When 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

Introduction

Digital technology is causing a sea-change in all parts of the medical profession. 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".

Other deep learning researchers provide a more nuanced perspective. Sebastian Thrun (1, 2) argues that "... deep learning devices will not replace dermatologists and radiologists. They will *augment* professionals, offering the expertise and assistance".

Artificial Intelligence and Intelligence Augmentation

Using computers to augment human intelligence rather replace it is both tantalizing and 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, think of smartphones that are quickly becoming an inseparable part of our person.

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(3, 4). 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 people's mind.

What would IA look like when applied to medicine? that is the question we aim to answer here. We argue that an important ingredient of the answer is to introduce to AI agents a level of humility. Specifically, to design classifiers, such as DNNs, to say "I don't know".

Labels, ground truth and testing

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 (2). 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.

In their famous work, Esteva et al. set out to show that a classifier trained by machine learning can performs as well as or better than expert dermatologists. In this application of supervised learning each example consists of an input image of a skin patch and an output label that is "benign" or "melignant"

As they wanted to compare the system to human dermatologists they needed a better ground truth than that provided by the dermatologists. 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 depends on the human judgement of the pathologist.

However, even if we assume that pathologists labels are more reliable than dermatology labels, the requirement that each example corresponds to a biopsy introduces a significant bias. Under normal circumstances, patients get biopsied only if the dermatologist thinks there is a chance of malignancy. Therefore, the set of biopsied examples is 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.

Uncertainty in medicine

Medicine is rife with risk and uncertainty. An incorrect diagnosis or treatment can cost the patient his life and the doctor her license.

Uncertainty has many causes, we discuss some of those below.

Patient monitoring and alarm fatigue

A patient monitor is a bedside system equipped with various bio-sensors that record, display and distribute different biometrics, ranging from vital signs such as heart rate, oxygen saturation and blood pressure, to high-frequency waveforms such as electrocardiogram, respiratory signal and arterial blood pressure. Monitors are typically used in hospitals and clinics to closely monitor patients at risk. Most patient monitors come with an alarm system that alerts clinicians life-threating clinical events, such as asystole, ventricular fibrillation, or an intubated patient being disconnected from the ventilator. However, such systems often suffer from a high rate of false alarms, which causes 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 (9, 10).

Yoav : What is the point of this paragraph? It is plausible that utilizing as much data as possible from the patient monitor could drive medical innovation and improve the healthcare. However, in this setup, besides the obvious data quality issue, like noise, the data calibration and validation issues are often less discussed. Due to its proprietary nature, researchers usually cannot calibrate or validate the recorded signals but assume the high data quality. As a result, it has been a long debate if the recorded biosignals are suitable for scientific research (19–22). Without a proper calibration or validation, it is even possible that the more data massively collected without proper calibration and validation, the more biased the developed intelligent system will be.

Yoav : PTT seems to me to be too much in the weeds for this popular article In (23), some delicate artifacts have been reported regarding the pulse transit time (PTT) analysis. **What is PTT and why is it important?** PTT is defined to be the phase latency between the cycles in the electrocardiogram and the photoplethysmogram. It has been shown that PTT contains rich information about the blood pressure (24). It is thus natural to include it to an intelligent system, by learning how it is related to clinical outcomes, to more closely monitor the hemodynamics. However, it was unintentionally found that in *some* patient monitors, the PTT is contaminated by a sawtooth artifact that *might* come from some hardware manufacture procedure. Since such non-physiological artifact is not universal, the usual statistical tools like variable selection cannot help. As a result, the intelligent system might be confused and lead to unpredictable uncertainties.

Yoav : I think the following should be partitioned into two blocks: "Protocols and their limitations" and "Inter-rater agreement and disagreement"

Protocol limitation

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 (25, 26). 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 (27). 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.

Quantification of inter-rater agreement rate

+Cohen's kappa, **Yoav :** Introduce Hoen kappa. Instead of mathematical definition, I would interpret some specific values: the value that corressponds to perfect agreement for positive perfect agreement for negative, the values that corresond to random-level agreement for positive and negative, etc.

Bladder and sphincster diagnosis

Urodynamic studies provide the best bladder and sphincter functional data for urologists to decide how to treat patients at risk for renal damage (28). 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 (28). 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 (29, 30).
Yoav : What are the cohen-kappa numbers

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 (31). In other situation, the variability among subjects is so big that it limits the development of a more quantitative protocol (29). In some situations, when the needed information is missing, it is challenging to make a differential diagnosis (32).

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 (33, 34).

Hautieng : should we jump into GDPR? **Yoav :** what is GDPR?

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. Depending 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.

We consider the management of uncertainty from the medical decision making process point of view(35). Following the "thinking fast, thinking slow" dualism Kahaneman and Tversky, it is generally agreed that two distinct mental processes are involved in choosing a diagnosis. Recognition is a fast, typically non-verbal, mental process in which the doctor identifies a pattern in the symptoms and instinctively makes a diagnosyis. On the other hand, elimination is a slow deliberative process through which the doctor methodically eliminates diagnostic possibilities. To make a decision based on elimination, slow thinking with focused attention is critical (36). This process is like taking a math examination, it takes time and effort, and it is exhaustive. In the every end, depending on the physician experience, he/she might end up with multiple possibilities. He/she could either guess and proceed, or say IDK and consult a higher level experts or discuss with other experts. **Yoav :** I like the last paragraph very much, it makes a lot of sense. I downloaded the vordermark book. I found a lot of good stuff. But I did not find anything about doctors consulting each other, majorities, concensus etc. It would be very relevant to find information both about how decision are made in today's hospital, and how they *should* be made to combine the best of fast and slow thinking. All of this before saying anything about using ML

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.
Yoav : Can you describe a particular interesting / illuminating / convincing case?

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.

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.

Uncertainty vs. accurracy
using ROC curves

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

Agency and Augmentation

Computers are already an integral part of medicine, 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*.

It is not enough to describe the desired end state. In this final section we chart a *migration path* from the current limited role of computers in diagnostics to a more central role. For this to take place, caregivers must benefit from the new technology. Setting the goal to be replacing human doctors with machines is both unrealistic and self-defeating.

Doctors and nurses are humans, they are not diagnostic machines. The personal and emotional connection between doctor and patient is critical for effective treatment. A good doctor combines their medical knowledge with a personal understanding of the patient to choose a treatment plan and discuss it with the patient to get their consent.

It is debatable whether a computer will ever be able to make a meaningful emotional connection with a patient. It is quite clear that such capabilities will not exist in the forseable future. We refer to the ability to connect and to act in a self concious way *agency* and separate it from *intelligence*. We suggest that computers can augment humans in intelligence tasks and leave agency to humans.

Easy and hard diagnoses. As described above, diagnostics is a process of elimination. It starts with a set of possible diagnoses which are gradually eliminated as evidence is gathered. As diagnoses are eliminated, the benefit and risk of different treatment plans is evaluated. Deciding on treatment and continually monitoring it is where the doctor’s agency is most important.

Our suggestion is that machine learning helps the doctor eliminate diagnoses and evaluate treatment options, while leaving the the decisions to the doctor.

Diagnosis vary in their difficulty. Consider a sequence of patients visiting a clinic. Suppose the clinic has four doctors, two trainees, and six nurses. Each patient first gets seen by a nurse, then by a doctor, and possibly by a trainee. Some of patients have a simple diagnosis, one that all twelve members of the clinic will state with confidence. Other patients have more complex diagnosis that requires the attention of a doctor. Finally, some diagnosis are so complex that the doctor needs to consult other doctors and possibly ask for additional tests. In other words, the clinic acts as an ensemble of classifiers. Easy cases result in a unanimous diagnosis, harder ones results in a clear majority, and the very hardest results in disagreement which requires consultation and additional tests.

The machine learning ensemble of classifiers follows a similar logic.

- **Computer aided diagnostics** Especially with very large data: ecg for 14 says....

  Pathology.

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

  Teaching young diagnostics

- **Confidence, Trust and adoption of technology**

Summary

1. Siddhartha Mukherjee. A.i. versus m.d.: What happens when diagnosis is automated? *The New Yorker*, April 2017.
2. 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.
3. W Ross Ashby. An introduction to cybernetics. 1957.
4. Douglas C Engelbart. Augmenting human intellect: A conceptual framework. *Menlo Park, CA*, 1962.
5. Sangeeta Mehta, John Granton, Stephen E Lapinsky, Gary Newton, Kristofer Bandayrel, Anjuli Little, Chuin Siau, Deborah J Cook, Dieter Ayers, Joel Singer, et al. Agreement in electrocardiogram interpretation in patients with septic shock. *Critical care medicine*, 39 (9):2080–2086, 2011.
6. Monika Atiya, Tobias Kurth, Klaus Berger, Julie E Buring, and Carlos S Kase. Interobserver agreement in the classification of stroke in the women’s health study. *Stroke*, 34(2): 565–567, 2003.
7. Alistair E W Johnson, Mohammad M. Ghassemi, Shamim Nemati, Katherine E. Niehaus, David Clifton, and Gari D. Clifford. Machine Learning and Decision Support in Critical Care. *Proceedings of the IEEE*, 104(2):444–466, 2016.
8. William Fleischman, Bethany Ciliberto, Nicole Rozanski, Vivek Parwani, and Steven L Bernstein. Emergency department monitor alarms rarely change clinical management: An observational study. *The American journal of emergency medicine*, 2019.
9. Maria Cvach. Monitor alarm fatigue: an integrative review. *Biomedical instrumentation & technology*, 46(4):268–277, 2012.
10. 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.
11. Joachim Behar, Julien Oster, Qiao Li, and Gari D Clifford. Ecg signal quality during arrhythmia and its application to false alarm reduction. *IEEE transactions on biomedical engineering*, 60(6):1660–1666, 2013. alarm fatigue prevention for arrhythmia.
12. Yong Bai, Duc Do, Quan Ding, Jorge Arroyo Palacios, Yalda Shahriari, Michele M Pelter, Noel Boyle, Richard Fidler, and Xiao Hu. Is the sequence of superalarm triggers more predictive than sequence of the currently utilized patient monitor alarms? *IEEE Transactions on Biomedical Engineering*, 64(5):1023–1032, 2016. a new alarm system called superalarm to avoid alarm fatigue.
13. Wei Zong, Larry Nielsen, Brian Gross, Juan Brea, and Joseph Frassica. A practical algorithm to reduce false critical ecg alarms using arterial blood pressure and/or photoplethysmogram waveforms. *Physiological measurement*, 37(8):1355, 2016.
14. Bradford D Winters, Maria M Cvach, Christopher P Bonafide, Xiao Hu, Avinash Konkani, Michael F O’Connor, Jeffrey M Rothschild, Nicholas M Selby, Michele M Pelter, Barbara McLean, et al. Technological distractions (part 2): a summary of approaches to manage clinical alarms with intent to reduce alarm fatigue. *Critical care medicine*, 46(1):130–137, 2018.
15. Xiao Hu. An algorithm strategy for precise patient monitoring in a connected healthcare enterprise. *NPJ digital medicine*, 2(1):1–5, 2019.
16. Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018. sepsis treatment via AI.

17. Vijay Huddar, Bapu Koundinya Desiraju, Vaibhav Rajan, Sakyajit Bhattacharya, Shourya Roy, and Chandan K Reddy. Predicting complications in critical care using heterogeneous clinical data. *IEEE Access*, 4:7988–8001, 2016. ML for complication prediction in ICU.
18. Alexander Meyer, Dina Zverinski, Boris Pfahringer, Jörg Kempfert, Titus Kuehne, Simon H Sündermann, Christof Stamm, Thomas Hofmann, Volkmar Falk, and Carsten Eickhoff. Machine learning for real-time prediction of complications in critical care: a retrospective study. *The Lancet Respiratory Medicine*, 6(12):905–914, 2018. ML for complication prediction in ICU.
19. Keith J. Ruskin and Kirk H. Shelley. Patent medicine and the "black box". *Anesthesia and Analgesia*, 100(5):1361–1362, 2005. ISSN 00032999.
20. Jeffrey M. Feldman. Can clinical monitors be used as scientific instruments? *Anesthesia and Analgesia*, 103(5):1071–1072, 2006.
21. Kirk H. Shelley and Steven J. Barker. Disclosures, what is necessary and sufficient? *Anesthesia and Analgesia*, 122(2):307–308, 2016.
22. Maxime Cannesson and Steven L. Shafer. All boxes are black. *Anesthesia and Analgesia*, 122(2):309–317, 2016.
23. Yu-Ting Lin, Yu-Lun Lo, Chen-Yun Lin, Martin G Frasch, and Hau-Tieng Wu. Unexpected sawtooth artifact in beat-to-beat pulse transit time measured from patient monitor data. *PloS one*, 14(9), 2019.
24. Heiko Gesche, Detlef Grosskurth, Gert Küchler, and Andreas Patzak. Continuous blood pressure measurement by using the pulse transit time: comparison to a cuff-based method. *European journal of applied physiology*, 112(1):309–315, 2012.
25. 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.
26. 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.
27. 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.
28. Paul Abrams. Describing bladder storage function: overactive bladder syndrome and detrusor overactivity. *Urology*, 62(5):28–37, 2003.
29. 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.
30. 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.
31. 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.
32. 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.
33. W Price and II Nicholson. Black-box medicine. *Harv. JL & Tech.*, 28:419, 2014.
34. Roger Allan Ford, W Price, and II Nicholson. Privacy and accountability in black-box medicine. *Mich. Telecomm. & Tech. L. Rev.*, 23:1, 2016.
35. Jonathan S Vordermark II. *An Introduction to Medical Decision-Making: Practical Insights and Approaches*. Springer Nature, 2019.
36. Jeffrey B Michel. Thinking fast and slow in medicine. In *Baylor University Medical Center Proceedings*, volume 33, pages 123–125. Taylor & Francis, 2020.