English version of the paper "Kunstig Intelligens og Medicinsk Etik: Tilfældet *Watson for Oncology*", to be printed in the volume *8 Cases i Medicinsk Etik* (eds. Ezio Di Nuzzi, Rasmus Thybo Jensen and Jeannette Bresson Ladegaard Knox, Munksgaard. Expected publication: 2020).

Ethics of medical AI: the case of Watson for Oncology

Ezio Di Nucci, Rasmus Thybo Jensen & Aaro Tupasela

Introduction

No doubt, medical students and professionals are often motivated by the prospect of helping people in need. But let's be honest, a big motivator for studying medicine is its job prospects: namely plenty of well-paid safe jobs. That is why medical artificial intelligence (medical AI) should scare you: because it is coming after your jobs.

In this chapter we will discuss *IBM Watson for Oncology* (from now on just *Watson* for short) as a case study in the emergence of medical AI. We will analyse some of the most interesting ethical and philosophical questions raised by medical AI in general and Watson in particular.

Watson is "a decision-support system that ranks cancer therapeutic options" (Di Nucci 2019: 1) based on machine learning algorithms, which are computer systems that are, according to cognitive scientists, able to "figure it out on their own, by making inferences from data" (Domingos 2015: xi).

So you can double down on your fear already, dear medics: those machines are coming after your jobs and they are also coming after the jobs of their own programmers – that's how greedy they are. They clearly won't stop until they have taken over the whole world, which is in fact what technophobes and their extremist friends, the *techno-apocalypsts*, are afraid of (see for example Tegmark 2017, especially the first chapter).[1]

How does Watson work? Based on its access to up-to-date medical research publications and patient's health records, Watson's algorithm – developed by IBM engineers together with oncologists from the *Memorial Sloan Kettering (MSK) Cancer Center* in New York - generates cancer treatment recommendations that oncologists can review and use in consultation with patients.

Here some more details on how Watson functions, from Rosalind McDougall's recent analysis in the *Journal of Medical Ethics*:

> Watson for Oncology is designed and marketed as a tool for clinicians, to assist them to 'zero in on the most promising care options' in an age where the available literature on cancer is huge and fast-moving. The system extracts clinical information from the patient's medical record, such as gender, age, stage and type of cancer, family history, notes from previous visits, test results and comorbidities. The doctor is prompted to verify this extracted information and add additional relevant information. The information is then analysed, based on the computer's training by oncologists at Memorial Sloan Kettering Cancer Center in New York. The system accesses over 300 medical journals and over 200 textbooks. According to IBM's marketing video, Watson 'identifies a prioritized list of treatment options based on Memorial Sloan Kettering expertise and training, and provides links to supporting evidence'. Watson presents a ranked list of treatment options and a synthesis of the existing published evidence relevant to that clinical situation. The treatment options are divided into three colour-coded sections: green for recommended treatments, amber for treatments to consider and red for treatments that are not recommended. There may be multiple options in each section. The options are ranked based on outcome statistics presented in terms of 'disease-free survival'. For each treatment option, there are two literature tabs available to the clinician. One tab gives links to literature that supports that treatment option, identified by Memorial Sloan Kettering clinicians. The other tab gives links to Watson-identified literature relevant to that clinical situation. Information on toxicities associated with the treatment (such as vomiting, anaemia, diarrhoea and so on) is also available to clinicians. The system also has the capacity to be customised to the specific

[1] There are also, it must be said, more optimistic voices, such as Eric Topol's *Deep Medicine*.

> geographic context where it is being used. For example, local clinical guidelines and availability of drugs are included for clinicians (McDougall 2019: 157).[2]

Are you scared now? The kind of medical advice and recommendations that patients used to receive from oncologists would seem to have been delegated to computer algorithms – and intelligent ones at that, who can program themselves (if you believe those cognitive scientists). On a superficial reading, everybody should be scared: patients because their treatment is being decided by computers and doctors because their jobs are being outsourced away.

Automation (Wajcman 2017, Susskind & Susskind 2015) will have come full-circle if even the most privileged and better paid jobs such as those in healthcare and law are lost to machine learning algorithms.[3]

What's the big deal, really? Delegation and Performance

We are going to suggest that the above argument is too quick: that there are indeed serious ethical and philosophical issues to consider, but that first we need to properly analyse Watson's tasks in order to be able to evaluate such innovation from an ethical point of view.

What is it – exactly – that we are delegating to Watson? There is a big difference, for example, between delegating decision making to a computer system and delegating advice or – as it is said within this debate – 'decision-support' to the algorithm. And in fact given the details above about

[2] You can see IBM's marketing material here: https://www.youtube.com/watch?v=8_bi-S0XNPI (accessed on 4.6.19). Some more details on Watson are available at the following links to recent articles in popular scientific publications:
http://fortune.com/2018/07/27/ibm-watson-cancer/
https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/
https://www.engadget.com/2017/06/01/ibm-watson-cancer-treatment-plans/?guccounter=1

[3] And in fact you will be relieved to know that it's not just medics, lawyers are under threat too (Fry 2018).

the functioning of Watson, it would look like that what we are delegating is decision-support instead of decision-making. As in, Watson comes in to support the clinician's decision-making rather than to do its own decision-making instead of, or on behalf of, the oncologist. At least in theory, then, the decision-making capacity and, crucially, authority still rests with the clinician, together with patients. However, it is worth noting that Watson is in fact designed so as to come up with solutions to the task that it is said to support humans in solving. The way Watson is to support the human experts' decision about which treatments to recommend is in part by itself coming up with a prioritized list of treatment recommendations. This means that, at least in theory, there is continuum of usages of Watson starting with letting Watson having very little say in the human decision making and ending with giving Watson full control over the decision about which treatment to recommend the patient. In practice, these possible usages mean that we are constantly faced with the question of how far we are willing to go in our delegation of tasks to Watson.

The simple theoretical consideration above – with its distinction between decision-making and decision-support and its application to Watson's case – shows that our analysis needs to be structured in ways that are a bit more sophisticated than the superficial technophobic skepticism with which we started.

Let us try breaking our analysis down into the following questions:

1) What could be the potential reasons for delegating respectively decision-support and decision-making to Watson?
2) How good is Watson's performance and how can we measure the performance level of Watson both in absolute terms and in comparison to human clinicians?
3) Are there risks in using Watson that are independent of problems concerning how to measure Watson's performance level?

4) Finally, which tasks are we – ethically – allowed to delegate to Watson and which ought we, again from an ethical point of view, to keep for humans?[4]

This last question #4 is the overall question that we need to answer in providing an ethical analysis of our Watson case study; and in order to do that, it will be necessary to first address the other three questions – and some further ones along the way. This is what we turn to now, starting from Question #1.

Question #1: delegating what, why and to whom?

As question #1 suggests, control is a big part of the ethical equation when it comes to Watson: healthcare is too precious – especially when it comes to life and death decision-making such as in the context of oncology – to lose control to algorithms such as Watson. How much control we should want – in every context – depends on how much we care and should care. And we do care and should care a lot about cancer, both because it kills people and because progress within cancer research can save many lives – recent statistics say that every second person alive today will experience cancer within their lifetime but that – at the same time – half of those will survive it.[5] So it matters. And therefore we don't want to lose (human) control.[6] Applied to our case, this reasoning should make us reluctant to put machines in control of the actual decision making concerning which treatment we should recommend to a particular patient.

[4] Here note the classic ethical distinction between being morally allowed to do something and being morally obliged to do something. See Di Nucci (2018, Chapter 3) for more on this distinction.

[5] Here some more precise numbers from Cancer UK, for example:
https://www.cancerresearchuk.org/health-professional/cancer-statistics/risk/lifetime-risk

[6] See Santoni & Van der Hoven (2018) for a philosophical analysis of the notion of "meaningful human control". See Grote & Di Nucci (forthcoming) for more on the importance of the issue of control in debates about algorithmic decision-making.

One possible objection to this conclusion is that we shouldn't a priori rule out any kind of delegation of control to an AI-system, since the loss of control might be outweighed by the fact that the AI-system compared with the best human experts improves the performance of the relevant tasks significantly. This is a fair objection but as formulated here it is in need of clarification.

We need to take a closer look at what improved performance can even mean in the particular context of Watson, because there are at least two different ways in which such technological systems can provide enhanced performance.

On the one hand, the system to which we delegate a task that was previously performed by a human can do a better job than the human (ever, maybe?) could. This is what we call *strategic delegation*, where we delegate to someone or something else because they are supposed to be better; they are – if you like – the experts. Strategic delegation happens not just between humans and technological systems but especially between different human agents, such as when we delegate to the plumber repairing piping in our kitchen.

In the case of Watson, we find IBM stressing a specifically strategic reason for using Watson when they in their promotion material write as follows: "With the ability to read over 200 million pages in three seconds, IBM Watson® for Oncology is learning the language of oncology and detecting patterns that humans may not have seen."[7] Here it is suggested that Watson has an ability to detect patterns relevant for treatment that exceeds that of human experts and this is one way in which the theoretical idea of strategic delegation could apply to Watson.

There is an alternative to strategic delegation, which we refer to as *economic delegation*: that's when we delegate some task to someone else or something else not because we believe that they will be better at it than ourselves (or anyway the previous person or system tasked with it) but

[7] https://www.ibm.com/downloads/cas/AONROW12

because we believe that it is more resources-efficient to delegate the task; cleaning is a straightforward example of this, where a lot of economically privileged people do not necessarily delegate cleaning because they believe that they wouldn't be as capable of doing it (even though they wouldn't, actually) but because they, in short, can't be bothered – which is a not very nice way of saying that they believe that it is more resources-efficient for them to delegate cleaning to someone else. Generally speaking, having a strategic reason for delegating a certain task does not necessarily depend on enhancing the quality of the performance of that specific task but on whether the delegation will save a given system (e.g. the health care system) resources, in a way that allows an optimization of the system's overall performance.

How does economic delegation apply to Watson? Instead of thinking that Watson can do things better than human oncologists can, we could alternatively argue that human oncologists – as stressed and pressed for time as they are - can be freed up for more meaningful tasks by having Watson scan medical literature. So that the expensive and limited time of human oncologists can – through the use of Watson – be deployed more efficiently – and then doctors might end up having more time for patients, for example. It is such an economic reason that IBM provides when they say as follows: "Recent studies continue to demonstrate that Watson for Oncology 'agrees' with physicians around the world in the vast majority of cases – so experts can focus on what they do best— deliver care" (IBM 2018).[8]

[8] It is worth noticing that when IBM says that the use of Watson allows the experts to focus on delivering care, they seem to imply that a reaonable usage of Watson may go beyond delegation of mere decision-support. However, as we shall see below, it is a good question whether studies of congruence-rates with human experts are at all apt to be used in support of delegation of decision-making to Watson.

The idea is that, even if Watson performed just as well as human doctors rather than better – which would mean that there would be no argument in favour of using Watson from the point of view of strategic delegation – there might be other independent arguments in favour of using such technological systems coming from economic delegation.

Once we have introduced and understood the distinction between strategic delegation and economic delegation, we can then also see that performance is crucial but is also not the whole story – there is obviously a context which contributes to performance evaluation. And it is in fact the familiar distributive context of efficiencies within limited resources. That's one thing that technology does not change, even though it might alleviate the healthcare burden – at least if we delegate responsibly.

In addressing question #1 we have learned about at least two important distinctions which are relevant to the ethical evaluation of our case study: the distinction between decision-making and decision-support which speaks to which precise tasks are being delegated and the distinction between strategic delegation and economic delegation which speaks to the different reasons for delegation. We can also begin to see how there might be an internal connection between our two distinctions. We said above that cancer treatment is something we care and should care deeply about, and that therefore we should not hand over control in the form of regular decision-making capacities to technologies like Watson. And we raised the possible objection that we shouldn't *a priori* rule out any possible delegation of tasks, since whether we should delegate or not ought to depend on whether the machine outperforms the human experts. But what kind of outperforming is relevant here? Is it the one connected to strategic reason or the one connected to economic reasons? The ultimate reason why we care about which cancer treatment we recommend to the patient is that we want the patient to receive the best possible recommendation, and not that we want to provide recommendations in a way that saves resources. If, for instance, we could know that Watson was significantly better at estimating the patient outcomes of different treatment, then this might speak in favor of delegating more control to Watson.[9] This tells us that if are we looking for reasons that could tell in favor of delegating even the decision competence to an AI-system like Watson, then we should be looking for strategic reason.

These tools in hand, we are now ready to address the crucial question #2 about Watson's performance.

Question #2: How do we measure Watson's performance? Problems of Concordance

As we have just seen different levels of performance becomes relevant dependent on whether one is arguing for delegation to Watson for strategic or for economic reason. For us to have strategic reasons for delegating we must have reason to think Watson outperforms human experts, whereas economic reasons only require that Watson lives up to the standard of the best human experts or at least approximates that standard. Often arguments for the delegation of a task to a certain technology present a mix of the two kinds of reasons, it is, however, important to keep in mind that the two kinds of reasons can function independently of one another. We may have purely strategic reasons to implement a health technology in cases where the technology will raise the quality of the treatment but without any economic benefit. And we may have purely economic reasons in cases where the technology will save us precious resources but without improving the quality of the treatment. It is paramount that we do not confuse these two kinds of reasons, because this could

[9] Many thanks to Isaac Wagner for alerting us to the kind of objection discussed here and for raising this particular version of it.

mislead us into thinking that we have better a reason for delegating a task to a given technology than we in fact have.

So what do we know about the performance level of Watson? Here are some results from two of the first attempts to measure the performance-level of Watson, both of which measured how concordant Watson's recommendations are with the recommendations of a specific group of oncology experts:

> Overall, treatment recommendations were concordant in 96.4% of lung, 81.0% of colon and 92.7% of rectal cancer cases. By tumour stage, treatment recommendations were concordant in 88.9% of localized and 97.9% of metastatic lung cancer, 85.5% of localized and 76.6% of metastatic colon cancer, and 96.8% of localized and 80.6% of metastatic rectal cancer (Somashekhar et al 2017).

> The overall concordance rate was 83%; 89% for colorectal, 91% for lung, 76% for breast, and 78% for gastric cancer. Similar concordance rates were observed when retrospective and prospective cases were analyzed separately. Discordance was attributable in part to local oncologists' preferences for non-U.S. guidelines for certain cancers, especially gastric cancer (Suwanvecho et al 2017).

Concordance with expert oncologists is used both as a selling point by IBM, as we saw under Q1, and as the most significant quality parameter in the attempts to scientifically test the system. This make it pertinent to ask exactly what we can learn from such studies. But before we address this crucial question, there are a few things to note about these two early studies about Watson's performance. First of all – and easiest of all – we should point out that the first of these studies discloses funding from IBM, namely the company which develops and markets Watson; hardly independent funding, in short (you can find the relevant finding disclosures here).[10]

---

[10] https://coi.asco.org/Report/ViewAbstractCOI?id=193610

In fact, there are IBM affiliations also in the other study, but those appear to be co-authors who are IBM employees; while that might also lead to some questions, some level of collaboration is probably to be expected in these early studies, even if just because of access to the technology. We are not claiming that these funding disclosures and affiliations disqualify those studies, but it is important to be aware of this particular kind of small print.

Secondly, while Watson has been developed, as we have already seen, as a collaborative effort between IBM and the *Memorial Sloan Kettering Cancer Center* in New York, the two studies have been carried out in different healthcare systems (India and Thailand). This element of diversity raises its own complex medical, economic and political questions, which we will address later. Thirdly, it is worth noting that what is meant by concordance in these studies is that the treatment suggestion of the oncology experts was to be found amongst the treatments suggested by Watson in either the "Recommended" category or the "For Consideration" category. In other words, concordance does not necessarily mean that there was a match between the treatment recommended by the human experts and the treatments Watson categorized as recommended.

Having made these three preliminary points about those studies, what about the data itself? What do these studies tell us about Watson? Is it sufficiently reliable, or should we actually be scared about delegating medical decisions or decision-support to it? With concordance levels between a minimum of 76% and a maximum of 97.9%, what should we make of Watson's performance? Is 76% good enough? Is 97.9% exceptional?

Numbers alone – as it is often the case – don't tell us very much.[11] And before we can even begin to answer such questions, we need to address a more fundamental question: Should we be using concordance as an evaluative standard in the first place?

The answer to this first question crucially depends on whether we are interested in evaluating the potential economic or the potential strategic reasons for delegating tasks to Watson. Given that these studies compare the performance of Watson with that of human experts, one might think that the studies could serve as ammunition for strategic delegation. However, as we shall see in fact the opposite is the case. Assuming that delegating to Watson will save us resources, a high degree of concordance seems exactly fit to play the role of an economic reason for delegation. What matters for economic reasons is that the level of performance of the technology we delegate to is on par with the human experts whose tasks are being delegated. And a high degree of concordance can exactly serve as evidence for the claim that Watson performs at a level similar to that of the human experts.

However, if we are looking for strategic reasons, concordance does not seem to be the right measure. A high degree of concordance can at most show us that the system is no worse than the relevant human experts; it cannot, just on its own, serve as evidence for the superiority of the recommendations provided by Watson. In fact, if we are aiming for a system that can out-perform humans then per definition we must also be aiming for a system that to some extent should be in disagreement with the relevant human experts. Only so would there be a chance that the system gets right what the humans gets wrong. The problem with the present studies, however, is that we do not know if the lack of concordance is due to human failing where the system succeeds, the system failing where humans succeeds, or both parties being perhaps equally far from succeeding in the

[11] It should be mentioned that other local studies show a lower congruence rate. For instance, studies from South Korea show a relatively low level of congruence (~ 40%) (Choi et al. 2019; Lee et al. 2018). The involved South Korean hospitals ended up not implementing Watson (Choi 2018).

finding the optimal treatment. In short, we lack a standard that is itself independent of both the verdict of the relevant group of experts and the output of the system, and yet is a standard we have reason to believe is more objective than both Watson and the experts.[12]

Could we get our hands on such a more objective standard? In some cases of using IA-systems for diagnostic purposes, it makes perfect sense to speak of a more objective standard. We can test the reliability of respectively the IA-system and the human experts by asking them to make a diagnosis of cases where we already know, via methods that are more or less bullet proof, what the correct answer is. However, when it comes to what the best cancer treatment to a given patient is, it is not so easy to provide such an independent "gold standard".

Probably our most reliable method for establishing the likely outcome of a given treatment to a concrete patient is by comparison with the results of prospective randomized clinical trials involving patients similar to the concrete patient on relevant parameters. However, in many concrete cases we do not possess sufficient evidence-basis from such trials to establish with a high degree of certainty the likely outcome of different kinds of treatment. This is one reason why we can find disagreement between experts when it comes to assessment of which treatment should be recommended to a given patient. And it is part of the reason why IBM uses the recommendations of the experts at MSK Cancer Center as input when training Watson.[13]

Even if we had much more relevant clinical trials available the application of such general results to a particular case would still involve an ineradicable element of clinical judgement in a way that for instance certain methods used as gold standards in a diagnostic context does not. However, this

[12] See Tupasela and Di Nucci (*submitted*) for a further discussion of this *problem of concordance.*

[13] See the critical news article by Ross and Swetlitz (2017) for a description of the training sessions with Watson. Also see the already cited Somashekhar 2018: "Finally, lack of concordance does not necessarily provide evidence regarding whether the MMDT or WFO was 'correct' in its recommendation. As previously discussed, disagreements can have many valid explanations such as differences in how patients with comorbidities and increasing age are treated. There is no 'gold standard' beyond expert opinion, which studies have shown can vary profoundly during the evaluation of CDSSs."

should not lead us to rule out that an AI-system such as Watson could become better than the best human expert at predicting the outcome of specific treatments in concrete cases. If Watson, as IBM hopes for, becomes able to detect otherwise hidden relevant patterns through the processing of enormous data-sets, this capacity might end up beating the clinical experience of human experts. Our epistemic question is, however, how we could ever know that the machine outperforms the human. In lack of an independent gold standard, perhaps the best indication we could have for such outperforming would be that the human experts on due consideration of the evidence presented by the machine in support of its alternative proposals are willing, in a significant amount of cases, to admit that the system's recommendations should be followed. However, that is not where we are at now, where it is to a large extend the human experts at Memorial Sloan Kettering Cancer Center who, by feeding Watson with their preferred recommendations, set the standard for Watson.

To be fair, the fact that concordance-studies cannot serve as strategic reasons for implementing Watson, does not imply that there are no strategic reasons in favor of using Watson. Even if we assume that Watson's recommendations are not as superior to the human experts, they might still help qualify the decisions made. Given Watson's ability to scan massive amounts of scientific literature in a few seconds, it's only to be expected that it might come up with otherwise unknown evidence that can confirm or challenge the human experts' initial judgement thereby enhancing the quality of the final decision. Here, however, it is a specific part of the task of providing decision support that Watson might be better at than humans, and not the task of providing treatment recommendations as such. This is important because it means that we are a still far from having a general strategic reason in favor of delegating treatment decisions making to Watson. And in so far as what is needed if we are to so much as consider delegating decision control to a system like Watson is, as we have argued under Q1, exactly strategic reasons, we can conclude that we are not in possession of any reasons for considering such outsourcing at the moment.

All these complexities and epistemic difficulties should not let us lose sight of a much more simple observation: if Watson delivers advice that the oncology scientific community agrees is mistaken and less than safe – then that must mean that this particular medical AI system is not yet fit for purpose. And there is evidence that Watson has in fact proven problematic in such a manner: there were multiple media reports in 2018 according to which IBM's own internal documents suggested that Watson had often given 'unsafe and incorrect' advice.[14] Furthermore, not all concordance studies show high concordance rates. Furthermore, in concordance studies that show a high degree of concordance, we still lack a good analysis of what might explain the dis-concordance we nevertheless find in such studies.

Questions #3: Additional risks of delegating to algorithms

So far we have focused on epistemic issues about how to evaluate the performance of medical AI systems such as Watson. But on top of the uncertainty and difficulties that the epistemic issues in the previous section bring, what other risks are there with medical AI? One problem is what in the literature is often referred to as the problem of 'black box' algorithms (Pasquale 2015) and the related issue of opacity, which is in a slogan the idea that we don't understand how these algorithms function. Another easy way of understanding the opacity and black box concerns is to state the obvious semantic fact that opacity is the opposite of transparency. And that transparency – also as a

[14] See for example the following three, especially the first one which appears to be the direct source of the other media reports:
https://www.statnews.com/2018/07/25/ibm-watson-recommended-unsafe-incorrect-treatments/
https://www.telegraph.co.uk/technology/2018/07/27/ibm-watson-ai-criticised-giving-unsafe-cancer-treatment-advice/
https://www.dailymail.co.uk/sciencetech/article-6001141/IBMs-Watson-suggested-inaccurate-unsafe-treatment-recommendations-cancer-patients.html
https://gizmodo.com/ibm-watson-reportedly-recommended-cancer-treatments-tha-1827868882

necessary condition for informed consent (Di Nucci 2018, Chapter 6) – is particularly crucial when it comes to medicine and healthcare, so that opacity concerns have particular weight for medical AI.

Burrell (2016) identifies three different kinds of opacity concerns for machine learning algorithms:

(i) opacity in terms of proprietary protection or corporate secrecy;

(ii) opacity as technical illiteracy (lack of ability to read the code); and

(iii) opacity as a "mismatch between mathematical procedures of machine learning algorithms and human styles of semantic interpretation" (Burrell 2016: 3).

The three kinds of opacity refer to three different ways in which we might be left in the dark as to how an AI-system solves its task. Opacity due to proprietary protection and corporate secrecy is a real concern in cases like IBM's Watson where the company's economic interests can be in obvious conflict with our interest in having the best possible basis for a rigorous scientific testing of the reliability of the system. Opacity due to technical illiteracy becomes relevant here since most doctors do not have the technical knowledge to understand the inner workings of Watson. This could be said to be the natural result of sensible division of labour, but because we do not have a simple objective "gold standard" against which we can measure Watson's output, it matters that doctors understand how Watson reached its conclusions, in terms of reasons they themselves can recognize as good reasons. Without such acknowledgement by human experts of the rationality of Watson's conclusions, we would basically be putting our trust in Watson blindly. The third kind of opacity has to do with exactly the principled possibility of translating the "rationality" of machine learning algorithms into a kind of reasoning that can be followed by humans. The worry is that the complexity of such systems' pattern recognition might rise to a level, where it is no longer just a lack of technical skill that can make them opaque but rather a proper incommensurability between the workings of system and the way humans reason via natural language. This is a worry that is of huge practical and theoretical importance, but also one that it would take us to far from our specific focus to pursue further here.

Generally, we can say that a high degree of epistemic transparency, allowing the human experts to check whether Watson's recommendations are grounded in good evidence is of utmost importance since we have no way of shortcutting this element of human evaluation via an independent reliability measure. It remains to be seen whether this possibility of checking will decrease, perhaps for principled reasons, as the capacity of Watson to detect patterns in huge datasets increases.

Watson doesn't just solve a theoretical task, as for instance figuring out what a patient's most probable diagnoses is, but also gives practical recommendations as to which treatment should be offered. This means that it is not just the epistemic standards of Watson that need to be transparent, but also what Watson so to speak values as a good outcome of a treatment. This value-element comes with its own risks that go beyond the opacity issues above. McDougall (2019) has for example emphasized that currently Watson's main criterion is 'disease-free survival'; as in, each option is ranked according to the chances of 'disease-free survival' and those numbers are based on existing evidence within the medical literature surveyed by Watson's algorithm.

Here it is helpful to focus on 'disease-free survival' because even such a basic and seemingly uncontroversial criterion poses fundamental questions: should 'disease-free survival' always be the fundamental goal of treatment, no matter the costs? The whole field of palliative care and end-of-life decision making testifies to the fact that 'disease-free survival' statistics might not always be the whole story.[15]

[15] More on these life and death issues in Di Nucci 2018, Chapter 5. Here we could alternatively think in terms of QUALYS, Quality-Adjusted-Life-Years (again thanks to Isaac Wagner for the suggestion).

The problem, though, is not only which variable we settle on or on which grounds we choose a certain variable over others – those decisions are, after all, contested and difficult whether or not artificial intelligence is involved. The additional problem that the involvement of algorithms brings to the table is that of altering the dialectic of contested variables such as ‘disease-free survival’. Once IBM chooses ‘disease-free survival’ for Watson then already that corporate decision alone shifts the power dynamics around oncological decision-making by making a particular option that used to be contested the new default (similarly to the concern that smartphone and app developers make design choices which set a particular (perhaps non-ideal) option as the default and thus have a huge (negative) impact on how humans interact with the tech they have designed).

Put it another way: numbers already have the advantage (especially in a technological society) of us being able to operationalize them; and this advantage becomes entrenched when powerful actors choose variables that can be operationalized over variables that are more difficult to operationalize. And the very reason for choosing one variable over another is obviously whether it can be operationalized. We have come full-circle but we have to be careful not to do so by begging the question: namely the decisive reason why we choose one variable over another cannot be alone the possibility of operationalizing that particular variable, in this case ‘disease-free survival’ chances. Otherwise technology risks silencing the lively debate around end-of-life decision making, for example. More broadly, technology would then end up justifying itself if a certain variable is only valuable because it can be operationalized and is thereby tech-valuable, so to speak. In a slogan, there must be value to it beyond its value to technology.

There are other risks associated with Watson which do not depend on whether or not it is successful or on how to measure its possible success or failure. One particularly interesting risk that we want to focus on here is Watson’s global target: as we pointed out, the early evidence on concordance for Watson comes often from second- or third-world healthcare systems. On the one hand, that’s great news: if through software we can bring medical expertise within oncology to countries and healthcare systems which would otherwise not have the resources to develop such expertise themselves, then Watson brings with it great developmental promise in terms of global health and global justice and we should celebrate that.

On the other hand, though, a few cautionary notes that must be stressed: first of all, if those healthcare systems didn't previously have access to such medical expertise within oncology because of lack of funds, buying Watson licenses might end up exacerbating their financial problems. Here we must also consider the fact that – in a country like India say - scare resources might be better invested in medical education than foreign technology, at least from a long-term sustainability point of view.

Secondly, if those countries and healthcare systems did not previously have the relevant oncology expertise, on which grounds are they going to be able to assess Watson? In such cases, it is clear that concordance cannot be the correct epistemic measure since the point of introducing Watson would be to improve on their current expertise in oncology, i.e. what we are looking for are strategic, not economic reasons for adopting Watson. But we must also be careful that they will not be put under unduly pressure to accept the authority of outside expertise resulting in a form of medical colonialism As long as Watson's output is to a high degree the result of the training by local experts at MSK, we cannot expect Watson to for example take into account how cultural (including for instance eating habits) and environmental differences can make a difference to which treatment should be recommended.

In conclusion to this section we want to add a fourth risk category on top of the opacity-related risks, the value-related risks and the risks related to the politics of globalization in healthcare: the risk that doctors, whether in New York, India or Copenhagen, start relying too much on Watson.[16] This phenomenon is well-known within aeronautics, where pilots are often asked to land planes manually even if the auto-pilot would be in a position to land (so-called 'autoland'), just to make sure that pilots do not forget how to land a plane and are kept up to date through practice.[17]

How do we make sure that oncologists do not end up relying too much on Watson? Surely before being able to answer this question we must deal with the related question of defining what would it mean for a human oncologist to use Watson too much. In the current implementation of Watson as we have described it here, we can all agree that using Watson too much would be constituted, for example, by a human oncologist uncritically passing on Watson's advice to the patient without having herself first assessed and evaluated the plausibility of Watson's advice. That way the oncologist would for example be culpable of contributing to perpetuating and exacerbating possible *algorithmic biases* about therapeutic options which may be contained within the code.[18]

There are other possible scenarios that would count as relying too much on Watson: for example if a human oncologist would answer the patient's questions on the soundness of a certain therapeutic options with "It was Watson's first choice" or "It's what Watson recommends". This would be just as bad – in fact possibly actually worst – then if the doctor replied "It's my first choice" or "It's what I recommend". Neither reaction meets basic communication and informed consent requirements but the latter at least allows the patient – in theory anyway – to challenge the source of

[16] There has in fact already been a pilot trial of Watson at Rigshospitalet in Copenhagen. See Tupasela & Di Nucci (submitted) for more details on this trial. Based partly on poor congruence Rigshospitalet decided not to implement Watson (se also Ross & Swetlitz 2017).

[17] https://eu.usatoday.com/story/travel/columnist/cox/2014/02/09/autoland-low-visibility-landings/5283931/

[18] For more details on algorithmic bias, see the following: Hajian et al. 2016, Bozdag 2013, Baeza-Yates 2016, Kirkpatrick 2016, Lambrecht & Tucker 2018, and Danks & London 2017.

the advice, the oncologist with whom they are talking; while the former defers to an absent authority – the hidden algorithm – that the patient can neither see nor be expected to understand.

There are also more subtle ways in which oncologists might end up relying too much on Watson: they might, for example, become habituated to Watson and to certain outcomes that its algorithms deliver, so that the human oncologists through habituation would end up trusting Watson's outcomes too much and questioning them too little. That would be a particularly frightening prospect because algorithmic bias would generate further human bias thereby creating a whole vicious circle of human and technological biases and mistakes.

Even though it will be difficult and beyond the scope here to precisely define what exactly will count as relying too much on Watson, we can agree that there is such a thing as relying too much on it, as the examples above show. And that is why it is important to establish processes that ensure that human oncologists do not rely too much on Watson's authority and are always capable to critically assess and evaluate Watson's advice.

Question #4 and concluding remarks

In conclusion, we can now at least attempt to answer the basic ethical question of what Watson is good for and what we ought not to use Watson for. We have distinguished between delegating oncology decision-making itself to Watson and only delegating decision-support (namely advice for the clinician) to Watson and we have recommended that, at least at this early stage, only the former (if at all) should be delegated to an algorithmic system such as Watson.

We have further distinguished between so-called 'strategic delegation' in which the aim is to improve performance and 'economic delegation' in which the aim to is have comparable performance for less invested resources. While we have identified multiple possible problems and risks with ‘strategic delegation’ to Watson – including the concordance issues that we have discussed at length – the system appears to be more promising in terms of ‘economic delegation’, namely if we restrict its use to basic tasks and thereby free some of the human oncologist’s precious time.

Finally, we have identified possible problems and risks that are independent of issues concerning the soundness of Watson’s advice and how to measure whether Watson’s advice is in fact sound; risks related to such issues as opacity, transparency, end-of-life decision-making and also global health justice; and finally algorithmic bias and the risk that healthcare systems end up relying too much on Watson thereby losing critical human skills.

We hope that this attempt at laying the groundwork for a full ethical assessment of Watson has provided a useful case study in at least three general directions: in the first instance, how to decide whether a particular information technology system can be safely and successfully adopted with healthcare; secondly, how ethical questions often depend upon and are intertwined with other complex philosophical questions such as the epistemological ones we have identified in this chapter; thirdly and more broadly, we hope to have provided a practical example of how to *use* ethics and philosophy in the real world.[19]

[19] Many thanks to Isaac Wagner for written comments on an earlier draft.

References

Baeza-Yates, R. (2016, May). Data and algorithmic bias in the web. In *Proceedings of the 8th ACM Conference on Web Science* (pp. 1-1). ACM.

Bozdag, E. (2013). Bias in algorithmic filtering and personalization. *Ethics and information technology*, *15*(3), 209-227.

Burrell J 2016 How the machine 'thinks': Understanding opacity in machine learning algorithms. *Big Data & Society* 3(1) :1-12.

Youn I Choi, Jun-won Chung, Kyoung Oh Kim, et al. 2019. Concordance Rate between Clinicians and Watson for Oncology among Patients with Advanced Gastric Cancer: Early, Real-World Experience in Korea. *Canadian Journal of Gastroenterology and Hepatology*, 2019, Article ID 8072928. https://doi.org/10.1155/2019/8072928.

Choi, MH. 2018. "Major Hospitals in S. Korea not very interested in Watson.2 March 28, 2018. http://www.businesskorea.co.kr/news/articleView.html?idxno=21308. Visited May 14.

Danks, D., & London, A. J. (2017). Algorithmic bias in autonomous systems. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence* (pp. 4691-4697).

Di Nucci E. 2018. *Ethics in Healthcare*. Rowman & Littlefield (danish edition : *Ethics4Medics*, Munksgaard 2018).

Di Nucci E. 2019. Should we be afraid of medical AI? *Journal of Medical Ethics* (forthcoming).

Di Nucci E. *The Control Paradox* (book manuscript).

Domingos P 2015 *The Master Algorithm*. Basic Books.

Fry H. 2018. *Hello World*. Penguin.

Grote & Di Nucci (forthcoming) Algorithmic Decision-Making and the Problem of Control. In: Beck B & Kühler M (eds) *Anthropology, Technology and Responsibility*. Metzler.

Hajian, S., Bonchi, F., & Castillo, C. (2016). Algorithmic bias: From discrimination discovery to fairness-aware data mining. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 2125-2126). ACM.

Kirkpatrick, K. (2016). Battling algorithmic bias: How do we ensure algorithms treat us fairly? *Communications of the ACM*, *59*(10), 16-17.

Lambrecht, A., & Tucker, C. E. (2018). Algorithmic bias? An empirical study into apparent gender-based discrimination in the display of STEM career ads.

Lee et al. 2018. Assessing Concordance With Watson for Oncology, a Cognitive Computing Decision Support System for Colon Cancer Treatment in Korea. *JCO Clinical Cancer Informatics*, 2(2): 1-8.

McDougall R. 2019. Computer knows best? The need for value-flexibility in medical AI. *Journal of Medical Ethics* 45 (3): 156.

Mittelstadt, B. D., & Floridi, L. (Eds.). (2016). *The ethics of biomedical big data* (Vol. 29). Springer.

Pasquale F 2015. *The Black Box Society*. Harvard UP.

Santoni, F. & Van den Hoven, J. (2018). Meaningful Human Control over Autonomous Systems: A Philosophical Account. *Frontiers in Robotics and AI, 5*, 15.

Somashekhar et al 2017. J Clin Oncol 35, 2017 (suppl; abstr 8527)

Susskind R & Susskind D (2015), *The Future of the Professions*. Oxford UP.

Suwanvecho et al 2017. J Clin Oncol 35, 2017 (suppl; abstr 6589)

Thaddeus Beck et al 2017. J Clin Oncol 35, 2017 (suppl; abstr 6501)

Tegmark M. 2017. *Life 3.0*. Penguin.

Topol E. 2019. *Deep Medicine*. Basic Books.

Tupasela A & Di Nucci E. (submitted). The Problem of Concordance as Evidence in Oncology Decision-Support Systems. *AI & Society.*

Wajcman J 2017 Automation : is it really different this time ? British Journal of Sociology 68 (1) : 119-127

STUDY QUESTIONS

I) What is the difference between decision-making and decision-support? And more specifically: what is the difference between delegating decision-making to a technological system and delegating decision-support to a technological system?

II) What is the difference between strategic delegation and economic delegation?

III) List at least three possible risks of Watson and assess the likelihood and plausibility of each of these three risks.

IV) Discuss, also based on the Watson case study, whether we are increasingly losing the human touch in healthcare.

V) Compare the case of Watson with another recent technological implementation within healthcare.