5. Delegation, relinquishment, and responsibility: The prospect of expert robots

Jason Millar* and Ian Kerr**

I don't quite know how to phrase my problem. On the one hand, it can be nothing at all. On the other, it can mean the end of humanity.

Stephen Byerley, World Coordinator[1]

1. INTRODUCTION

If the creators of *Jeopardy!* ever decide to adopt the *Who Wants to be a Millionaire?* feature that allows contestants to "call a friend" for help answering a question, they could prepare a single speed-dial option: 1-8-0-0-W-A-T-S-O-N. In the 2011 public spectacle that pitted the two all-time *Jeopardy!* champions, Ken Jennings and Brad Rutter, against IBM's robot *cogitans*, Watson, the "deep question answering" supercomputer made counterparts of its human counterparts. Watson's win in the *IBM*

* Jason Millar would like to thank Michael Froomkin, Patrick Hubbard, Evan Selinger, and the great participants at the inaugural We Robot (2012) who provided generous feedback on drafts of this chapter, as well as Sergio Sismondo and Elena Ponte for feedback and help with drafts. He also wishes to recognize the Social Sciences and Humanities Research Council (SSHRC) and the Canadian Institutes for Health Research (CIHR) for generously funding research activities that inspired his pursuit of this topic.

** Ian Kerr wishes to extend his gratitude to the Social Sciences and Humanities Research Council and the Canada Research Chairs program for the generous contributions to the funding of the research project from which this chapter derives. Thanks also to his wonderful colleagues Jane Bailey, Chloe Georas, David Matheson, Madelaine Saginur, Ellen Zweibel, and the super six pack: Eliot Che, Hannah Draper, Charlotte Freeman-Shaw, Sinzi Gutiu, Katie Szilagyi, and Kristen Thomasen for their inspiring thoughts on this important emerging subject.

[1] ISAAC ASIMOV, *The Evitable Conflict*, in I, ROBOT 447 (2004).

Challenge was decisive and momentous. With it, IBM's system obliterated the human monopoly on natural language, rendering Watson the world's go-to expert at *Jeopardy!*. As Jennings, the game show's foremost human expert said of Watson, "I, for one, welcome our new computer overlords."

Jennings's quote was prescient, if ambivalent. Watson's success raises questions about what role humans will occupy once robots are capable of performing a multitude of tasks traditionally delegated to human experts – and performing them well. On the one hand, Jennings seems enthusiastically to accept that Watson is the successor to human dominance in the game of *Jeopardy!*. On the other, he suggests that a central tradeoff of creating highly skilled robots takes the form of a loss of human control.

Given the benefits that robots might some day confer on humanity, *should* we, indeed *can* we, remain in control once they emerge superior with respect to particular abilities?

Responses to such scenarios vary by degrees of dystopia. Here is one portrayal:

> First let us postulate that the computer scientists succeed in developing intelligent machines that can do all things better than human beings can do them. . . . Either of two cases might occur. The machines might be permitted to make all of their own decisions without human oversight, or else human control over the machines might be retained.
>
> If the machines are permitted to make all their own decisions, we can't make any conjectures as to the results, because it is impossible to guess how such machines might behave. . . . It might be argued that the human race would never be foolish enough to hand over all the power to the machines. But . . . the human race might easily permit itself to drift into a position of such dependence on the machines that it would have no practical choice but to accept all of the machines' decisions. As society and the problems that face it become more and more complex and machines become more and more intelligent, people will let machines make more of their decisions for them, simply because machine-made decisions will bring better results than man-made ones.[2]

These are the words of Theodore Kaczynski, better known to most as the Unabomber. Bill Joy made the passage famous when he quoted it in its entirety in his famous *Wired* essay, "Why the Future Doesn't Need Us."[3] In that essay, Joy, the former chief scientist of Sun Microsystems, set out his concerns over the unanticipated consequences of "GNR" technologies

---

[2] Ted Kaczynski, The New Luddite Challenge (2001[1995]), *available at* http://www.kurzweilai.net/the-new-luddite-challenge.

[3] Bill Joy, *Why the Future Doesn't Need Us*, 8 WIRED, (2000). Joy learned of Kaczynski's passage from Ray Kurzweil, who also quoted this entire passage in his THE AGE OF SPIRITUAL MACHINES (2000).

(genetic, nanotechnology, robotics), expressing amazement and surprise at the rapid rate of their development, ultimately calling for relinquishment. Not relinquishment of human control; relinquishment of GNR.

Despite Bill Joy's well demonstrated, near-term concerns, the prospect of having to make policy decisions about relinquishing control to robots might strike some as far-off or even far-fetched.

In this chapter, we suggest that the precipice of having to decide whether to relinquish some control to robots is near. In the not-too-distant future, we will have to make some difficult choices. On the one hand, we might choose to accept the relative fallibility of human experts and remain in total control. Alternatively, we may decide to forge our robot "overlords" and relinquish some control to them for the greater good. Although such choices do not entail the extreme outcomes of the Unabomber's dystopia, we aim to demonstrate that even Kaczynski's basic logic is neither far-off nor far-fetched.

In order to better understand the paths forward and some of the moral choices each path will present, we offer a kind of logical narrative – reason's road from here to there. In so doing we explore two important questions regarding the expert systems that will drive tomorrow's robots if we take that turn. First, at what point are we justified in relinquishing control of certain highly specialized expert tasks to robots? Second, how would answers to the first question bear on the question of determining responsibility, particularly when expert robots disagree with their expert human coworkers with undesirable outcomes?

We argue that, given the normative pull of *evidence-based practice*, if we go down the path of developing Watson-like robots, we will be hard-pressed to find reasons to remain in control of the expert decisions at which they excel. If, instead, we choose to remain in control, we might deliberately be advocating a status quo in which human experts deliver less than optimal outcomes, *ceteris paribus*, to what *co-robotics*[4] might otherwise achieve. Even if today it is not immediately clear whether either decision is anything to worry about at all or presages the end of humanity, it remains our responsibility to face these difficult choices, head on, before we find ourselves wondering how we got to where we are. Consequently, we need to carefully consider the right way to think about these robots. Are they merely tools or are they something more? Speculating about the near- and mid-term future of Watson-like robots, we suggest, may push us to reconsider established legal categories.

---

[4] We use the term *co-robotics* to refer to any situation in which human and robot experts are working alongside one another.

2. WHAT KINDS OF ROBOTS?

Given its expertise in building expert systems, it is not surprising that IBM has big plans for Watson. For starters, Watson will receive ears and a new voice.[5] This will eventually allow Watson to interact with people in ways it could not (without help) during the *IBM Challenge*. Watson is being used as a fashion coach, helping people pick outfits that work well for them;[6] as a culinary consultant, designing novel lists of ingredients that work well together and that could be used as the basis of new recipes;[7] and even scouring the intellectual property literature to assist patent lawyers in their work.[8] IBM has also transformed Watson into a medical expert. Watson has been programmed with "clinical language understanding," a field of natural language processing focused on "extract[ing] structured, 'actionable' information from unstructured (dictated) medical documents."[9] IBM has installed its expert system at several major health care organizations in the United States in order to test Watson on the front lines.[10] At the New York Genome Center and the Memorial Sloan-Kettering Cancer Center, Watson is using genetic information about patients and their tumors in order to help doctors identify the best candidate cancer drugs for a patient from the vast medical (and other scientific) literature that is produced every year.[11] That is the challenge of medical practice that Watson is designed to overcome: the volume of medical scientific evidence is doubling every five years, making it difficult if not impossible, for humans to stay on top of the latest and greatest in medical knowledge, let

---

[5] IBM, IBM to Collaborate with Nuance to Apply IBM's "Watson" Analytics Technology to Healthcare (2011), *available at* http://www-03.ibm.com/press/us/en/pressrelease/33726.wss.

[6] J. Taylor, *IBM's Watson Aims to Pick the Fashion Trends*, ZDNet, Sept. 2, 2014, *available at* http://www.zdnet.com/ibms-watson-aims-to-pick-the-fashion-trends-7000033182/.

[7] J. Jackson, *IBM Watson Cooks Up Some New Dishes*, PCWorld, August 29, 2014, *available at* http://www.pcworld.com/article/2600780/ibm-watson-cooks-up-some-new-dishes.html

[8] IBM, IBM BAO Strategic IP Insight Platform (SIIP) (2014), *available at* http://www-935.ibm.com/services/us/gbs/bao/siip/.

[9] Nuance, Clinical Language Understanding (2014), *available at* http://www.nuance.com/for-healthcare/resources/clinical-language-understanding/index.htm.

[10] Matthew Herper, IBM's Watson Attempts to Tackle the Genetics of Brain Cancer, Forbes, Mar. 19, 2014; Tom Groenfeldt, Big Data Delivers Deep Views of Patients for Better Care, Forbes, Jan. 20, 2012.

[11] IBM, Memorial Sloan-Kettering Cancer Center, IBM to Collaborate in Applying Watson Technology to Help Oncologists (2012), *available at* http://www-03.ibm.com/press/us/en/pressrelease/37235.wss.

alone incorporate it into practice.[12] Watson's ability to "understand 200 million digital pages, and deliver an answer within three seconds"[13] makes it an attractive expert robot candidate.

In the not-too-distant future, IBM will be marketing Watson's descendants to healthcare providers worldwide.[14] If they perform as well as Watson did on *Jeopardy!*, Watson's descendants will become the go-to medical experts in their fields. One can imagine doctors consulting Watson with the same frequency and reliance that *Jeopardy!* contestants would, if they could.

Watson's success on *Jeopardy!*, and in other knowledge domains, stems directly from having been programmed in a very particular way – *one that makes Watson's answers unpredictable to its programmers*. When it comes to *Jeopardy!*, Watson "knows" how to formulate questions in response to the clues, rather than simply being programmed to respond to a known set of inputs. Watson works by scouring a set of data that, in theory, could span the entire Internet, for information that it deems relevant to the particular task, then learns over time how best to "make sense" of that information. By shifting Watson's parameters toward medical, legal, culinary, fashion, and other sets of information, IBM hopes that Watson will do to those knowledge domains what it has already done to the world of *Jeopardy!*.

Watson can be considered a forerunner to an era that Chris Anderson calls "the end of theory"[15] – an age in which we come to rely on robotic prediction machines in place of human experts. Not because the robots' software provides better theoretical accounts of the relevant phenomenon (in fact, they don't provide any). Rather, we will rely on robots without really knowing why – simply because their algorithms provide the greatest number of successful outcomes. We have already seen this in Google's search approach. Neither Larry nor Sergey (nor any other Google employee) knows exactly why one particular web page is a better result than another. When the click patterns say it is, that's good enough. No semantic or causal analysis is required. Like the oracles of previous times, Google's search engine and IBM's Watson are prototypes for prediction

---

[12] *Id.*

[13] *Id.*

[14] M. Castillo, *Next for Jeopardy! Winner: Dr. Watson, I Presume?*, TIME, Feb. 17, 2011, *available at* http://www.time.com/time/business/article/0,8599,2049826,00.html.

[15] Chris Anderson, *The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, WIRED, June 23, 2008, *available at* http://www.wired.com/science/discoveries/magazine/16-07/pb_theory.

bots that divine without knowing. Like the ancients, we will, quite rationally, come to rely upon them, knowing full well that we cannot necessarily explain the reasons for their decisions.[16]

IBM is not alone in its willingness to relinquish knowledge and control to the machines. Google has similar plans with their Google Driverless Car (GDC) project.[17] Equipped with an array of sensors and the predictive code to make sense of the highly contextual interactions taking place in a driving environment, the GDC is poised to radically improve driving safety. Eventually, the tradeoff for improved safety will be a requirement that humans must let go of the wheel. The prospect of letting go might be a straightforward proposition for the "Googly" roboticist who believes "the data can make better rules."[18] But letting go may, at first, be a tougher sell to humans harboring a nagging distrust of Googly roboticists and their toys.[19] Keeping in mind that our safety is at stake, the prospect of riding on the rails of expert systems like (Dr.) Watson or the GDC raises a number of questions about our delegation of human tasks such as driving and diagnosing to expert machine systems.

3. UNPREDICTABLE BY DESIGN

We begin our analysis by defining the kind of robotic systems that we are concerned with: those for which unpredictability in its operations is a feature and not a bug. Watson exemplifies such systems and is the descriptive focal point of our analysis.

Watson's inherent unpredictability derives in part from the fact that its inputs are, in principle, the complete set of information "out there" on the Internet. That set of data is constantly changing as a function of the behavior of millions of individuals who constantly contribute new bits of information to it, the content of which is also unpredictable. When sets of algorithms act on such a vast set of ever-shifting inputs, the outputs become unpredictable.

The set of targeted health information available to a Watson-like system will grow significantly as various jurisdictions implement "digital

---

[16] Elena Esposito, *Digital Prophecies and Web Intelligence*, in PRIVACY, DUE PROCESS AND THE COMPUTATIONAL TURN: THE PHILOSOPHY OF LAW MEETS THE PHILOSOPHY OF TECHNOLOGY (Mireille Hildebrandt & Katja De Vries eds., 2013).

[17] T. Vanderbilt, *Let the Robot Drive*, WIRED, Feb. 2012.

[18] *Id.*

[19] J. Millar, *You Should Have a Say in Your Robot Car's Code of Ethics*, WIRED, Sept. 2, 2014, *available at* http://www.wired.com/2014/09/set-the-ethics-robot-car/.

health systems." These digital information repositories have the potential to house every piece of health information related to the individuals (patients and healthcare providers) accessing that system, a process that is either underway or completed in most of the world. According to Kaiser Permanente, one of the largest health maintenance organizations (HMOs) in the United States, at least three terabytes of data is generated every day as their 14 million patients access health-related services.[20] Watson's ability to scour that fluid set of data and glean meaningful information from it, in a way that is largely unpredictable to its designers, is what gives it an edge over its human competitors.

Watson-like systems stand in contrast to simpler robots that are designed with strict operational parameters in mind. Industrial robots, such as those that paint cars or weld joints, are programmed to function in highly predictable ways. Unlike Watson, their every movement is predetermined and structured to eliminate variation between "jobs." Successful programming, for these robots, means that they never surprise their programmers; for these robots, experiencing fewer surprises is considered better design.

Watson, on the other hand, is designed to surprise. Though software is written in comprehensible lines of code, software functions that parse and operate on massive, constantly changing data sets, deliver results that no programmer can fully anticipate.

Today's examples of these kinds of software systems include, in addition to Watson and the GDC, the various predictive data-mining engines built into Google, Facebook, Amazon, and iTunes.[21] Depending on the wording of an email being read by a user at a particular instant, or the webpages that a user has visited in the previous few minutes, those predictive algorithms will deliver up different results. No programmer could predict the outputs of such a system with any degree of accuracy, and certainly not within the time constraints that makes the difference between judging a predictive system "successful" or "unsuccessful."

Unpredictability is being amplified in newer "machine learning" computer chips that are designed to mimic the plasticity of the human brain.[22]

---

[20] P. Webster, *Electronic Health Records: An Asset or a Whole Lot of Hype?*, 182 Can. Med. Ass'n J. (2010).

[21] Ian Kerr, *Prediction, Presumption, Preemption: The Path of Law After the Computational Turn*, in Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology (Mireille Hildebrandt & Katja De Vries eds., 2013).

[22] J. Volpe, *IBM's New Supercomputing Chip Mimics the Human Brain With Very Little Power*, Engadget, Aug. 7, 2014, *available at* http://www.engadget.com/2014/08/07/ibm-synapse-supercomputing-chip-mimics-human-brain/.

These chips will not run programs per se; instead, they will independently, and over a period of time, form tiny interconnections between internal pieces of hardware-based logic, resulting entirely from continuous stimulation at the inputs. Like a human brain learning from environmental stimuli delivered via nerves, next-generation robot brains will change over time. Thus, they will not be programmed in the traditional sense of the term, making them even less predictable than today's most sophisticated and unpredictable robots.

4. EXPERT ROBOTS?

How might we describe robots that are capable of performing complex tasks, often unpredictably?

For starters, if programmers are able to understand Watson's lines of code, but are unable to predict Watson's inputs and outputs based on them, then we can say that Watson's lines of code – the rules upon which it operates – *underdetermine* its resulting behavior. That is, Watson's ability to win at *Jeopardy!* cannot be fully explained by reference to the many lines of code that make up its programs. Interestingly, a great deal of the sociology of "expertise" focuses on the role of underdetermination in delineating experts from non-experts.[23] That human experts are unable to fully describe their actions in terms of "rules" plays a big role in understanding what their expertise consists of.[24] That Watson-like computers similarly cannot be fully understood by reference to their lines of code opens the door to, one day, describing them as *expert robots*.

Some of the pioneering work linking underdetermination with expertise comes from Harry Collins's observations made while studying scientists at work in laboratories. A classic example of scientific expertise is the ability to reproduce another scientist's experiment. One way of attempting to reproduce an experiment is to read the detailed methodology (lets

---

[23] Harry Collins & Robert Evans, Rethinking Expertise (2007). Collins and Evans provide a great deal of the pioneering work on the sociology of expertise, which forms a branch of Science and Technology Studies (STS). Earlier discussions on underdetermination stem from Collins's pioneering work on expertise: Harry Collins, Changing Order (1992). See also: Harry Collins & Martin Weinel, *Transmuted Expertise: How Technical Non-Experts Can Assess Experts and Expertise*, 25 Argumentation 401–13 (2011).

[24] In STS, explanations of underdetermination are based on a particular interpretation of Ludwig Wittgenstein's work on game theory (*see* Collins, 1992, *supra* note 23, for a fuller explanation).

call those the rules) contained in a published scientific paper and then attempt to do what the other scientist(s) did. In writing up a methodology scientists, one might assume, should be able to transcribe their expertise into a set of rules that could then successfully be followed by any another scientist with expertise in that scientific area. Collins notes, however, that when scientists attempt to reproduce experiments, scientific papers and even lab notes – the detailed rules – are typically insufficient for successful reproductions.[25] The set of rules, he concludes, does not adequately determine the actions required to demonstrate expertise in carrying out *that* experiment. The codified rules are not where the expertise resides.

Scientists trying unsuccessfully to reproduce an experiment typically interpret failure as an indication that a different type of knowledge is required. In other words, there is an understanding among scientific experts that codified descriptions of how to do an experiment are insufficient for actually doing it. The kind of knowledge needed to "fill in the blanks" comes via direct mentoring: scientists must work alongside other scientists in order to understand how actually to do the experiment.[26] In fact, Collins and Evans's claim is that only by *doing* can one develop high levels of expertise in any field, because the doing is what provides the *tacit knowledge* – "things you just know how to do without being able to explain the rules for how you do them"[27] – that makes one an expert. The upshot of their work on expertise is that tacit knowledge helps to delineate seasoned from novice experts, and experts from non-experts: the greater the level of expertise one has achieved, the greater the amount of *tacit knowledge* that person possesses. Tacit knowledge can be seen as the difference between merely *describing* how to complete an expert task and *actually being able to do it*.

When we turn to the question of whether robots might someday be described as experts, there are certain analogies that can be drawn between human experts and robots like Watson or the GDC, and others that cannot. Though we will certainly not be able to claim that the current slate of robots is capable of forming communities of expertise in the strong linguistic sense emphasized by Collins and Evans, we can locate a weak form of tacit knowledge among robots that we think could qualify talk of them as expert robots.[28]

---

[25] Collins, (1992) *supra* note 23.

[26] *Id.*, Collins & Evans, *supra* note 23.

[27] Collins & Evans, *supra* note 23, at 13.

[28] John Searle's famous argument against strong artificial intelligence (*Minds, Brains, and Programs*, 3 Behavioral and Brain Sciences 417–57 (1980)) is a good example of what we mean here. We do not claim that robots have intentionality

Human expertise is most often described as the result of a process of socialization into a group of human experts; expertise is acquired through strong language-based interactions with other humans.[29] In stark contrast to most humans, Watson and the GDC cannot enter into traditional, language-based apprenticeships with other human experts. However, Watson and the GDC are able to receive mentoring of a sort from human experts, who help refine their algorithms based on expert judgments of when the robots have gotten things correct, and when they have erred. Those expert human judgments are "interpreted" by machine learning algorithms, allowing the robot to form new associative rules between data points and subsequently affecting the robot's behavior. This suggests that human experts are able, in a weak sense, to help the robots improve their functioning with respect to specific tasks usually associated exclusively with human expertise. Though the mentoring is not language-based in the traditional sense, it is directed at improving the robots' performance beyond the rank of novice.

As a result, Watson and the GDC are able to independently extract meaningful information from large sets of unstructured information, even in novel situations, in a way that human experts interpret as "correct" or "incorrect." This indicates that a sort of bi-directional communication is possible. When Watson correctly interprets the meaning of a subtle play on words in *Jeopardy!*, Watson is communicating *something more than just an answer*. It is most often the *correct* answer to a question that would normally require a high degree of language-based expertise to generate. Moreover, Watson is able to deliver correct answers with a higher degree of success than the humans experts against whom it is playing. Thus, the weak form of mentoring and communication seems able to foster a high degree of *ability* in robots to function in tasks otherwise reserved for human experts.

If it were the case that Watson or the GDC were operating according

---

(desires, and the like), nor are we claiming that when robots communicate information they understand what they are doing in the sense that a human expert would. We are happy to stick to a modest, *weak* sense of robot communication (and mental life), such that when we say robots engage in expert-like communication we simply mean they are able to act on unstructured information, extract meaningful information from it, and convey meaningful information to humans, in the same way that a human expert would from the recipient's perspective.

[29] Collins & Evans, *supra* note 23, Collins, *supra* note 23, Collins & Weinel, *supra* note 23; M. Casper & A Clarke, *Making the Pap Smear into the "Right Tool" for the Job: Cervical Cancer Screening in the USA, Circa 1940–95*, 28 Social Studies of Science 255 (1998); Thomas Kuhn, The Structure of Scientific Revolutions (1962).

to highly predictable programs with relatively constrained sets of inputs and outputs, then we would be inclined to describe them merely as sophisticated tools rather than experts. But Watson and the GDC are able to achieve high degrees of something like "expertise" by acting on sets of rules (their programs) that underdetermine their success. Thus there is a critical descriptive gap to be filled if we are to explain how it is that Watson-like robots do what they do so well. Like human experts, we suggest that what fills that gap between the codified programs and Watson's abilities is, in a weak sense, expert tacit knowledge.

Rather than getting stuck on a requirement that experts must communicate verbally in a community of expertise, as Collins and Evans do, we suggest that a robot's ability to function like human experts in certain tasks, unpredictably and underdetermined by any readily available description of its underlying programs, is enough to qualify it, if only in a weak sense, as an expert. Thus, we argue that it might make sense to talk of expert robots.

When would we call a robot an expert? Humans become experts when they are accepted into a community of human experts. According to Collins, Evans, Ribeiro and Hall,[30] this requires novices to pass Turing-like language tests to prove themselves. Take academic expertise as an example. Recognized experts (professors) interact with candidate experts (students) until such time as those recognized experts confer the status of expert (PhD) on the candidates. This requires the candidates to convince the recognized experts that they have reached the required level of proficiency in certain well-defined tasks (writing papers, arguing, evaluating other students, identifying important problems, and so on). When there is strong evidence that the candidates are able to perform those tasks at an expert level, as defined by the relevant other experts, the students are seen as experts in their own right.

Similarly, we suggest that a robot can be considered an expert when there is strong evidence it is capable of consistently performing a well-defined set of tasks traditionally associated with human expertise, with results that are, on average, better than the average human expert. Thus, for example, a robot that consistently beats the best human *Jeopardy!* experts qualifies, if only in the weak sense described above, as an expert at *Jeopardy!*. Similarly, a robot that can operate a vehicle with a safety record that surpasses the average human driver by some (expertly agreed upon)

---

[30] Harry Collins, Robert Evans, Rodrigo Ribeiro & Martin Hall, *Experiments with Interactional Expertise*, 37 Studies in History and Philosophy of Science 656–74 (2006).

predetermined amount ought to qualify, *ceteris paribus*, as an expert driver.

One might argue that without the strong language component robots cannot qualify as experts in the same sense as human experts. This objection might be underscored by the fact that human experts can explain how they perform certain tasks whereas robots cannot, explanation being an important aspect of expertise. It is true that experts are regularly called upon to explain their decisions, so how could a robot function as an expert if it cannot provide coherent explanations?

For example, a doctor will often be asked by patients or other health professionals to explain why he is recommending a certain intervention. Explaining is a common and important function that a doctor must perform in providing expert care. However, explanations tend to function more as *examples* of expertise, *rather than as tests* of expertise. Doctors are experts *by virtue of their ongoing membership in a community of other doctors*, that is by demonstrating an ability to make appropriate medical decisions, by having been trained as doctors for many years while working alongside other doctors, by having a name diploma that says "Dr." and so on. Explanations might be demonstrations of expertise, but only one among many.

Interestingly enough, Watson was able to provide its own cursory explanations to the audience during its *Jeopardy!* stint. Watson was programmed to display its top three answer choices, ranked according to its "confidence" in each answer. Watson would only answer a question if its confidence in any one answer passed a predetermined threshold. Thus, in a weak sense, Watson was explaining its "reasoning" to the audience. Regardless, Watson's expertise would not be primarily based on its ability to provide such explanations; its rate of success at performing particular tasks would seem to do the heavy lifting where expertise is concerned. Its explanations would act as a means of providing meaningful insight into the inner workings of its unpredictable algorithms, at best. Of course, the same could be said of human explanations.

Still, the demand for detailed explanations will persist, especially when expert humans question expert robots' decisions. In some cases we will have the luxury of time in which to assess Watson's "explanations" for decisions that we question. Assessing Watson's cancer treatment predictions would perhaps be such a case. In other situations, such as operating fast-moving vehicles in complex traffic situations or choosing among alternative interventions in rapidly progressing illnesses, human experts might receive explanations, but may not be able to evaluate them in the time required to make a decision. We examine those cases in more detail below.

5. WHAT IS GAINED BY CALLING A ROBOT AN EXPERT?

We have argued that describing Watson-like robots as though they are experts, more and more, involves a descriptive accuracy that other descriptors simply lack. Alternatively, we could ignore the similarities between Watson and human experts and call Watson a mere tool; a really smart toaster.[31] But the many ways in which Watson surprises its designers with unpredictability, the extent to which Watson's code underdetermines its behavior, and the complexity and "humanness" of the tasks it successfully performs, set Watson and other robots like it apart from mere tools. So much so that we might feel more comfortable, even morally compelled, to delegate to such robots tasks traditionally carried out only by human experts. Relinquishing control and coming to rely on robots instead of humans, we speculate, might change the social meaning attached to our interactions and relations to these machines. It might one day result in our adoption of a different kind of language – especially if we are interested in getting the description right; the facts may buck, and so we may find ourselves unsaddled from more familiar talk of machines as mere tools. At the very least, as Ryan Calo suggests, we might consider these and other kinds of robots in a regulatory category apart from other tool-like machines, owing to the unique characteristics they possess.[32] Others argue that we ought to recognize a whole new ontological category for robots – not quite tools, not quite agents.[33] Indeed, to adopt mere tool-talk in reference to future Watson-like robots may require its own justification – in order to explain away the expert-like characteristics that these machines display. As they become more sophisticated, Watson's descendants will only broaden the behavioral gap between mere tools and themselves, making it ever more difficult to find comfort in tool-talk when describing them.

Although we are not proposing that Watson ought to be considered on an ontological par with human experts, critics of our view might ask: *what is gained by describing future Watson-like robots as experts rather than mere tools?* To this we respond by reiterating that we realize a philosophical gain, in that our descriptive account hits the mark in a way that tool-talk does not. Even with today's robots – to describe Watson as being

---

[31] As Richards and Smart would advise. *See* Neil M. Richards and William D. Smart, *How should the law think about robots?*, this volume, Chapter 1.

[32] R. Calo, *Robotics and the New Cyberlaw*, 103 California L. Rev. (2015), *available at* http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2402972.

[33] P.H. Kahn, et al., The New Ontological Category Hypothesis in Human-Robot Interaction, Proceedings of HRI (Mar. 6–9, 2011).

like an expert at *Jeopardy!* is to account for both its unique abilities and its social meaning in the context of *playing Jeopardy!*. Yes, Watson was designed and programmed by humans. But we say that Watson played a unique role, quite independently of its designers, in winning the game.

As discussed below, we recognize that a shift toward treating robots as experts could at the same time wreak havoc on the law by disrupting traditional legal categories. This is something regulators and policy-makers should be aware of as we continue to delegate more and more human decision-making to machines. While it may be true that treating robots as though they were experts is, by today's standards, a kind of fiction, what if tomorrow's robots shift those standards? Should we, as Wittgenstein famously pondered (in another context), stay in the saddle no matter how much the facts buck?

To the contrary, we argue that tool-talk reduces Watson-like robots to something quite other than what they are; it strips them of a legitimate descriptive richness in order to fit them into comfortable metaphors that suggest established categories of liability even though those categories may one day soon no longer be fitting. In fact, referring to Watson-like robots as mere tools, as we are driven further down the road of automation, could become the fiction. Although speculative, it is important to see that this possibility would complicate the law. It certainly complicates our practical ethics.[34] But to talk of tools in reference to tomorrow's Watsons may be to sacrifice accuracy for tradition, precision for metaphor.

6. THE NORMATIVE PULL OF EVIDENCE

What effect could the prospect of expert robots have on the question of relinquishing control of expert decision-making to machine systems? As we have defined them, we can consider a robot an expert only once there exists strong evidence that the robot is capable of consistently performing a well-defined set of tasks, tasks traditionally associated with human expertise, with results that are, on average, better than the average human expert. It stands to reason that if a robot is better at a particular set of tasks than the average human expert, we are faced with the decision of whether or not to let it perform those tasks in actual practice. Of course, that could

[34] J. Millar, *Technology as Moral Proxy: Autonomy and Paternalism By Design*, Proceedings of the IEEE International Symposium on the Ethics of Engineering, Science and Technology, (May 22–24, 2014), *available at* https://ethicstechnologyandsociety.files.wordpress.com/2014/06/millar-technology-as-moral-proxy-autonomy-and-paternalism-by-design.pdf.

mean relinquishing control to the robot, especially for time-sensitive tasks where a thorough review of the robot's decision is unfeasible (e.g., driving a car in busy traffic).

The central argument in support of delegating decision-making to expert robots comes from an understanding of evidence-based practice, which has become the gold standard of practice in healthcare and other fields of expertise.[35] Generally speaking, according to evidence-based practice, if there is good evidence to suggest that a particular action produces the most favorable outcomes, then that action is the most justifiable one. In health care, for example, evidence-based medicine is "the conscientious, explicit, and judicious use of current best evidence in making decisions about the care of individual patients."[36] A good evidence-based decision is, therefore, one that combines the individual expertise of the clinician with the "best available external clinical evidence from systematic research."[37] The underlying rationale for adopting evidence-based practice has to do with the normative pull of evidence that, for the most part, is self-evident: if the best available evidence suggests that option x is the most likely to produce desirable outcomes, then one ought to pursue option x. Indeed, the normative pull of evidence is such that to ignore the best available evidence would seem to beg questions about an individual's expertise.

Robots like Watson are meant to exemplify, and amplify, the model of evidence-based practice. Watson was designed specifically to overcome cognitive and time-related limitations that humans suffer with respect to accessing, reading, understanding, and incorporating evidence into their expert practice.[38] There is simply too much information for humans reasonably to digest, and the situation worsens as the rate of evidence production increases.[39] It is significant to recognize that, from a normative perspective, evidence suggesting a Watson-like robot can perform better at certain well-defined tasks than a human expert, is also evidence that relinquishing control to Watson is a better way of doing evidence-based practice.[40]

---

[35] D. Sackett et al., *Evidence Based Medicine: What it Is and What it Isn't*, 312 Brit. Med. J. 71 (1996).

[36] *Id.*

[37] *Id.*

[38] IBM, *supra* note 11.

[39] *Id.*

[40] Of course, this claim becomes complicated once Watson's operations transcend human comprehension, at which point the only evidence is success of outcome since we no longer understand Watson's decision-making process well enough to see it as evidence-based.

Keeping the normative pull of evidence in mind, the normative implications of expert robots become clearer. Once there are expert robots, it will be easier to argue in some instances that they *ought* to be used to their full potential, because the evidence will suggest that in those instances they will, on average, deliver better results than human experts. It will likewise be harder to argue that they ought not to be used to their full potential. That is, the normative pull of evidence will provide a strong justification for our relinquishing control of decision-making to expert robots in the same way that it suggests pharmacists ought to recommend acetaminophen in response to a high fever, rather than some other, less effective medication. Moreover, the normative pull of evidence will make it harder to choose not to relinquish control to expert robots in such cases, since evidence would suggest that keeping humans in control would increase the likelihood of undesirable outcomes.

7. HUMAN-ROBOT DISAGREEMENT

The normative pull of evidence suggests that expert robots, when they emerge, should be considered sources of decision-making authority, not merely sources of supporting knowledge to be taken into account by human experts. With respect to those tasks at which they are most expert, robots will deliver the most desirable outcomes. Of course, once expert robots emerge, we do not expect a rapid transition from human decision-making authority to robot authority, regardless of how "expert" any particular system is proven to be. Normative pull notwithstanding, humans are likely to want to remain in the saddle. But, as we have suggested, it will eventually be difficult to justify refusing to relinquish at least some control. In the following sections we evaluate several scenarios in which we cast both robot and human experts, in order to tease out the ethical difficulties of keeping humans in control of decision-making once we go down the path of expert robots.

In many situations human experts will find themselves working alongside their robot counterparts, perhaps to complement the robots' expertise with areas of expertise that remain dominated by humans. Early on, it may even be necessary to keep humans in the loop as a kind of fail-safe to prevent egregious robot errors from occurring.[41] We have referred to these situations, in which human and expert robots work alongside, as

[41] One can only imagine the kind of tragedy that might 'ensue if Dr. Watson made similar egregious errors of the sort that the *Jeopardy!*-playing Watson made

co-robotics. In co-robotics it is easy to imagine that human experts will, on occasion, disagree with the decisions made by the expert robot. Cases of disagreement between robot and human experts pose interesting situations from which to evaluate questions about which expert ought to be delegated decision-making authority. Cases of disagreement will naturally amplify our (human) concerns over whether or not we ought to relinquish control by delegating certain decisions to the machines.

Certain cases of disagreement will provide the time necessary for human experts to gather, understand the sources of disagreement, and make decisions based on an examination of the underlying rationales (both robot[42] and human) that resulted in the divergent expert opinions. Those cases will be relatively unproblematic.

Other cases will be less accommodating. As we have mentioned, cases that are time-sensitive – critical emergency room admissions, perhaps, or cases where GDCs need to make split-second decisions about how best to navigate rapidly evolving traffic situations – might afford human experts the time to disagree with the robot, but little or no time to evaluate the underlying rationales to come to anything resembling a meaningful conclusion about the sources of disagreement. In short, the human expert might have time to disagree with the expert robot, but not have time to develop a clear justification for choosing one underlying rationale over the other. Coupled with our knowledge of the evidence in favor of delegating authority to expert robots, these cases will challenge our intuitions about whether or not to relinquish control to them.

One could object that we are going over familiar ground here, that we already experience cases in which computers make mistakes (they are called bugs, or malfunctions). In such cases we are clearly justified in letting human experts override buggy or malfunctioning computers. In the cases of disagreement between Watson-like expert robots and human experts that we are concerned with, however, there is no clear malfunction: no mistaking Toronto for a U.S. airport. We are interested in examining cases of expert disagreement. Those are cases where, for example, Watson makes an expert recommendation, and the human expert makes a different one. These are cases of two experts disagreeing with one another. Though we recognize there will be cases where robot and human experts disagree, and where one will be in clear error, for the sake of this argument

---

during its match. (Watson famously referred to Toronto as a U.S. airport.) We deal with the nature of egregious robot errors more fully below.

[42] Recall Watson's ability to provide a rationale to underpin its confidence in answers.

we are trying to focus on cases of genuine expert disagreement, where we can assume there are competing rationales, rather than one rationale and one relatively straightforward case of error.[43]

It may be the case someday that expert robots are able to articulate rationales that, upon close examination even by a panel of human experts, result in a lingering disagreement between human and expert robots. In other words, it may someday be the case that robot and human experts disagree in much the same way that human experts are able to disagree with one another. Such cases, we think, will only act to make the question of when to relinquish control more pressing. This is because they will present cases where time-sensitivity plays little or no role in underscoring the nature of the disagreement. But until robots and humans can genuinely disagree, cases in which time-sensitive decisions must be made, we think, approximate genuine expert disagreement quite well, as they are cases where decisions cannot be made based on a deeper understanding of the rationales underpinning any one expert suggestion.

8. CASES OF ROBOT-HUMAN EXPERT DISAGREEMENT

In order to further illustrate the kinds of cases that matter the most, and to guide a more nuanced discussion of the normative pull of evidence-based practice in time-sensitive decision-making, let us consider four possible decision-making scenarios. In each of the cases we describe we have an expert robot working alongside a human expert – a case of co-robotics. Two of the cases are relatively unproblematic in terms of both their features and their outcomes, so we dispense with them first. In the

[43] It is worth noting that even in cases where a *post hoc* examination ends up revealing a critical error in the rationale underpinning an expert robot's decision, there may be no clear fix to the software, no obvious bug to eliminate. This is because the outcomes of Watson-like computers are unpredictable, and might not necessarily be linked to logic errors in the underlying program. The algorithms leading to the error in "reasoning" (in the weak sense) might be performing well, despite the occasional error in the output. It is quite possible that programmers would be reluctant to change any of the code for fear of causing some other problem with the expert robot's performance. These might be thought of as cases of robot "inexperience," similar to the kinds that human experts encounter when dealing with novel situations within their scope of expertise. Correcting such errors might require human experts to "train" the problematic behavior out of the robot, much like they would a mistaken human. The difference between this kind of scenario and buggy code might seem subtle, but it is not trivial.

first case, we have an expert robot and a human expert both suggesting an action that produces a "desirable" outcome, while in the second case, both experts suggest an action that produces an "undesirable" outcome. Unanimous decisions resulting in desirable outcomes are of little interest in this discussion. Similarly, unanimous decisions producing undesirable outcomes would generate little controversy of interest, as the human would appear at least as blameworthy as the robot.

Two other cases are not so straightforward. Both types describe cases of disagreement between an expert robot and a human expert. We suggest that cases of disagreement are good ones to focus on, because they draw questions of delegation, relinquishment, and responsibility into sharper focus: when robot experts and human experts disagree, to which should we delegate decision-making authority, and at what point are we justified in relinquishing control to the machines, if ever? In each of these two cases, the outcome will differ depending on which expert's suggestion is ultimately adopted.

8.1 When Expert Robots Get It "Right"

Consider a case in which an expert robot suggests an action that would produce a desirable outcome, while a human expert, contradicting the robot, suggests an action that would produce an undesirable outcome. With Watson, such a case of disagreement could be one in which Watson gets a time-sensitive diagnosis correct, while a human expert does not. With driverless cars, one could imagine a situation where the car makes a decision that would ultimately avoid an accident, whereas the human driver, if he were in control of the vehicle, would act in a way to cause (or prevent from avoiding) that accident. The outcome of this type of case – desirable or undesirable – depends on which expert's judgment is considered authoritative.

If the expert robot is granted decision-making authority, then all is well, the patient gets the intervention that saves her life, the car and driver avoid the accident, and we have a bit of anecdotal evidence bolstering the empirical evidence that our robot is, indeed, an expert. We can say that in these cases, relative to the human the expert robot "gets it right."

It is possible that a person might raise concerns about the route by which the desirable outcome was obtained. For example, a patient might question the human expert's decision to delegate to the robot and "trust the machine." But faced with such a challenge the human expert would have a standard, evidence-based, justification for his actions: "There's good evidence to suggest that Watson produces better outcomes than does his average human counterpart, so I trust it."

Interestingly, a human expert would likely have a harder time explaining his own expert judgment in this case, especially because it would have resulted in an undesirable outcome had the human expert been considered authoritative. A patient could quite reasonably ask why the human's judgment differed from the robot's, and might further question the human expert's credibility as an expert owing to the fact that the human expert "got it wrong." It is likely that if the human expert were granted decision-making authority, resulting in, say, a misdiagnosis or car crash, legitimate demands for explanations and justifications would be swift. In such a case no evidence-based justification like the one available in the previous case would be available. A decision to grant human experts authority over expert robots (i.e., experts that evidence suggests are better than humans at getting that particular job done) would seem to run contrary to evidence-based decision-making, a fact that would feature large in the ensuing debate. Reasonable demands for justification could be made on whichever individual(s) decided to grant human experts authority over expert robots, demands that would be difficult to meet.

What justification would be available for granting human expert decision-making authority over expert robots? It might be argued that it is possible to anticipate certain cases where it would be obvious to human experts that there is good evidence contradicting the expert robot's "opinion." For example, one might anticipate cases where an emergency room physician considers the expert robot's judgment and decides that it contradicts certain evidence that he has, and that he considers "clearly" the better evidence upon which to base a judgment. It could be the case that the robot was in clear error, as was Watson when he referred to Toronto as a U.S. airport. The alternative is that we have a straightforward case of *expert disagreement*, in which we have one expert judgment that is contrary to another expert judgment, both of which are evidence-based, with some underlying rationale. However, both types of disagreement – errors and expert disagreements – are going to feature experts who believe they have good reasons (perhaps evidence) that seem "obviously" to support their judgments. Without some overriding consideration upon which to base a decision, the claim that one expert's opinion is "clearly" the right one is of little help in deciding which judgment to act on. Unless there is good reason, for example, to think that one of the experts has a tendency to produce desirable outcomes more consistently than the other (perhaps a senior staff physician disagreeing with a physician that is far less experienced), then each expert's opinion could reasonably be considered as "clearly" authoritative as the other. But in cases of "equivalent expert" disagreement, that is, cases with no clear overriding considerations, we might say of an undesirable outcome that it was a simple fact of the

complexities of expert collaboration, say in the practice of medicine, where expert disagreements are common and outcomes are often uncertain.

Owing to the evidence in their favor (stipulated by definition), it is more appropriate to think of expert robots as above average in their ability to make decisions that will produce desirable outcomes. This fact suggests that granting a general decision-making authority to human experts will be problematic once expert robots are properly on the scene. It might seem justifiable to grant "override" authority to human experts in situations where there appears to be "clear" evidence contradicting the expert robot's judgment, but even this would be contra-evidence-based. Furthermore, it would beg important questions about what weight ought to be placed on claims of "clear" evidence, based on the features of human–human expert disagreements. Expert disagreements tend to be characterized by a lack, rather than excess, of clarity.

8.2 When Expert Robots Get It "Wrong"

Cases of disagreement of this sort differ from the previous cases in that the expert robot is now suggesting an action that would result in an undesirable outcome, whereas the human expert is suggesting an action that would result in a desirable outcome. The possibility of these cases of disagreement can produce curious reactions. Wallach and Allen suggest we might hold robots to higher standards than human experts, perhaps because of the fear that humans could be "overridden" by "mistaken" robots.[44] Thus, an evidence-based decision to grant expert robots decision-making authority could appear problematic because of the mere fear that a machine, rather than human, might "get it wrong." Granting a blanket decision-making authority to expert robots that we know will occasionally err (though, by definition with less frequency than humans) could, quite predictably, raise the ire of individuals negatively affected by an undesirable outcome. Perhaps, as Wallach and Allen suggest, we are more willing to accept an undesirable outcome that is the result of a "mistaken" human expert, than the same outcome that was robot generated. Though that may be the case, the question remains: Would we be justified in granting human experts decision-making authority over expert robots just because of worries that the expert robot might produce an undesirable outcome?

We think not. Undesirable outcomes stemming from a "mistaken" expert robot could be justified with an appeal to evidence. That fact cannot be overstated (despite our best efforts). Prior to knowing the outcome, the

[44] Wendell Wallach & Colin Allen, Moral Machines 71 (2009).

kinds of disagreements between human and robot experts that we are focusing on here are very similar in their features: each is a case of expert disagreement in which a time-sensitive decision must be made.[45] A decision to grant human experts blanket decision-making authority over expert robots would be to treat the expert robots on par with human experts, despite the existence of evidence that they are more likely to produce desirable outcomes.

What could be said about a decision to grant decision-making to a human expert in a case where the robot "gets it wrong"? Would such a decision not indicate that there are sometimes benefits to overriding expert robots? It would certainly seem to do just that. But the occasional benefit ought not to trump solid evidence-based reasoning. The problem is this, cases of disagreement where the human expert turns out to be right *could* be legitimate examples of human expertise outperforming robot expertise in *that* case, but if one accepts the normative pull of evidence-based practice, then they are *always* cases of *moral luck*.[46] Evidence-based practice suggests that we ought to act according to the best available evidence, and, in cases of robot–human expert disagreement, that means we ought (ethically) to delegate decision-making authority to the robots when we know that they outperform human experts. Cases in which human experts override expert robot decisions are, *ceteris paribus*, ethically problematic. That on occasion a human expert might override an expert robot's decision and produce desirable outcomes does not provide any *systematic* criterion for generating the best outcomes. Evidence-based practice, on the other hand, is meant to accomplish just that. It is only by *post hoc* analysis of cases of disagreement (or any case involving co-robotics involving expert robots) that we can assess the competing possibilities relative to one another. Prior to the outcome, that is, at the time when we are forced to make decisions, both choices look identical – there is no systematic overriding consideration upon which to base a decision other than the expert robot's evidence-based track record. When moral luck is the distinguishing factor between cases where humans override an expert robot and produce desirable outcomes, and cases where overriding the robot produces undesirable outcomes, we cannot systematically justify overriding an expert robot's decision.

Of course, one could bite the bullet and try to justify undesirable outcomes that are the result of overriding expert robot decisions. But

---

[45] As we have said, we readily acknowledge that some cases of disagreement will arise because either the robot or human is simply mistaken, perhaps "obviously" so. But these cases will be difficult to identify in the moment, and will be normatively colored by the evidence in the robot's corner.

[46] Thomas Nagel, Mortal Questions (1979), and Bernard Williams, Moral Luck (1981).

that would require an additional argument against the normative pull of evidence-based practice. We suspect such an argument would be difficult to produce. It would have the same flavor as justifying the use of one medication over another, in a time-sensitive situation, despite having evidence that the other medication would likely produce more desirable outcomes. True, things might turn out all right, but that is no justification for the decision.

9. RESPONSIBILITY

Having carefully analyzed core instances of human–expert robot disagreement, we conclude that it is not difficult to imagine a smooth and simple logic that would lead a society like ours to delegate increasingly significant decision-making to future Watson-like robots. The cases we have discussed likewise illustrate possible reasons in favor of relinquishing significant levels of control to robots that might, more and more, become understood as experts. As we have tried to demonstrate, the logic that leads us from here to there is neither revolutionary nor radical. In fact, there is a calm banality about it. Robot decision-making could be normalized in much the same way as classic Weberian bureaucratic decision-making: invoking rules, regulations, and formal authority mechanisms such as statistical reports, performance appraisals, and the like to guide performance and to regulate behavior and results.

If this is correct, it becomes difficult to conceive of innovative accountability frameworks (outside of existing institutional structures) both for preventing things from going badly wrong and for assessing liability once they do. After all, we will be told, the expert robot was just doing its job in a highly speculative enterprise not well understood by even the brightest of human experts. When thinking about what happens when things go wrong, unlike cases involving more primitive automated systems, these will generally not be cases of mere product liability. That kind of regime works well where robots are designed as mere tools, and when those tool-like robots do not do what they are supposed to, due to some kind of defect or malfunction. With Watson-like robots, there is neither defect nor malfunction in the usual sense. Nor are these situations of the sort one might imagine with near term semi-autonomous software bots – such as those that might search, procure, negotiate, and enter into contracts on one's behalf but, in doing so, exceed authority.[47] Although this latter sort

[47] For an early example of these kinds of legal problems, *see, e.g.*, Ian Kerr, *Spirits in the Material World: Intelligent Agents as Intermediaries in Electronic*

of case likewise involves the intentional adoption of a system whose future operations and outcomes are to some extent unpredictable, in those cases, the robot ultimately does something that exceeds the intentions of those who delegated control to it.

Although the trope of the robot run amok is a common dystopic theme, it is not our primary concern. The cases we are imagining are ones in which the entire point of engaging the robot is because we are limited in knowing what to do and the robot has a better track record of success than we do. Consequently, when time-sensitive decisions must be made and human and robot experts disagree, and where an undesirable outcome is the result of the decision because either the expert robot or human expert was in error, it will be difficult to assess liability in any established way.[48]

On occasion, we might draw on first principles or useful common law analogies. For example, imagine a medical center that diagnoses illnesses using a Watson-like robot. Imagine that the expert robot produces an undesirable outcome. Here, it might make sense to try to assess the liability of the hospital in a manner similar to the liability analysis that would take place if the undesirable outcome resulted from a human expert's decision. Assuming that the medical center clearly owed a duty of care to its patients, the liability question arising from a human expert's decision would be whether the human expert breached the appropriate standard of care in formulating the diagnosis. This issue would be resolved in the usual manner: divergent human experts would be called to give testimony about the diagnostic decision, explaining as clearly as possible how and why the decision was made and whether it was sound. Eventually, a judge would weigh the evidence of the competing experts and decide whether the standard of care was breached or not.

In the analogous expert robot case the chief difficulty, of course, would be in determining the appropriate standard of care for the robot. The problem is not merely that there is no preexisting standard as there might be in the case of mistaken human diagnosis. Nor is it necessarily a problem about assessing what a "reasonable robot" would have done (although that might well be a big problem!). The challenge is that it will be difficult if not impossible for anyone to offer an explanation on behalf

---

Commerce, 22 Dalhousie Law Journal 189–249 (1999). For a more comprehensive treatment of these sorts of issues, see generally S. Chopra & L. White, A Legal Theory for Autonomous Artificial Agents (2011).

[48] For a good explication of the Problem of Responsibility see: Peter Asaro, *A Body to Kick, but Still no Soul to Damn: Legal Perspectives on Robotics*, in Robot Ethics: The Legal and Social Implications of Robotics 169–86 (Patrick Lin, Keith Abney, & George Bekey eds. 2012).

of the medical center's reliance on the expert robot. This may be because the robot is not programmed to explain such things (in which case its programmers might be called upon to do their best to explain). But it could also be because the robotic algorithm is somewhat inexplicable or not likely to be fully (or even partially) understood by the human experts who built it, as could be the case with expert robots designed using the kinds of machine-learning processors described earlier. In such a case, the only evidence-based rationale involves reference to the previous track record of the robot as compared with the previous level of human success.

Here we are confronted with a paradox: the normative pull leading to a decision to delegate to the robot – namely, evidence-based reasoning – generates a system in which we now have no obvious evidentiary rationale for explaining the outcome generated by the expert robot. All we have is a hindsight case where the advice of a human expert was not followed to the detriment of the patient. Such cases make it easy to imagine fictional medical characters like Dr. Gregory House or even Dr. Leonard "Bones" McCoy of *Star Trek* eschewing expert robot decision-making, favoring the intangible qualities of human intuition and wisdom. Even though one of the two authors of this chapter is deeply sympathetic to such an approach, it is conceded that this doesn't get us very far in terms of assessing liability – especially if the robot got it right in nine out of ten such cases and the human tends to score a seven.

10. CONCLUSION

The moral of the story so far is not that lawyers should work with roboticists to ensure that future expert robots can sufficiently explain their operations in case there is a lawsuit. Although such a legal demand might have the salutary effect of providing a safeguard to ensure that co-robotics never exceeds human control, such a rule might also unduly limit the good that expert robots might one day contribute to humankind. Hence we find ourselves back where we began: wondering whether the risks associated with delegating decision-making and relinquishing human control are justified by the benefits expert robots may one day offer.

If our chapter was successful, we will have convinced our readers of at least four things: (1) there is an important and relevant sense in which robots might be understood as experts and that to understand them as merely "tools" is, more and more, descriptively inaccurate and unhelpful; (2) there is a logical impetus for delegating some expert decisions to robots; (3) cases of disagreement between human experts and expert robots generally speak in favor of delegating decision-making to the robots; (4) our

current models for assessing responsibility are not easily applicable in the case of sophisticated robot decision makers.

Many issues remain. For example, little thought has been given in this chapter to the means by which human control might be maintained alongside delegated robot decision-making and why that might be a good thing. Further thinking is also necessary in terms of how to ensure trust and reliability in situations where human control has been relinquished. We have also barely scratched the surface regarding potential liability models. These and other issues are sure to unfold as a critical mass of human experts emerges around this nascent topic. Our central aim was to provide a sufficiently rich descriptive framework and logical narrative, in order to transform what has until now been seen as dystopic science fiction, into a set of living ethical and legal concerns that are likely to emerge if the prospect of expert robots is embraced.