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

June 24, 2020

Abstract

The meteoric rise of AI in general and Deep Learning in particular is generating great excitement throughout academia and commerce, and in particular in medicine[**?**, **?**]. With some some high-profile claims [] that AI will soon replace humans in many medical specialties.

In this position paper we present an alternative view. We contrast *Artificial Intelligence* with *Intelligence Augmentation* and argue that the second is more likely to benefit the patient than the first. We provide evidence to this argument and present a vision in which easier decisions are delegated to computers, while the more difficult ones are handled by humans.

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 [**?**]. 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 [**?**, **?**] argues that "... deep learning devices will not replace dermatologists and radiologists. They will *augment* professionals, offering the expertise and assistance".

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".

AI vs IA

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[**?**, **?**]. 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.

Quantification of inter-rater agreement rate	**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**

Urodynamic studies provide the best bladder and sphincter functional data for urologists to decide how to treat patients at risk for renal damage [?]. While it has been extensively studied and applied in clinics, the main issue that plagues

In general, diagnosis is performed by comparing observed symptoms to past experience. Roughly speaking there are two ways to make this comparison: *recognition* and *elimination*. Recognition is a fast, typically non-

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 de-

Medical de- vices use a va- ri- ety of bio- sensors that record, dis- play and dis- tribute dif- fer- ent bio- met- rics, rang- ing from vi- tal signs such as heart rate, oxy- gen sat- u- ra- tion and blood pres- sure, to high- frequency wave- forms such as ECG, EEG, res- pi- ra-

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 biosig-

Certainty and conditional probability**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**

Uncertainty vs. accurracyUncertainty vs. accurracy using ROC curves

The scarcity of ground truth

Supervised machine learning is a voracious consumer of **labeled data** ([**?**] page XXX). Unfortunately, in most medical applications of machine learning, labeled data, also called "ground truth" is not available.

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

Skin cancer diagnosis using Deep Neural Network

One of the papers that provided evidence that deep neural networks might be able to outperform humans is the work of Esteva et al [**?**]. They trained a Deep neural network to classify images of skin into three categories: benign, malignant and non-

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 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.

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 [**?**]. In other situation, the variability among subjects is so big that it limits the development of a more quantitative protocol [**?**]. In some situations, when the needed information is missing, it is challenging to make a differential diagnosis [**?**].

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 [**?**, **?**].

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.

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.

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.

Human decisions and Intelligence augmentation

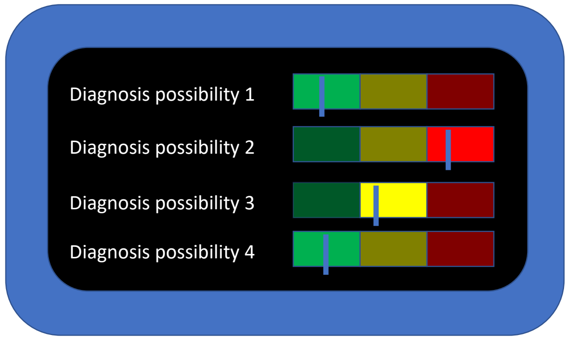
Computers are an integral part of medical practice. From electronical medical records to medical instrumentation to billing, hospitals and cliniques cannot function without computers. By some measures computers can already make better diagnosis that human doctors. The question is not *whether* computer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

Summary