment. In increasingly more

complex cases, the doctor might ask for multiple tests and visits, refer the patient to a specialist, consult colleagues, journals and books etc. In most cases this process of narrowing will lead to a treatment plan. Ideally, followup visits and tests will confirm that the patient is recovering. In most cases, all that we can know about the patient is whether or not they recovered. This is a far cry from knowing what was their correct diagnostics when they first cam in the door.

For the digital doctor to be effective, it should incorporate this notion of uncertainty. It needs to correctly classify the easiest cases and out "I don't know" on the hardest cases. Before describing how this might be achieved, lets consider some of the many sources of uncertainty in medical diagnostics.

Protocol limitation

Yoav : For readers that are not MD, we should explain what are protocols, how they are generated, and whether all or some of their functionality ca be taken over by a computer. Also, I would put "extrapolation error" in here The American Academy of Sleep Medicine (AASM) publishes criteria for manual sleep stage and sleep apnea annotation from the gold standard sleep study instrument, the polysomnogram (PSG). This annotation is based on manual analysis of biosignals recorded from the PSG (11, 12). The AASM is a protocol that has been extensively applied, with rigorous scientific support, and updated regularly according to latest evidences. A detail sleep profile is critical for sleep quality enhancement, or even medical condition improvement. However, it is well known that even with the well established protocol, the inter-rater agreement rate of sleep stage annotation among experienced experts is only about 76% over normal subjects and about 71% over subjects with sleep apnea (13). Among many reasons, the one that is directly related to the intelligent system development is how the criteria are "described" in the protocol. For example, it is described in the protocol that if the delta wave occupies more than 20% of a given 30-second epoch of the electroencephalogram during sleep, that 30-second epoch is defined to be the N3 stage. 20% of a given 30-second epoch is 6 seconds. What about if the delta wave occupies 5.99-, or 6.01-seconds? What about if the delta wave sustains for 10 seconds, but it is divided into two consecutive 30-second epochs? When sitting on the "gray area" that is inherited from the protocol, sleep experts need to make a decision based on their experience or the information they have at hand, and this leads to medical uncertainties, and hence the inter-rater, or even intra-rater disagreement.

Yoav : can we add some specific κ values related to specific diagnostics?

Uncertainty due to signal quality

Medical devices use a variety of bio-sensors that record, display and distribute different biometrics, ranging from vital signs such as heart rate, oxygen saturation, temperature and blood pressure, to high-frequency waveforms such as ECG, EEG, respiratory signal and arterial blood pressure. ~~suffer from variety of problems~~ usually suffer from artifacts and other signal quality problems, some of which depend on the patient. ~~Reducing these problem often requires~~ In some cases, these problems can be easily handled by a human expert. In some cases such as artifact removal of EEG is still an active research problem (7).

Patient monitors are bedside medical devices that monitor patients at risk, freeing the medical staff to attend to the patients that need care at the moment. However, Patient monitors suffer from signal quality issues and tend to generate false alarms at a high rate. These cause the medical staff to ignore the alarms, rendering them useless. This phenomenon, called alarm fatigue (or alarm overload) is a major problem in hospital care (8, 9).

Another common source of signal quality issue is how the sensor is placed. While there have been several standards, ranging from the well-known ECG systems () and EEG systems () to recently smart clothing system for telemedicine (10), it is not always possible to achieve a precise sensor placement for biomedical signal collection due to various reasons. This uncertainty might be tolerable for some clinical applications; for example, an imprecise ECG sensor placement might not impact the identification of some types of arrhythmia from the ECG signal, like atrial fibrillation. However, this uncertainty might cause troubles in other applications; for example,

Quantification of inter-rater agreement rate

A common measurement of inter-rater reliability (or intra-rater reliability) for categorical quantities is the Cohen's kappa coefficient, usually denoted as κ. Compared with the percent agreement calculation, it considers the possible agreement occurring by chance. Specifically, while the percent agreement, denoted as $0 \leq a \leq 1$ is defined as the percent agreement among raters, the Cohen's kappa is defined as the ratio of $a - c$ and $1 - c$, where $0 \leq c \leq 1$ is the agreement occurring by chance. Clearly, the largest κ is 1, which means a complete agreement among raters, even under the possibility of agreement by chance. If the agreement is totally by chance, that is, $c = a$, then κ is 0. If the agreement is worse than agreement by chance, then κ can be negative. An interpretation of κ recommended by Cohen (14) is that when $\kappa \leq 0$, there is no agreement, when $0 < \kappa \leq 0.20$ as none to slight agreement, $0.2 < \kappa \leq 0.40$ as fair agreement, $0.4 < \kappa \leq 0.60$ as moderate agreement, $0.6 < \kappa \leq 0.80$ as substantial agreement, and $0.8 < \kappa \leq 1.00$ as perfect agreement. However, depending on scenarios, the meaning of agreement might be different.

Medical uncertainty as manifest low inter-rater agreement consequence, can be found in many clinical problems (see blocks on ...)For example, the low agreement might come from the "extrapolation error"; that is, when we apply the developed protocol to the population different from the population that we collect the evidence for the protocol (18). In other situation, the variability among subjects is so big that it limits the development of a more quantitative protocol (16). In some situations, when the needed information is missing, it is challenging to make a differential diagnosis (19).

A direct consequence of the low inter-rater agreement rate is that the trained intelligent system might be questionable. It is clear that such intelligent system is questionable and might raise concerns. Recently, various regulations in this regard have been proposed (20, 21).

Now, suppose we are able to eliminate all challenges from data calibration and validation issues, and we can provide as much information as possible to train the intelligence system. Even under this assumption, it is clear that the system still suffers from the protocol limitation or knowledge gap issues. Can such system be useful in clinics? To answer this question, we should not forget that physicians also follow the same protocol and have knowledge gaps. Depend-

knowledge gap

Yoav : is this an example of uncertainty because of knowledge gap? Wouldn't it be better to use something better known such as Covid-19? Urodynamic studies provide the best bladder and sphincter functional data for urologists to decide how to treat patients at risk for renal damage (15). While it has been extensively studied and applied in clinics, the main issue that plagues this field of urodynamics is the lack of precise definition of a detrusor contraction or overactive contraction (15). The lack of a well defined definition is due to the lack of quantitative study from the pathophysiological perspective, so the definition is still based on "expert opinion". For example, usually an overactive contraction represents itself as a "bump" in the detrusor pressure signal. However, what is the breath and height of a bump should we call it an overactive contraction? How to distinguish a true overactive contraction from an artifact? While there have been several reference information, like abdominal pressure, that could help us identify artifacts, but it can only explain a small portion of them.

Unsurprisingly, this fundamental issue has led to a significant inter-rater disagreement (16, 17).

ing on the clinical problems, and the experience of physicians under consideration, the agreement rate varies. Usually, intern doctors know the least, while a senior attending knows the most. It is natural that we trust a senior expert more, but it does not mean that we do not trust a junior intern doctor.

Managing uncertainty in medicine. Medical diagnosis is often uncertain or inconclusive, on the other hand, A doctor responsible for a patient's health has to make decisions in spite of this uncertainty. If the uncertainty present a sufficiently small risk, the doctor can choose a treatment. Otherwise the doctor might consult other doctors, a medical journal or a book.

To better understand the process and the possible place of AI in it, we turn to the Kahaneman's (22) "Thinking Fast Thinking Slow" and to Vordermark book on medical decision making (23).

Medical diagnosis can be divided into two main types: recognition and elimination. Recognition is a fast mental process that is mostly unconcious where the one correct diagnosis reveals itself in the doctors mind. Recognition does not lend itself to verbal description and is therefor hard to debate or document. As it typically points to a single diagnosis there is a danger that the recognized diagnosis will hide other possible diagnoses. Elimination, on the other hand, is a slow deliberate process where the doctor starts with all possible diagnoses and gradually eliminates impossible ones based on patient history, examination and test results. As Elimination is delibirative, it is easier to discuss and document it.

Inter-Rater agreement
A direct consequence of this low inter-rater agreement is a questionable trained "artificial intelligence". It is possible that we magically obtain a dataset that contains information that is sufficient for the decision making, while the information is too subtle so that it is not considered in the protocol, and we also magically obtain labels from a magical master that can see though all the information and provide the correct decision. However, by doing a simple math, we shall not count on such a magic and should come back to the protocol itself.

All medical decision making is based on experience, in which we include medical cases as well as knowledge learned from lectures Both recognition and elimination depend on past experience, if we include in past experience

IA can aid the doctor both in Recognition and in Elimination.

Sources of uncertainty in medical diagnosis.

- **The diagnostic process of elimination**
- **Data Quality, Calibration, resolution** Discuss issue as placement of sensors, .

Hiding Uncertainty

- **Psychological reasons** Both doctor and patient prefer the projection of certitude.
- **Protocols** –done
- **diagnostic devices** Secrecy of the internal code limits the trustworthiness of the alarms.–done
- **Alarm Fatigue**–done

How to quantify IDK? We should discuss how to quantify the confidence, or certainty, a physician has when making a decision.

Clearly, experience leads to confidence. With more experience aggregated, diagnostic options that contradict the accumulated experience are eliminated, and hence more problems that need to be handled by the elimination process can be handled by the recognition process. However, facing our complicated human body, it is almost not possible for any single physician to aggregate all necessary experience to be confident about anything, so IDK is still an option. A practical and simple way to increase diagnostic certainty is to solicit the experience of a diverse group of doctors via discussion. If there is a clear majority for one diagnostic outcome, then the overall confidence in that diagnostics is high. While this voting procedure might be guarantee the optimal outcome, it eliminates the uncertainly during the whole procedure. With this certain procedure, even if the outcome is negative, it can be traced back and accumulate evidence and experience.

Certainty and conditional probability
This certainty is very different from the the conditional probability of the disease given the diagnostic. The first is akin to saying: 95% of the dermatologists would give the same diagnostics. The second defines the probability that, if we had access to ground truth, then 95% of the patients that receive this diagnostics have the corresponding condition.

3. Uncertainty in Machine Learning

One can define "confidence" in machine learning. The definition follows a similar logic to the one used for human diagnosticians in the previous section. The yardstick by which we measure confidence of predicting a label is "how much do alternative labels contradict previous experience?". More formally, we ask how much do we need to change the training data so that it supports an alternative label.

- Bootstrap samples.
- Samples from different hospitals.
- Easy and hard cases.

4. Human decisions and Intelligence augmentation

Computers are an integral part of medical practice. From electronical medical records to medical instrumentation to billing, hospitals and cliniques cannot function without computers. By some measures computers can already make better diagnosis that human doctors. The question is not whether computer diagnostics will become part of medical practice, the question is how.

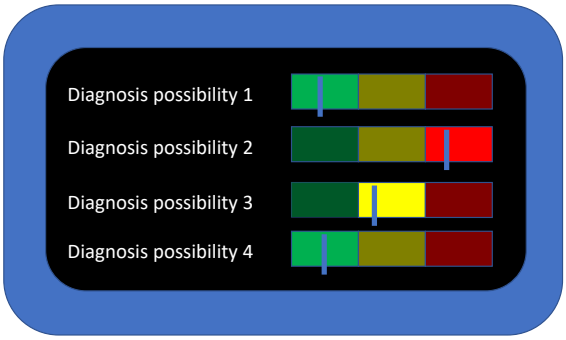
Some claim that human doctors and nurses are heading to extinction, following the fate of manufactoring jobs and bank cashiers. Our prediction is that computers will change the nature of medical work, but that it will increase, rather than decrease, the number of healthcare workers, especially in the care of chronic disease and aging.

We believe computers can perform accurate diagnosis for cases where different doctors are likely to agree. In other cases which are diagnostic gray area the computer will output "I don't know" and transfer the responsibility to the doctor. In most cases, the doctor cannot say "I don't know" because she is responsible for the patients health. On the other hand, resolving the diagnostic question is not her only choice. She can consult another doctor or the literature, ask for additional tests, or decide on a treatment based on available information. Deciding between these options requires much more than diagnostic information. It

involves understanding the patient's emotional, mental and financial state, the patient's support system, the strengths and weaknesses of the hospital in which this is taking place etc.

Over time, computers will be able to take into consideration more and more of this complex information. However, for the forseable future, it is unlikely that computers will be given the responsibility to make medical decisions. Computers will take on much of the diagnostics and alarm tasks, improving the accuracy and timeliness of the doctors actions. Computers will output IDK in gray areas and will leave the decision making to the human doctor. Giving the computer the authority to make decisions currently done by human doctors will deprive the patient the human attention of the doctor.

Some of the digitization of the medicine has come between patients and doctors. The need to record all activity into EMR system require doctors to spend more time at the keyboard, reducing the amount of time of physical examination an discussion. We believe that IA can move medicine in the opposite direction, letting the computer make the common noncontroversial diagnostics and giving the patient more time to interact with the patient.



For IA technology to be widely adopted, the nurses and doctors that use them should experience an improvement in their practice. Suppose that the display of the diagnostics computer uses a three color code for each . Green indicates a confident negative diagnostic, red corresponds to a confident positive diagnosis. Finally, yellow corresponds to IDK, meaning that the computer cannot confirm or reject the diagnostic outcome.

The thresholds which define the three ranges ....

We finish this section with a few application areas which seem ready for applications of IA.

- **Computer aided diagnostics for large-scale data**
  Medical imaging devices such at digital X-ray, CT, EMR and scanning microscope generate many gigabytes of data for each patient. Radiologists and pathologists spend their days analyzing these images to diagnose the patient. The large size and high resolution of the images on the one hand, and the time limitation on the analyst on the other imply that the analyst has to quickly narrow down the suspicious region, increase the chance of missing dangerous abonormalities.

  IA can help the pathologist by suggesting locations in the high resolution image that might contain cancer nodules ().

  directing her attention to the parts of the image that are

- **Adaptive Patient montors**

- **Dissemination of expertise** Computers, trained by experts, can help novices. Serves a function similar to score-cards.

  Teaching young diagnostics

5. Summary

1. Siddhartha Mukherjee. A.i. versus m.d.: What happens when diagnosis is automated? The New Yorker, April 2017.
2. W Ross Ashby. An introduction to cybernetics. 1957.
3. Douglas C Engelbart. Augmenting human intellect: A conceptual framework. Menlo Park, CA, 1962.
4. Andre Esteva, Brett Kuprel, Roberto A Novoa, Justin Ko, Susan M Swetter, Helen M Blau, and Sebastian Thrun. Dermatologist-level classification of skin cancer with deep neural networks. Nature, 542(7639):115–118, 2017.
5. Andrew Ng. What artificial intelligence can and can't do right now. Harvard Business Review, 9, 2016.
6. Eric Topol. Deep medicine: how artificial intelligence can make healthcare human again. Hachette UK, 2019.
7. Md Kafiul Islam, Amir Rastegarnia, and Zhi Yang. Methods for artifact detection and removal from scalp eeg: A review. Neurophysiologie Clinique/Clinical Neurophysiology, 46(4-5):287–305, 2016.
8. Maria Cvach. Monitor alarm fatigue: an integrative review. Biomedical instrumentation & technology, 46(4):268–277, 2012.
9. EXECUTIVE BRIEF. Top 10 health technology hazards for 2020. 2020. URL https://www.ecri.org/landing-top-10-patient-safety-concerns-2020. Top 10 patient safety concerns of 2020 from ECRI Institute.
10. Krisjanis Nesenbergs. Architecture of smart clothing for standardized wearable sensor systems. IEEE Instrumentation & Measurement Magazine, 19(5):36–64, 2016.
11. C. Iber, S. Ancoli-Isreal, A. Chesson Jr., , and S. Quan. The AASM Manual for Scoring of Sleep and Associated Events-Rules: Terminology and Technical Specification. American Academy of Sleep Medicine, 2007.
12. R. B. Berry, D. G. Budhiraja, and et al. Rules for scoring respiratory events in sleep: update of the 2007 AASM manual for the scoring of sleep and associated events. J Clin Sleep Med, 8(5): 597–619, 2012.
13. Robert G Norman, Ivan Pal, Chip Stewart, Joyce A Walsleben, and David M Rapoport. Interobserver agreement among sleep scorers from different centers in a large dataset. Sleep, 23(7):901–908, 2000.
14. Mary L McHugh. Interrater reliability: the kappa statistic. Biochemia medica: Biochemia medica, 22(3):276–282, 2012.
15. Paul Abrams. Describing bladder storage function: overactive bladder syndrome and detrusor overactivity. Urology, 62(5):28–37, 2003.
16. Mika Venhola, Mikko Reunanen, Seppo Taskinen, Tuija Lahdes-Vasama, and Matti Uhari. Interobserver and intra-observer agreement in interpreting urodynamic measurements in children. The Journal of urology, 169(6):2344–2346, 2003.
17. Anne G Dudley, Mark C Adams, John W Brock III, Douglass B Clayton, David B Joseph, Chester J Koh, Paul A Merguerian, John C Pope IV, Jonathan C Routh, John C Thomas, et al. Interrater reliability in interpretation of neuropathic pediatric urodynamic tracings: an expanded multicenter study. The Journal of urology, 199(5):1337–1343, 2018.
18. Maria Brosnan, Andre La Gerche, Saurabh Kumar, Wilson Lo, Jonathan Kalman, and David Prior. Modest agreement in ecg interpretation limits the application of ecg screening in young athletes. Heart Rhythm, 12(1):130–136, 2015.
19. Diana Carolina Moncada, Zulma Vanessa Rueda, Antonio Macías, Tatiana Suárez, Héctor Ortega, and Lázaro Agustín Vélez. Reading and interpretation of chest x-ray in adults with community-acquired pneumonia. The Brazilian Journal of Infectious Diseases, 15(6):540–546, 2011.
20. W Price and II Nicholson. Black-box medicine. Harv. JL & Tech., 28:419, 2014.
21. Roger Allan Ford, W Price, and II Nicholson. Privacy and accountability in black-box medicine. Mich. Telecomm. & Tech. L. Rev., 23:1, 2016.
22. Daniel Kahneman. Thinking, fast and slow. Macmillan, 2011.
23. Jonathan S Vordermark II. An Introduction to Medical Decision-Making: Practical Insights and Approaches. Springer Nature, 2019.