Delegation, Relinquishment and Responsibility:
The Prospect of Expert Robots

Jason Millar* & 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]

I. 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, it is very clear what will happen. Pretty much every contestant will dial-up the same expert: 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 counter parts of its human counterparts. Watson's win in the *IBM 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 Jeopardy!'s foremost human expert said of Watson, "I, for one, welcome our new computer overlords."[2]

Jennings' quote was prescient, if ambivalent. Watson's success raises questions about what role humans will occupy once robot experts 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 robot experts takes the form of a loss of human control.

---

* CIHR Science Policy Fellow (eHealth), and Ph.D. (Cand.) Queen's University Philosophy Department: jasonxmillar@gmail.com. The author wishes to recognize the Social Sciences and Humanities Research Council (SSHRC) and the Canadian Institutes for Health Research (CIHR) for funding research activities that
** Canada Research Chair in Ethics, Law and Technology, University of Ottawa: iankerr@uottawa.ca. The author 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 article derives. Thanks also to my 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] Asimov, I. (2004). "The Evitable Conflict." in *I, Robot.* (New York: Bantam Dell):447.

[2] This phrase was popularized in a 1994 Simpsons episode in which Kent Brockman, a local news anchor, mistakenly believes Earth is being taken over by a powerful race of giant ants. Fearing his life, he announces on air: "I, for one, welcome our new insect overlords".

This ambivalence is reminiscent of the set of concerns expressed by Isaac Asimov in *The Evitable Conflict*,[3] a wonderful short story in which he portrays the global economy managed by highly capable, utilitarian "Machines" designed to maximize human happiness. The society Asimov asks us to imagine has recognized that Machine expertise outstrips human thinking and doing on a number of fronts. In an effort to avoid perceived inevitable conflicts generated by war, famine, the environment, etc., key economic and political decision-making is delegated to the great Machines. The plot depicts regional political leaders discovering that something has very badly gone wrong, that the Machines have been "off" in their calculations. The World Coordinator, Stephen Byerley, is charged with investigating what has gone wrong. In the end, he determines that the Machines have purposely skewed the results, knowing that the humans, perceiving error, were attempting to "override" machine-based decisions. In other words, the Machines were intentionally producing "perceived errors" as a preemptive set of corrective measures for their correctly predicted human errors. At the end of the day, we learn that human delegation to the Machines did in fact render war and human strife "evitable" but, at the same time, rendered human dependency on the Machines "inevitable". Hence, for Asimov, the inevitable becomes evitable, and the evitable, inevitable.[4]

Given the benefits that expert 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 another typical 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. In that case presumably all work will be done by vast, highly organized systems of machines and no human effort will be necessary. 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. We only point out that the fate of the human race would be at the mercy of the machines. It might be argued that the human race would never be foolish enough to hand over all the power to the machines. But we are suggesting neither that the human race would voluntarily turn power over to the machines nor that the machines would willfully seize power. What we do suggest is that 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

[3] Asimov (n.1).

[4] Despite correcting them over and over, our students almost always call this story "The Inevitable Conflict", which we find most telling.

> 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. Eventually a stage may be reached at which the decisions necessary to keep the system running will be so complex that human beings will be incapable of making them intelligently. At that stage the machines will be in effective control. People won't be able to just turn the machines off, because they will be so dependent on them that turning them off would amount to suicide.[5]

These are the words of Theodore Kaczynski, better known to most as the Unabomber.[6] The passage was made famous by Bill Joy, who quoted it in its entirety in his famous WIRED essay, "Why the Future Doesn't Need Us."[7] In that essay, Joy, the former Chief Scientist of Sun Microsystems, set out his concerns about the unanticipated consequences of "GNR" technologies (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 article, we suggest that the precipice of having to decide whether to relinquish some control to expert 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, in Brad Jennings' parlance, 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 dystopic fiction, 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 must fill in some details of the backstory Asimov so cunningly plays out. Because that is roughly where we operate today, in Asimov's backstory. And so we offer a kind of logical narrative for Asimov's backstory—reason's road from here to there. In so doing we explore two important questions regarding the expert systems that will drive

---

[5] Kaczynski, T.. (2001[1995]). "The New Luddite Challenge." www.Kurzweilai.net (last accessed March 30, 2012: http://www.kurzweilai.net/the-new-luddite-challenge).

[6]Ted Kaczynski (infamously known as the Unabomber) was a mathematician and professor at the University of California, Berkeley, before he moved to a small, unpowered, cabin in the woods. Over the course of twenty years he sent letter bombs around the U.S., killing three people and injuring 23. In 1995 he sent a letter to the New York Times promising to give up terrorism in exchange for having his manifesto (known as the Unabomber Manifesto) published. In it, Kaczynski argues that human freedoms are being eroded by technologies, or industrial-technological systems, that require the large-scale coordination of human efforts, and that we should resist the development of those technologies and systems if we are to remain free and in touch with what is natural and human.

[7] Joy, B. (2000). "Why the Future Doesn't Need Us." *WIRED 8*(4). Joy learned of Kaczynski's passage from Ray Kurzweil, who also quoted this entire passage in his *The Age of Spiritual Machines*. (Kurzweil, R. (2000). *The Age of Spiritual Machines.* (New York: Penguin).)

tomorrow's robots if we take that turn: (i) at what point are we justified in relinquishing control of certain highly specialized (expert) tasks to robots?; and (ii) how would answers to the first question bear on the question of determining moral 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 expert 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*[8] might otherwise achieve. Like Asimov's World Coordinator, it is not immediately clear whether either decision is anything to worry about at all, or the end of humanity. Either way, it is our responsibility to face these difficult choices, head on, before we find ourselves wondering how we got to where we are.

II. What Kinds of Robots?

Given its expertise in building expert systems, it is perhaps not surprising that IBM has big plans for Watson. For starters, Watson will receive ears and a new voice.[9] This will allow Watson to interact with people in ways it could not (without help) during the *IBM Challenge*. With these new capacities in place, IBM is transforming Watson into a medical expert. Watson is being programmed with "clinical language understanding", a field of natural language processing focused on "extract[ing] structured, 'actionable' information from unstructured (dictated) medical documents."[10] Currently, IBM has installed its expert system at Seton Health Care Family, the top ranked health care system in Texas, in order to test Watson on the front lines of health care.[11] At the Seton facility, Watson will be analyzing troves of unstructured health data to uncover trends that can be used to make novel predictions, the goal being to improve the efficiency of health delivery, and patient outcomes. Watson is also being installed at the Memorial Sloan-Kettering Cancer Center (MSKCC) to "help doctors everywhere create individualized cancer diagnostic and treatment recommendations for their patients based on current evidence."[12] MSKCC prides itself on its model of evidence-based practice. The problem is that medical evidence is doubling every 5 years, making it difficult if not impossible for humans to stay on top of the

[8] We use the term *co-robotics* to refer to any situation in which human and robot experts are working alongside one another.
[9] On the heels of Watson's victory, IBM announced a partnership with Nuance, a company that designs sophisticated voice recognition and generation software: [Press release at: http://www-03.ibm.com/press/us/en/pressrelease/33726.wss]
[10] From Nuance's web site: http://www.nuance.com/for-healthcare/resources/clinical-language-understanding/index.htm
[11] Groenfeldt, T. (2012). "Big Data Delivers Deep Views of Patients for Better Care". *Forbes.* (Jan. 20).
[12] IBM. (2012). "Memorial Sloan-Kettering Cancer Center, IBM to Collaborate in Applying Watson Technology to Help Oncologists." (last accessed Apr. 4, 2012: http://www-03.ibm.com/press/us/en/pressrelease/37235.wss).

latest and greatest in medical knowledge, let alone incorporate it into practice.[13] Watson's ability to "understand 200 million digital pages, and deliver an answer within three seconds"[14] makes it an attractive expert robot candidate.

In the not-too-distant future, IBM will be marketing Watson's descendants to health care providers worldwide.[15] And, if they perform as well as Watson did on Jeopardy!, Watson's descendants will become the go-to medical experts. It is not hard to imagine doctors consulting Watson with the same frequency and reliance that Jeopardy! contestants would, if they could.

Watson's success on Jeopardy! stems directly from having been programmed in a very particular way—*one that makes Watson's answers unpredictable to its programmers.* Watson "knows" how to formulate questions in response to Jeopardy! clues, rather than simply being programmed to respond to a known set of inputs. Watson scours a set of data that, in theory, could span the entire Internet, for information that it deems relevant to the clues, then learns over time how best to "make sense" of that information. Watson is now capable of extracting structured, actionable information from an ever-shifting slurry of information that is out there, and apply it successfully to the problem of playing Jeopardy!. By shifting Watson's parameters towards medical information, IBM hopes that Watson will do to the world of medicine and healthcare what it has done to the world of Jeopardy!. And if this turns out to be correct, no one will be more surprised and delighted than its programmers.

If so, Watson is a forerunner to an era that Chris Anderson calls "the end of theory"[16]—an age in which we come to rely on robotic prediction machines in the place of human experts. Not because the robots' software provide 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 approach. Neither Larry nor Sergey (nor any other Google employee) knows exactly why one particular web page is better than another. But if the statistics of incoming links 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 the prototypes for prediction bots that divine without knowing. And, like the ancients, we will, quite rationally, come to rely upon them, knowing full well that we cannot explain the reasons for their decisions.[17]

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.[18] Equipped with

---

[13] Ibid.
[14] Ibid.
[15] Castillo, M. (2011). "Next for Jeopardy! Winner: Dr. Watson, I Presume?" *Time*. (February 17). (available online: www.time.com/time/business/article/0,8599,2049826,00.html).
[16] Anderson, C. (2008). "The End of Theory: The Data Deluge Makes the Scientific Method Obsolete." *WIRED*. (June 26). (Available online: http://www.wired.com/science/discoveries/magazine/16-07/pb_theory)
[17] Esposito, E. (forthcoming). "Digital Prophecies and Web Intelligence." in Mireille Hildebrandt & Ekaterina De Vries (eds.) *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, (London: Routledge).
[18] Vanderbilt, T. (2012). "Let the Robot Drive." *WIRED*. (Feb):86.

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 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."[19] But letting go may, at first, be a tougher sell for human drivers, who might harbor a nagging distrust of Googly roboticists and their toys. 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.

II. 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 principal, the complete set of information "out there" on the Internet. That set of data is constantly changing as a function of the behaviour of millions of individuals who constantly contribute new bits of information to it, the content of which is also unpredictable.

The set of targeted health information available to a Watson-like system will grow significantly as various jurisdictions implement eHealth systems (or Health Information Management systems). These digital repositories have the potential to house every piece of health information related to the individuals accessing that system, a process that is either underway or completed in most of the world. According to Kaiser Permanente, one of the largest HMOs in the US, at least three terabytes of data is generated every day as their 14M 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, for example those that paint or weld, are programmed to function in highly predictable ways. Unlike Watson, their every movement is predetermined and structured to eliminate variation between "jobs". In fact, they are monitored in such a way that they stop operating as soon as their performance falls outside of predetermined limits. Successful programming, for these robots, means that they never surprise their programmers.

Watson, on the other hand, is designed to surprise even its programmers. How can this be? Software is written in lines of code, each of which is comprehensible to the programmer, which seems to suggest that the programs are inherently predictable. True, the code can be traced stepwise, but any program operating on a vast, unpredictable set of inputs inherits

[19] Ibid.
[20] Webster, P. (2010). "Electronic health records: an asset or a whole lot of hype?" *CMAJ. 182*(4).

some of that unpredictability. Software functions that parse and operate on massive, constantly changing data stores, deliver results that no programmer can fully anticipate.

Today's examples of these kinds of software systems include, in addition to Watson, the various predictive data mining engines built into Google, Facebook, Amazon, and iTunes.[21] Those systems act upon enormous data sets to predict certain things about users in real time—the ad most likely to generate a "click", songs or books that the user will most likely want to purchase, or web pages that the user will most likely want to navigate to. Depending on the wording of an email being read by a user at a particular instant, or the few web pages 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" (those observant googlers will have noticed that Google publishes the "speed" of every search right before listing the results).

III. Expert Robots?

How else might we describe robots that are capable of performing complex and unpredictable tasks?

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.[22] 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.[23] That Watson-like computers similarly cannot be fully understood by reference to their lines of code opens the door to describing them as *robot experts*.

---

[21] Kerr, I. (forthcoming). "Prediction, Presumption, Preemption: The Path of Law After the Computational Turn" in Mireille Hildebrandt & Ekaterina De Vries (eds.) *Privacy, Due Process and the Computational Turn: The Philosophy of Law Meets the Philosophy of Technology*, (London: Routledge).

[22] Collins, H., Evans, R.. (2007). *Rethinking Expertise.* (Chicago: University of Chicago Press). 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: Collins, H.. (1985). *Changing Order.* (Chicago: University of Chicago Press). See also: Collins H., Weinel, M.. (2011). "Transmuted Expertise: How Technical Non-Experts Can Assess Experts and Expertise." *Argumentation 25*: 401-413.

[23] In STS, explanations of underdetermination are based on a particular interpretation of Ludwig Wittgenstein's work on game theory (see Collins, n. 11 for a fuller explanation). Wittgenstein famously demonstrated that the rules for continuing a mathematical series, say {1,2,4,8,…}, could always be "misinterpreted" by a clever (or stubborn) enough student. Wittgenstein seems to argue that it is impossible to generate a set of rules that would guarantee the correct continuation of a series, because the "correct" way of doing so is grounded in the social practice of doing it, not in the rules.

Some of the pioneering work linking underdetermination with expertise comes from Harry Collins' observations made while studying scientists at work in laboratories. A classic example of scientific rule following involves scientists trying to reproduce other scientists' experiments. One way of attempting to reproduce an experiment is to read the original methodology (the rules), as it is described in the published scientific paper, and attempt to do what the other researchers did. In writing up a lab report in the form of a scientific paper, scientists, one might assume, should be able to transcribe the expertise they employed in completing the experiment into a set of rules, or instructions, that could then successfully be followed by any another scientist with sufficient knowledge in the area under study. Collins notes, however, that when scientists attempt to reproduce experiments, scientific papers and lab notes—the rules—are typically insufficient for successful reproductions.[24] Collins describes how other scientists, using only those resources, are unable to design replicas of the original apparatus, unable to reproduce the original methodology, and therefore unable to reproduce the original results.[25] The set of rules, he concludes, do not adequately determine the actions required to demonstrate expertise in carrying out *that* experiment. The rules are not where the expertise resides.

Perhaps surprisingly, this inability to reproduce an experiment is typically not taken as a sign of having refuted the original findings, nor is it interpreted as a lack of scientific expertise; rather it is typically taken as an indication that more information is needed. In other words, there is an understanding among scientific experts that descriptions of how to do an experiment are insufficient for actually doing it. The kind of information needed to "fill in the blanks" comes via direct mentoring: scientists working with other scientists in order to understand how actually to do the experiment.[26] In fact Collins & Evans's claim is that only by *doing* can one develop high levels of expertise in any field, and that 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 with respect to a particular set of knowledge and/or tasks. 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*.

There are numerous examples of tacit knowledge at work in human expertises. Though it qualifies as an example of what Collins & Evans call *ubiquitous* expertise (an expertise shared by a relatively large group of people), rather than *specialist* expertise (an expertise shared by a relatively small group of people), the ability to speak a language is a good example of tacit knowledge at work. The many failures of the artificial intelligence community to produce a computer that can converse in natural language can be taken as an indication of the considerable amount of tacit knowledge involved in the use of natural language.[28] Collins and Evans also point to the "art" of driving, in which one *has a sense* of

[24] Collins, H. (n. 11).
[25] Ibid.
[26] Ibid. Also, Collins & Evans. (n. 11).
[27] Collins & Evans. (n. 11):13.
[28] Selinger, E., Dreyfus, H., Collins, H.. (2007). "Interactional Expertise and Embodiment." *Studies in History*

what is going on with the car, the traffic, the road conditions, and so on, without really being able to articulate that *sense* in any meaningful way.[29] Expert drivers can feel the ground through the various steering mechanisms, while the very expert driver knows when the car is losing traction and about to go into a skid. No set of descriptive rules or instructions can seem to express that kind of knowledge—the knowledge is in the driver if anywhere.

Steven Epstein provides a detailed account of how AIDS activists in the late 1980s developed a level of expertise that allowed them to interact with medical researchers at a very high level.[30] Among other forms of tacit knowledge, they developed the ability to discriminate between credible and incredible sources of knowledge, a skill that is associated with very high levels of expertise. Experts working in a particular community of specialist expertise, say pathophysiologists, will tend to "accept" or "reject" claims made in papers dealing with their specialist area, for reasons that are not apparent to the untrained observer, and that can be very difficult to describe.[31] Indeed, experts in one group will sometimes accept or reject claims made by similar kinds of experts in another (geographically or culturally separated) group for reasons that are not easily articulated.[32] The ability to discriminate is sometimes described as an effect of experts' socialization into communities of practice, as those communities will tend to accept or reject particular claims along group boundaries.[33]

When we turn to the question of robot expertise 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 are capable of forming communities of expertise in the strong linguistic sense emphasized by Collins & Evans—robots will not be mentoring other robots in the near future[34]—we can locate a weak form of tacit knowledge among robots that seems to qualify them as robot experts.[35]

---

and Philosophy of Science 38: 722-740.

[29] Collins & Evans (n. 11).

[30] Epstein, S. (1995). "The Construction of Lay Expertise: AIDS Activism and the Forging of Credibility in the Reform of Clinical Trials." *Science, Technology & Human Values 20*(4):408-437.

[31] Collins. (n. 11).

[32] Casper, M., Clarke, A.. (1998). "Making the Pap Smear into the "Right Tool" for the Job: Cervical Cancer Screening in the USA, circa 1940-95." *Social Studies of Science 28*:255.

[33] Ibid. Also, Epstein (n. 17); Casper & Clarke (n. 20). Thomas Kuhn also made similar claims in his famous description of "worldviews", within which, he claimed, scientists tended to accept claims that fit unproblematically with their particular worldview, and reject those that don't. Of course, understanding which claim fits and which doesn't takes a considerable amount of tacit knowledge, since "fitting" does not appear to reduce to mathematics or theory-based considerations: Kuhn, T.. (1962). *The Structure of Scientific Revolutions*. (Chicago: University of Chicago Press).

[34] This is still the stuff of Asimovian science fiction. See, e.g., 'That Thou Art Mindful of Him" in: Asimov, I. (1983). *The Complete Robot.* (New York: Harper Collins).

[35] John Searle's famous argument against strong artificial intelligence (see Searle, J. (1980). "Minds, Brains, and Programs. *Behavioral and Brain Sciences 3*(3): 417-457) is a good example of what we mean here. We do not claim that robots have intentions (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

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.[36] In stark contrast to most humans, Watson and the GDC cannot enter into traditional, language-based apprenticeships with other human experts. That is, they cannot have conversations with humans, and do not understand (in the strong sense of the term) what they are doing. 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, which allows the robot to form new associative rules between data points, and subsequently affects the machine's outputs. This suggests that human experts are able, in a weak sense, to help the robots understand how to, and how not to, improve their functioning with respect to specific tasks usually associated exclusively with human expertise, such as winning at Jeopardy!, or driving in rush hour traffic on the freeway. Though the mentoring is not bidirectional and language-based, it is directed at improving the robots' performance beyond the rank of novice.

Watson and the GDC are also able to independently extract meaningful information from large sets of unstructured information in a way that human experts interpret as "correct", or "incorrect". In other words, they are able to apply their expert "knowledge", developed through a (weak) form of mentoring, to novel situations. When Watson correctly interprets the meaning of a subtle play on words, indicated by its answering a Jeopardy! question correctly with a high degree of confidence, Watson is communicating *something*. Though it is not sophisticated in its language, the information that gets communicated is not merely 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 to highly predictable programs with relatively constrained sets of inputs and outputs, then we would be inclined to describe them merely as sophisticated coffee makers.[37] But Watson and the GDC are able to achieve high degrees of "expertise" by acting on sets of rules (their programs) that underdetermine their success. Thus there is a gap to be filled if we are to explain how it is that Watson-like robots do what they do so well.[38] Like human experts, we suggest that

---

same way that a human expert would from the recipient's perspective.

[36] Collins & Evans, (n. 11); Collins, (n. 11); Collins & Weinel, (n. 11); Casper & Clarke, (n. 20); Kuhn, (n. 21).

[37] This is not a slight against expert baristas, for whom one author of this article has incredible respect. Specialist tacit knowledge is undoubtedly what resulted in the many wonderful espressos he had during his last visit to Italy, which he reminisces about often, and which have resulted in his cursing the many "robotic" espresso machines currently employed in too many North American coffee bars.

[38] That gap was expressed nicely during the *Jeopardy! IBM Challenge*, in a video IBM aired featuring Dr. Jennifer Chu-Carroll, one of Watson's many creators, reacting to one of Watson's many correct answers. In reaction to Watson her look was one of disbelief—eyes popping, jaw dropped. She later commented, "Watson surprises me every time it plays." (Chu-Carroll, J. (2011). *Jeopardy!* (Episode originally aired February 14, 2011).)

what fills that gap between the programs and Watson's abilities is, in a weak sense, specialist 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 actions, is enough to qualify it, if only in a weak sense, as an expert. Thus, we argue that it makes sense to talk of expert robots.

When can we call a robot an expert?

Humans become experts when they are accepted as experts by other human experts. According to Collins, Evans and Ribeiro[39], this requires novices to pass Turing-like language tests[40] in communities of experts. Take academic expertise as an example. Recognized experts (university professors) interact with candidate experts (doctoral students) until such time as those recognized experts confer the status of expert on the candidates. This requires the candidate experts 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, 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 become experts in their own right.

Similarly, we suggest that a robot is an expert only when there is strong evidence it 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. Thus, 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) 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. But is it reasonable to consider explanations and justifications as the standard of expertise? Imagine you had a human who could perform certain tasks as well as or better than other experts, but who was unwilling or for some unknown reason unable to provide an explanation for any of her actions. For example, imagine you have a human who is clearly capable of learning the

---

[39] Collins, H., Evans, R., Ribeiro, R., Hall, M.. (2006). "Experiments with Interactional Expertise." *Studies in History and Philosophy of Science 37*: 656-674.

[40] Turing's original test, which he called the Imitation Game, was meant to answer the question, At what point can we say that a machine is intelligent? Turing claimed that we can make such a claim when a computing machine is able to win an imitation game. The imitation game involves a human interrogator, a human contestant, and a machine contestant. The interrogator's job is to submit a series of text-based (written) questions to each of the contestants to determine which is the computer and which is the human. Each of the contestants are trying to convince the interrogator into thinking it is the human. (See, Turing, A.M. (1950). "Computing Machinery and Intelligence." *Mind 59*: 433-460). Collins, Evans and Ribeiro argue that questions of expertise are settled by variations on the Turing Test, in that candidate experts must convince credentialed experts, via detailed language-based interactions, that they qualify as experts in the relevant sense.

tasks associated with a particular area of expertise, but who chooses not to communicate with others for some unknown reason. What would provide the basis for calling her an expert at those tasks? It would seem to be her ability to perform those tasks in the manner of other experts, rather than her ability to explain what she is doing, a requirement that has already been ruled out by the recognition of tacit knowledge. Experts often are unable to explain how they do what they do, but they do it nonetheless.

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 course of treatment, or how he came to a particular diagnosis. Explaining is a common, and important, function that a doctor must perform in providing expert care. In dealing with laypersons, however, expert explanations will rarely require more than a cursory reference to specialist knowledge, maybe some statistical evidence, in order to qualify sufficiently as an explanation. It is true that, in the Internet age, more and more non-experts are armed with greater and greater amounts of medical information. This certainly allows them to ask better and more intelligent questions of medical experts. That said, laypersons are generally no better off in their ability to question the *expertise* of the doctor—usually, they are merely seeking evidence of some underlying rationale. The doctor is expert by virtue of his being a doctor, having been trained as such for years and years while working in a clinic, having a name badge that says "Dr." and so on. In other words, explanations might be demonstrations of expertise, one among many, but requests for explanations are not challenges to an individual's *expertise*, especially when issued by laypersons.

Interestingly enough, Watson was able to provide its own cursory explanations to the audience during the 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. Presumably, Watson could be programmed to add complexity to its explanations, maybe by providing a chain of concepts that led it to particular answers, by providing statistical models of its weighing of data, or by providing documents that it weighted heavily in its expert "deliberations". 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 that work. Its explanations would act as a means of providing meaningful insight into the inner workings of its unpredictable algorithms, at best. Of course, it could be argued that human explanations function similarly.

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, we will not have the luxury of time; in time-sensitive applications such as operating fast-moving vehicles in complex traffic situations, or choosing between 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.

One might still argue that the human is the only possible expert because the human understands, in the strong sense, what he or she does. Again, this seems to set the bar too high for expertise, since it adds something on top of what seem to pass unproblematically as examples of expertise, namely the performing of expert tasks at a very advanced level, as judged by other experts.

IV. 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, robots are experts 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 an expert at something, and if it is better at that task than the average human expert, we are faced with the decision whether or not to let it perform those tasks in actual practice. Of course, that could 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 decision-making in health care and other fields of expertise.[41] 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."[42] 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."[43] The underlying rationale for adopting evidence based practice has to do with the normative pull of evidence which, 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.[44] Indeed, to ignore the best available evidence would seem to beg questions about an individual's expertise.

---

[41] Sackett, D., Rosenberg, W., Gray, J., Haynes, R., and Richardson, W.. (1996). "Evidence based medicine: What it is and what it isn't". *BMJ,* 312:71

[42] Ibid.

[43] Ibid.

[44] It is worth noting that in health care the choice of "most desirable outcomes" might vary depending on whose perspective you are assuming. A patient's perception of what is most desirable could vary significantly from a physician's. For the sake of this argument we are assuming that the context will dictate the definition of "most desirable outcomes": in health care it will tend to fall on the patient to ultimately decide what is most desirable, while in driving a car it will tend to be something like whatever option is least likely to result in an accident. Therefore, we do not take a particular view on what count as the most desirable outcomes in any expert decision-making context.

Robot experts 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.[45] There is simply too much information for humans reasonably to digest, and the situation worsens as the rate of evidence production increases.[46] It is significant from a normative perspective that 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.[47]

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 robot experts 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.

V. Human-Robot Expert Disagreements

The normative pull of evidence suggests that expert robots 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 human's in the

[45] IBM. (n.12).
[46] Ibid.
[47] 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.

loop as a kind of fail-safe to prevent egregious robot errors from occurring.[48] We have referred to these situations, in which human and robot experts work alongside, as co-robotics. While working alongside expert robots, 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[49] 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 robot expert, 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 favour 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. They 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 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.[50]

---

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

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

[50] It is worth noting that even in cases where a *post hoc* examination ends up revealing a critical error in the rationale underpinning a robot expert'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 robot expert's performance. These might be thought of as cases of robot "inexperience",

It may be the case that someday robot experts are able to articulate rationales that, upon close examination even by a panel of human experts, result in a lingering disagreement between human and robot experts. 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.

VI. 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 scenarios of decision-making. Table 1 describes some possible cases of agreement and disagreement between robot and human experts. In each case we have a robot expert working alongside a human expert, a case of co-robotics. Type I and Type IV cases are relatively unproblematic in terms of both their features and their outcomes. Both Type I and Type IV cases see agreement (==) between the robot expert (R_E) and the human expert (H_E). In Type I cases, both experts suggest an action, D, that produces a "desirable" outcome. In Type IV cases, both experts suggest an action, U, which produces an "undesirable" outcome.[51] Decisions resulting in desirable outcomes would tend not to result in anyone spending much time trying to assign blame or determine accountability. So Type I cases are of less interest in this discussion. Similarly, Type IV cases, in which a human and robot expert agree on doing something that produces undesirable outcomes, would generate little controversy, as the human would appear at least as blameworthy as the robot.

Type II and III cases are not so straightforward, and so deserve some more detailed examination. Both types describe cases of disagreement between robot and human experts. We suggest that cases of disagreement are good ones to focus on, because they draw questions of delegation and relinquishment into sharp focus: when robot 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 types of case, the outcome will differ depending on which expert's suggestion is ultimately adopted.

---

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 non-trivial.

[51] We are intentionally vague in our use of "desirable" and "undesirable." They are meant to indicate only that the outcome is one that would, in the case of D, not generally result in controversy, or a situation in which people sought to determine accountability, attribute blame and so on, whereas outcome U would generally result in those things. Our use of "desirable" and "undesirable" could also be cases of *more* or *less* desirable respectively, since a *post hoc* realization that a more desirable outcome was available could result in similar kinds of accountability seeking.

Type I	**Type II**
R_E == H_E Agree to do D Outcome == D	R_E ≠ H_E R_E suggests D H_E suggests U Outcome == ?
Type III	**Type IV**
R_E ≠ H_E R_E suggests U H_E suggests D Outcome == ?	R_E == H_E Agree to do U Outcome == U

Table 1: Cases of R_E-H_E Agreement and Disagreement

Note that we are starting from the position that we do not know which expert to consider authoritative. It is probably still correct to claim, for now, that humans are the default authority. But that is an accident of history—robot experts will be competing for jobs that are already filled, so to speak. We are interested in determining at what point that accident becomes unjustifiable, in other words, at what point ought we (humans) to relinquish control, if ever. In order to make that determination we start by clearing the historical slate. That way we can make the determination with a bit less baggage.

So, let us consider in each of the imagined cases that we have a robot expert that, according to the evidence, outperforms its average human counterpart when performing a particular set of expert tasks. The point at issue is just this: If faced with making a time-sensitive expert decision, which expert ought to be granted decision-making authority?

A. Type II Cases of Disagreement

In a Type II case we have a robot expert suggesting an action, D, which would produce a desirable outcome, and a human expert suggesting an action, U, which would produce an undesirable outcome. In the case of Watson, a Type II case of disagreement could be one in which Watson gets a time-sensitive diagnosis correct, while a human expert does not. Similarly with GDCs, 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. Of course, the outcome of a Type II case depends on which expert's judgment is considered authoritative.

If the robot expert 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 scientific evidence that our robot expert is, indeed, an expert. Let's call this a Type II_D outcome (the "$_D$" indicating a desirable outcome, see Table 2).

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 "trust the machine". In that case 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."

Interestingly, a human expert would likely have a harder time explaining his own expert judgment in a Type II case, especially because it would have resulted in an undesirable outcome had the human expert been considered authoritative. A patient could reasonably ask why the human's judgment differed from the robot's, and might further question the human expert's credibility as a result.

Desirable Outcome	**Undesirable Outcome**
Type II_D R_E Considered Authoritative Evidence-Based Decision	**Type II_U** H_E Considered Authoritative Contra-evidence-based decision
Type III_D H_E Considered Authoritative Contra-evidence-based decision	**Type III_U** R_E Considered Authoritative Evidence-based decision

Table 2: Cases of disagreement between R_E and H_E.

It is likely, on the other hand, that if the human expert is granted decision-making authority, resulting in a misdiagnosis or car crash, legitimate demands for explanations and justifications would be swift. In such a case, let us call it a Type II_U outcome (the "$_U$" indicating an undesirable outcome), no evidence-based justification like the one available following Type II_D outcomes would be available. The fact of a decision to grant human experts authority over robot experts, experts that evidence suggests are better at getting that particular job done, would feature large in the ensuing debate. It would place a considerable demand on whichever individual(s) decided to grant human experts authority over robot experts, to justify their decision.

What justification would be available for granting human expert decision-making authority over robot experts? 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 robot expert's "opinion". For example, one might anticipate cases where an emergency room physician considers the robot expert'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. Both types of disagreement—errors and expert disagreements—are going to feature human experts who believe they have good reasons (perhaps evidence) that seem

"obviously" to underpin their judgment. 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 (say we have a senior staff physician disagreeing with a physician that is far less experienced), then each expert's opinion could reasonably said to be as "clearly" authoritative as the other. In such cases of "average human expert" disagreement, we might say that a Type II_U outcome is unfortunate, but a fact of the complexities of expert collaboration, say in the practice of medicine, where expert disagreements are common and outcomes are often uncertain.

Still, one might resist relinquishing control to robot experts precisely because they can make straightforward errors; critics will want to emphasize that Toronto is "clearly" not a U.S. airport. We accept that cases of error will arise, but can say only that human experts will need to make a snap judgment in the moment whether or not to accept the decision-making authority of a robot. The features may seem clear enough, but distinguishing cases of error from cases of genuine expert disagreement will not necessarily be so easy, especially in time-sensitive decision contexts. In all cases the evidence in favor of Watson will always feature in the backdrop of the moment. If things turn out well, everyone breathes a sigh of relief.[52] If they don't, justifications for not relinquishing control to the robot could end up looking themselves like cases of human experts making mistakes, as they often do.

Owing to the evidence in their favor, it is more appropriate to think of robot experts 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 robot experts 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 robot expert'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.

B. Type III Cases of Disagreement

Type III cases of disagreement differ from Type II cases in that the robot expert 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. Understanding the possibility of Type III cases of disagreement can lead to curious reactions. Wallach & Allen suggest we might hold robots to higher standards than humans, perhaps because of the fear that humans could be "overridden" by "mistaken" robots, in other words, the fear of Type III_U outcomes.[53] Thus, an evidence-based decision to grant robot experts decision-

[52] See also our discussion of moral luck in Type III cases of disagreement.
[53] Wallach, W. Allen, C. (2009). *Moral Machines.* (Oxford: Oxford University Press): 71.

making authority could appear problematic because of Type III cases of disagreement. A general robot expert decision-making authority could result in Type III_U outcomes that, quite predictably, would raise the ire of individuals negatively affected by a Type III_U outcome. Perhaps, as Wallach & 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 computer generated. Though that may be the case, the question remains: Would we be justified in granting human experts decision-making authority over robot experts just because of worries over Type III_U outcomes?

We think not. Unlike the case of Type II_U outcomes, Type III_U outcomes could be justified with an appeal to evidence. That fact cannot be overstated (despite our best efforts). Prior to knowing the outcome, Type II and Type III cases of disagreement are very similar in their features: each is a case of expert disagreement in which a time-sensitive decision must be made.[54] A decision to grant human experts general decision-making authority over robot experts would be to treat the robot experts on par with human experts, despite the existence of evidence that they are more likely to produce desirable outcomes.

What of the Type III_D outcomes? Do they not indicate that there are benefits to overriding robot experts? Again, we argue that the answer has to be 'no'. 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*.[55] 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 robot expert decisions are, *ceteris paribus*, ethically problematic. That, on occasion, a human expert might override a robot expert decision thus producing desirable outcomes (the Type III_D outcome) does not provide any systematic guidance for generating the best outcomes. Evidence-based practice, on the other hand, is meant to accomplish just that. It is only in a *post hoc* analysis of Type III cases of disagreement (or any case involving co-robotics) that we can assess the outcomes relative to one another. Prior to the outcome, that is, at the time when we are

[54] 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.

[55] Moral luck is a philosophical concept, famously discussed by Thomas Nagel and Bernard Williams, which challenges our intuitions about how we go about morally evaluating actions. Suppose we have two individuals who attempt to murder a third person. Suppose the first would be murderer attempts and fails because the victim slips and falls so that the bullet intended for him misses. Suppose the second would be murderer is successful in killing the victim. One of our intuitions (and our practices) suggests we should seek a harsher penalty for the successful murderer, even though the only thing differentiating the two criminals seems to be the luck that made the first unsuccessful. As far as their intentions and actions go, each wanted to murder the victim. This kind of luck, which causes us to morally evaluate situations differently, is referred to as moral luck. Moral luck is often considered problematic because we would tend to argue that a person's actions should not be evaluated based on things over which he has no control. In other words, our conflicting intuition seems to suggest that we should not evaluate actions differently just because of moral luck. See: Nagel, T. (1979). *Mortal Questions.* (New York: Cambridge University Press), and Williams, B. (1981). *Moral Luck*. (Cambridge: Cambridge University Press).

forced to make decisions, both actions look identical—a human assumes authority despite the evidence that the robot will tend to make better decisions. Calling Type III_D outcomes justifiable forces us to consider what differentiates our assessment of them from Type II_U outcomes, after which we would question the human expert's decision, and the answer seems to be moral luck.

Of course, one could bite the bullet and try to justify Type II_U outcomes. But 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 knowing of 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.

VI. Responsibility

Having carefully analyzed core instances of human-robot expert 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 robots. These cases likewise illustrate possible reasons in favour of relinquishing significant levels of control to those robots. 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—in the sense that Hannah Arendt once used that term.[56] 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 behaviour 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. Such cases occur when a robot does not do what it is supposed to, due to some kind of malfunction. Here, there is no 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.[57] Although this latter sort of case likewise involves the intentional adoption of a system whose future operations and outcomes are to

[56] In her famous book *Eichmann in Jerusalem: A Report on the Banality of Evil* (New York: Viking Press, 1963), Hannah Arendt examines the idea that evil is not always the result of egregious moral wrongdoing or the insane acts of fanatics or sociopaths, but is often carried out by ordinary people who accept bureaucratic rules in an unquestioned manner and therefore participate with the view that their actions are normal.

[57] For an early example of these kinds of legal problems see, e.g., Kerr, I. (1999). "Spirits in the Material World: Intelligent Agents as Intermediaries in Electronic Commerce" *Dalhousie Law Journal 22:*189-249. For a more comprehensive treatment of these sorts of issues, see generally Chopra, S. and White, L. (2011). *A Legal Theory for Autonomous Artificial Agents* (Ann Arbour: The University of Michigan Press)

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 do not know 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 robot expert or human expert was in error, it will be difficult to assess liability in any meaningful way.[58]

On occasion, we might draw on first principles or useful common law analogies. For example, imagine a medical centre that diagnoses illnesses using a Watson-like robot. Imagine that the robot generates a Type III_U decision. Here, it might make a lot of sense to try to assess the liability of the hospital in a manner similar to the liability analysis that would take place if it had been a Type II_U decision. Assuming that the medical centre clearly owed a duty of care to its patients, the liability question in a Type II_U case 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 Type III_U 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 the "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 of the medical centre's actual robot expert. This may be because the robot is incapable of explaining such things (in which case its programmers might be called upon to do their best to explain). But it is more likely because there is an explanatory gap that exists between a robot's algorithm and its functioning that is largely inexplicable or not likely to be fully (or even partially) understood by the human experts who built it. In such a case, the primary evidence-based rationale involves reference to the previous track record of the robot as compared to the previous level of human success.

Here we are confronted with an Asimovian paradox: the normative pull leading to a decision to delegate to the robot —namely, evidence based reasoning—generates a system in which we might no longer have any 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 medical characters like Dr. Gregory House or even Dr. Leonard "Bones" McCoy

[58] For a good explication of the Problem of Responsibility see: Asaro, P. (2012). "A Body to Kick, but Still no Soul to Damn: Legal Perspectives on Robotics." In (P. Lin, K. Abney, and G. Bekey) *Robot Ethics: The Legal and Social Implications of Robotics.* (Cambridge: MIT Press): 169-186.

eschewing robot decision-making, favouring the intangible qualities of human intuition and wisdom. Even though one of the two authors of this article 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 wasn't even in the general ballpark.

VII. Conclusion

The moral of the story so far is not that lawyers should work with roboticists to ensure that 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 do for humankind. Hence we find ourselves right where we began: wondering whether then risks associated with delegating decision-making and relinquishing human control are justified by the benefits expert robots may one day offer.

If our article was successful, we will have convinced our readers of at least four things: (i) there is an important and relevant sense in which robots can be understood as experts; (ii) there is a logical impetus for delegating some expert decisions to robots; (iii) cases of human-robot expert disagreement generally speak in favour of delegating decision-making to the robots; (iv) our current models for assessing responsibility are not easily applicable in the case of expert robots.

Many issues remain. For example, little thought has been given in this article 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 logical narrative in order to transform what has till 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.