

## The Unintended Consequences of Chasing Electric Zebras

Rob is an excellent and expert radiologist. He is a physician who can use a wide set of advanced techniques in medical imaging like X-rays, CT, and MRI examinations to diagnose patients with many types of illness. Rob is in his late forties and many believe he will soon become the head of the radiology department of the hospital: when he makes a point in a meeting, few would contradict him. Many colleagues know that they can rely on him for an informed second opinion in case they have to cope with a difficult case. Some even believe that he is the most talented and knowledgeable radiologist of the hospital. For this job, Rob earns a quite good salary: last year this was \$380,000, which yet is not extraordinary for radiologists with his expertise; according to the U.S. Bureau of Labor Statistics, that amount is just the median salary for physicians in his position. That notwithstanding Rob does not work to get rich. His passion for his job is genuine. To become such a radiologist, the competition for Rob had been long and fierce: after the high school, he had to get accepted for a 4-year university and then obtain a bachelor's degree with excellent medical school grades and very high university scores; then he had to collect a number of enthusiastic letters of recommendation on his attitude and competences to enter a medical school as a resident, where he practiced general medicine and surgery for one year before applying for a further 4-year training program in the field. During this latter program, Rob had to interpret, both day and night, tens of thousands of imaging studies, while also counseling patients and communicating and discussing results with his referring colleagues. Over time, his skills became clear and widely recognized at work.

Specialists like neurologists and orthopedists who refer patient to radiologists for digital imaging examinations usually disregard the radiologists'



Figure 1. A zebra melting into a bar code. Artistic picture by Nevit Dilmen (2013, CC 0).

report when they consider the resulting images: they also know how to read images, do not want to be influenced in their decisions, and above all, do not want to reduce the analogical richness of an image to the discrete rigidity of verbal labels and ordinal categories reported in the radiological report. However, when the specialists of the hospital of Rob see that a report has been signed by him, they usually read it carefully and consider it trustfully.

Probably also for his reputation, one day Rob was involved in a reliability study in regard to the early diagnosis of Tuberculous Meningitis (TBM). This is a serious disease that usually kills two patients out of three and leaves one patient out of two with serious neurological consequences. In this study Rob, and other nine colleagues from several other hospitals, were supposed to assess a number of CT scans to assess and classify the entity of Basal Meningeal Enhancement (BME), that is a classic neuroradiologic features of TBM<sup>1</sup>. The results of this study were soon published and two particular aspects disquieted Rob. The intra-rater agreement (measured by means of the Fleiss' kappa statistic) assessed the extent Rob "agreed with himself" on the same CT scan being shown to him twice at different times during the day and in a random order with other control images: it was fair but not too high (0.55), i.e., he agreed with himself slightly more than 2 times out of three. However, the inter-

---

<sup>1</sup> This situation is inspired by the study reported in: Botha, H., Ackerman, C., Candy, S., Carr, J. A., Griffith-Richards, S., & Bateman, K. J. (2012). Reliability and

diagnostic performance of CT imaging criteria in the diagnosis of tuberculous meningitis. *PloS one*, 7(6), e38982.

rater agreement, that is the extent his findings were confirmed by the other colleagues involved in the study, was even lower (0.35), just half of the times. This result was not a thrashing of the radiologists' skills and knowledge. All the contrary, their diagnosis was still considered accurate as a whole, but showing some important variability: this could be traced back to a number of factors, like the reporting surgeon (and probably Rob was among the best raters), the device used to display the images, the radiologist's experience, and even the work shift (being agreement lower at the end of intense work shifts or by night). The authors of the study just commented on this point that the criteria for BME assessment needed to be standardized and validated in more thoroughly prospective cohort studies and that this kind of assessment was worthy of further study.

Some years later, the management of the hospital where Rob worked was contacted by an important IT company that proposed to provide the radiology department with a novel decision support system, called "Zebra Hunter", a state-of-the-art machine learning system, in exchange of the expertise necessary to optimize it and the availability for hosting a series of experiments on its accuracy and reliability. However, the Zeb (how it was dearly called by its designers) was already very good in detecting anomalies and relevant traits of digital imaging: its accuracy for TBM cases, for instance, was 90% by means of a complex convolutional neural network. The management understood that this partnership could bring positive publicity and accepted to adopt this system in the idea that it could only improve the performance and accuracy of the radiological department. Moreover, it was decided that the Zeb should be used by radiologists only (and not, for instance by the neurologists): expert radiologists, like Rob, had to validate any of the Zeb's suggestions and decide whether to integrate them in their reports or not: these reports should not indicate whether the diagnosis was performed with the aid of the Zeb or without it. In other words, nothing should look different at the terminals of the department, although patients and physicians knew that the radiologists had an extra

oomph for their daily job. Indeed, Zeb was used extensively and IT designers and radiologists worked intensively and closely together since its first deployment: its performance improved very soon, and its accuracy reached an impressive 98% on many types of images, while for others it was nevertheless above 85%: glowing figures that tied the best performances by the humans and that were totally irrespective of the work load, occasional resource shortages, the more or less frantic pace of hospital work. And mood swings.

The vignette can stop here. It already inspires us a number of questions. Some high-level questions are often addressed in both the Academic and white literature. One question is whether Zeb and systems like it will be used (and its bills be paid) more to increase hospital efficiency, prolong continuity of care, reduce health system costs, or improve care quality and outcomes. Moreover: are all these objectives positively correlated, or rather incompatible? Analysts also wonder if the almost sudden availability in the market of a number of "super Rob" (or at least, "tireless Rob") would cause a reduction of the salaries of expert radiologists like the real Rob, as well as of the demand by hospitals and hence the society for this medical role. As in the vignette above, we could think of the advent of a new medical role, probably extending the traditional skills of the radiologists, that can establish and maintain a trustworthy interface between the intelligent machine, kept as a sort of solitary oracle, and the rest of the socio-technical environment, that is anyone wanting to get access to the predictive services. This new radiologist would be someone who could properly feed in the machine and then interpret and make sense of its output and still filter and process it to make this "collaborative" advice be meaningful for the other physicians, the patients and all of the care givers. After all, approximately one century ago, this is exactly how the professional role of the radiologist was established and empowered: as an interface between the new X-ray equipment and the rest of the hospital physicians asking for the new type of

consultation<sup>2</sup>. Moreover, exactly 40 years ago, Maxmen, an influential psychiatrist and professor at Columbia University, foresaw and advocated that within 2025 doctors would be substituted by a new “Medic-Machine symbiosis”<sup>3</sup> for the better provision of healthcare services.

However, we are more interested in subtler and more eerie questions: for instance, we wonder whether, and how, Zeb will affect the decisions of the doctors who use it, including Rob. Maybe Zeb would help novices learn how to interpret both easy and difficult images; improve the residents’ skills more quickly also by challenging them with tricky simulations based on real cases; and teach also expert radiologists how to solve cases that before its arrival would be too difficult for a fast analysis and require several meetings, or worse yet, further examinations for the patient. Conversely, Zeb would perhaps undermine the self-confidence of the expert radiologists like Rob, after a few times that their interpretation differ from Zeb’s and this latter proved to be the correct one. Likewise, it could make the young or less brilliant radiologists more lazy and dependent on its recommendations. Thus the main point here is whether systems like Zeb have the potential to actually deskill or “spoil” the physicians in the long run. After all, Rob and colleagues know that they can make mistakes, especially under time pressure or after 10 hours of work shift, while Zeb, also thanks to their daily feedback, has become almost 100% accurate on most kinds of imaging. Unfortunately, the radiologists know that Zeb will misclassify (either in terms of false negatives, or false positives) almost one case every day: they cannot let down their guard. Zeb is not infallible, but nevertheless its opinion must be held in very high esteem, just like Rob’s opinion was, and still is, when he is at his best. More subtly, we also wonder if doctors will be allowed not to follow Zeb’s advice. We are not speaking of a legal limitation, as accountability and

responsibility will be at the human side for a long time still. Rather we are wondering whether ignoring Zeb will be socially or professionally blameworthy. Being against Zeb could seem a sign of obstinacy, arrogance, or presumption: after all that machine is right almost 98 times out of 100 for many pathologies, and no radiologist could seriously think to perform better than this. Probably novices and young radiologists would restrain themselves from defending their diagnostic hypothesis if different from the Zeb’s one; but what about the seniors like Rob? Will using Zeb, and any similar system, either increase or decrease the number of “zebras”<sup>4</sup> that doctors will pursue whenever they hear hoofbeats? After all, Zeb will mention many alternative explanations all together with its best recommendation, even those that no radiologist would ever think of. Furthermore, how will this kind of decision support change the “political” status of radiologists like Rob towards the hospital management and other heads; the reputation-based hierarchical relationships among radiologists; the trust relationships between the radiologists and the other specialists; the tensions and collaborative dynamics between the IT staff and the physicians? Lastly, can we exclude that doctors will also use systems like Zeb to practice a stronger and more surreptitious defensive medicine, that is to choose for the most plausible option that defends them against potential controversies (whereas plausibility is estimated by the machine), or worse yet, to make Zeb a scapegoat to indulge in excusing both personal and teamwork failures (i.e., those related to collaboration and communication failures)?

We acknowledge and purport the open nature of these questions, as well as of any related question that any ethically engaged designer of medical technology should address. These questions all regard the potential unintended (that is

---

<sup>2</sup> Reiser, S. J. (1981). *Medicine and the Reign of Technology*. Cambridge University Press.

<sup>3</sup> Maxmen, J. S. (1976). *The post-physician era: Medicine in the twenty-first century*.

<sup>4</sup> Zebras in the medical jargon are the anomalies, the odd signs, the diseases so rare that most doctors never encounter them in regular medical practice. This term comes after the aphorism “When you hear hoofbeats, think of horses not zebras”.

unexpected) consequences<sup>5</sup> of using powerful machine learning techniques in support of medical practice, especially for prognosis, diagnosis and treatment choice. Among these unintended consequences, we focus on the concept of *overreliance*. Overreliance on these new decision support systems (DSS) can be further distinguished in two kinds of more specific consequences: *overdependence* and *overconfidence*. The former one occurs when habitual users of these systems either forget, ignore or even stop conceiving any safety net, plan B, or contingency plan, that is any alternative system that could substitute the automated one when this fails, is interrupted or breaks down. Overdependence then relates to a lack of real autonomy by the human actor and also technology abuse, that is the use of the system beyond actual needs. On the other hand, overconfidence relates to three ways of thinking: thinking that the DSS will never fail; thinking that it will never harm; and thinking that it will never be wrong. Overdependence and overreliance have been initially discussed by Parasuraman, who speaks of automation-related complacency and automation bias<sup>6</sup>, respectively to denote the same phenomena from the human factor perspective. In particular, complacency is defined as “trusting automation to fulfill the function for which it was designed”; on the other hand, automation bias results in making both omission and commission errors when decision aids are imperfect (and to some extent they are *all* imperfect).

To this preliminary distinction we add the concept of *semiotic desensitization*. This is the progressive decrease of responsiveness and sensitivity of physicians with respect to material, bodily signs of

the patient who stands in front of them, in favor of the data proposed by electronic patient records, registries and decision support tools. These data are the digitized, quantified (or just categorized) counterparts of the patient’s signs, which are produced by any kind of automated probe, sensor and device. Therefore, semiotic desensitization is a consequence of the “quantified patient”<sup>7</sup>. This is just the necessary amount of data that “chasing electric zebras” requires to make machine learning tools accurate and reliable; that is to make systems like Zeb the optimal *intermediary* between the actual body of the patient who turns to doctors for help and care, and the required decisions that doctors have to make on how to intervene on that body to solve the patient’s problems.

In our research we are trying to understand this potential high-impact consequence of the datafication of healthcare and to find some effective antidote. In particular, we are investigating if visualization tools<sup>8</sup> that provide doctors with alternative, visual, number-less (i.e., analogical) representations of the patient’s conditions could support their decision making and also get them more wary of the validity of clear-cut categories and quantitative numbers. In so doing, we also aim to reinforce the idea that medical practice is primarily still an “art and science of the signs”. Exactly two hundred years ago, Landré-Beauvais<sup>9</sup> conceived and proposed to his colleagues this vision of what medicine is. Now we are wondering if it is important to preserve such a vision, in this age in which machines learn how to *datafy* medical signs better and faster than any brilliant student, as Rob was, actually can learn to interpret them.

---

<sup>5</sup> Ash, J. S., Berg, M., & Coiera, E. (2004). Some unintended consequences of information technology in health care: the nature of patient care information system-related errors. *Journal of the American Medical Informatics Association*, 11(2), 104-112.

<sup>6</sup> Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 52(3), 381-410.

<sup>7</sup> Smith, G. J., & Vonthethoff, B. (2016). Health by numbers? Exploring the practice and experience of datafied health. *Health Sociology Review*, 1-16.

<sup>8</sup> Cabitza, F., Locoro, A., Fogli, D., & Giacomini, M. (2016). Valuable Visualization of Healthcare Information: From the Quantified Self Data to Conversations. In *Proceedings of the International Working Conference on Advanced Visual Interfaces* (pp. 376-380). ACM.

<sup>9</sup> Landré-Beauvais AJ (1818). *Séméiotique, ou traité des signes des maladies*. Paris: J.A. Brosson