NEWSLETTER | The American Philosophical Association

Philosophy and Computers

FALL 2014 VOLUME 14 | NUMBER 1

ISSN 2155-9708

APA NEWSLETTER ON

Philosophy and Computers

PETER BOLTUC, EDITOR

VOLUME 14 | NUMBER 1 | FALL 2014

FROM THE EDITOR

Scavenger Science

Peter Boltuc

AUSTRALIAN NATIONAL UNIVERSITY

We have better and better reasons to believe that computer engineering and artificial intelligence broadly understood—robotics, genetics and other kinds of engineering—are the most likely areas to become the transformative science of the near future. Most people look back at the transformative science of the past, exemplified by Newton, Darwin, Einstein, or Heisenberg, and they look towards the new generation of super-colliders to bring in the next scientific revolution.[1] I think we should look closer; if we do so, we may notice an even greater opportunity emerges in front of us. Let us call this new area the "scavenger science."

Having plenty of data seems like a good thing; sometimes it is. Yet, one of the main weaknesses of contemporary society—from science labs, through marketing companies, to security agencies—is that we have too much data. Today we learn how to manage this abundance of good things. Big data mining shall allow us to utilize information, including scientific information, from the hips of previously discarded or overlooked data. One of the main problems is that even the best minds process only a fraction of new data available in their discipline, often those published in the top dozen journals. Research that did not make it to those top publications, including 99 percent of research that is published in languages other than English, goes unnoticed. Who gets published in those top journals follows a pattern of peer-review that rejects questionable approaches but also, oftentimes, most of the ground-breaking ones, especially if they come from younger scientists or those from less central institutions.[2] Semantic computing shall allow computerized search engines to get through the tons of articles published out of the main light, in smaller journals, maybe as university pre-prints, in other countries or languages. It will compare, even analyze, those data and ideas in a relatively impartial manner. Furthermore, discovery machines can generate entirely new ideas, both applied and theoretical, some of them so counterintuitive that no human brain would have ever generated them.[3]

The scavenger science shall allow for radically new, truly transformative discoveries. We may discover facts that occur with small frequencies that, for reasonably small samples (of a few hundreds, even thousands), are within the margin of error, but for hundreds of thousands of observations would become measurable, if discrete, regularities. We may be able to track patterns that appear only in hard-to-predict conditions, including environmental medicinal interactions. Some such research may teach us things that current science has no idea about, or even revisit the hypotheses explicitly rejected for lack of good evidence. There shall also be research conducted that exceeds human capacity to follow—at least the details—which some people, inspired largely by sci-fi, fear as the *singularity*. This powerful mega-research may still lead to different conclusions, depending on varieties of assumptions or methodologies (in particular, we may input different levels of *reality function* in imagitrons).

Let us consider a super-thinking-network based on imagitrons set within narrow disturbance frequencies of its reality function (we may call this system a Mega-Dennett) and another super-thinking-network based on imagitrons that work within broader frequencies of disturbances that lead to a more active creativity function (call it a Mega-Chalmers). Even if both networks work with the same information they shall still come up with different theories, as well as practical recommendations. Moreover, since those networks are also search (and maybe re-search) engines they would amass different sets of information; they would also follow different initial patterns of organizing such information (initial scientific theories). Undoubtedly, Mega-Dennett would lean towards more reductive explanations, would be more likely to view the demarcation problem as methodologically essential, and would have a tendency to razor off the pieces of information radically inconsistent with the leading theories. Mega-Chalmers, on the other hand, would be likely to build theories on the basis of merely statistical evidence and inference to the best explanation. Mega-Chalmers would be less likely to reject theories that pass a relatively low threshold of statistical confirmation. It might investigate inter-relations among such theories and compute conditional probabilities—of what initially looks like marginal theories—conditional on acceptance of the other low-probability theories (the theoretical context). Mega-Chalmers would be less keen on the problem of demarcation and invest some of its information-processing power to investigate some measurable interpretations of theories generally rejected by science; perhaps human ability to affect slight anomalies in randomizers, or even some version of astrology. Most of those attempts would turn out unsuccessful; yet the few that turn out to work out may give rise to ground-breaking changes. One last question emerges: Would it be better to build a Mega-Joe, a super-thinking-network with the reality function set just in between Mega-Dennett and Mega-Chalmers? For the development of *normal science*, especially engineering

innovations, this may indeed be the way to go. If, on the other hand, we care about generating ground-breaking transformative science, the Mega-Dennett/Mega-Chalmers combo may do a better job. Those familiar with the architecture of Thaler's discovery machines would see the analogy between the role of the sub-routine creatively generating rational permutations of the world (analogous to Mega-Chalmers) and the sub-routine selecting the subset of those images that make sense in accordance with narrower criteria (analogous to mega-Dennett). Depending on details of the cognitive architecture of the future super-thinking-networks, this complementarity of functions may need to be preserved in more sophisticated cognitive engines.

***

The above view on the transformative philosophical role of artificial cognitive architectures is perhaps the most ambitious framework in which philosophy and computers may operate. Let us see what developments the current issue of this newsletter brings.

In his note on the models for machine ethics, Tom Powers argues that machine ethics, in fact any ethics, should be testable on robots. This vital argument follows up on the project developed by Jim Moor and other philosophers trying to clarify ethical recommendations so that the ethical framework can be implemented in a consistent autonomous robotic agent for civilian or military purposes.[4]

The featured article by John P. Sullins—a distinguished member of our committee—includes a thorough analysis of the ethical trust which patients and doctors put in machines in the context of robot-assisted surgery. Sullins emphasizes the special, intimate status of the relationship between doctors and patients, which is clearly not replicable in relation to a machine. He also discusses, with great competency and sensitivity to detail, the consequences of frequent use of surgical robots that is likely to follow in the areas such as apprenticeship, professional development, and training. Sullins concludes his article with a note of cautious optimism. He notes the many benefits of robotic surgery but also the ramifications of this project for surgery ethics. I also want to turn the reader's attention to the author's brief discussion of the issues of future medical technology, such as human enhancements.

Applied ethics is also the topic of two papers that follow, both related to recent work by Luciano Floridi. In her elegant article, Mariarosaria Taddeo discusses the ethics of information warfare. Taddeo proposes Floridi's ethics of information as a way to bridge the ontological gap between the old-style understanding of harm as mainly physical and the approach of taking into account lost information and operational capacities, which seems more appropriate for assessing informational warfare. Under the proposed approach, information is viewed as an entity with independent ontological and moral status. The ethics of information presents a consistent lowest common denominator that allows comparisons among different morally valuable objects through the moral currency of informational losses. The topic of Floridi's ethics of information continues in the interesting article of Pompeu Casanovas. The author combines erudition from a number of fields that include intricate knowledge of philosophy of science, philosophy, and computer programming. Casanovas engages in a delightful dialogue with various ideas of the book. He ends with an important (yet somewhat unexpected, from the main argument in the paper) call for Floridi to take more strongly into account the socio-political ramifications of infosphere construction.

The next article uses biologically inspired cognitive architectures to show why we are unable to acquire new phenomenal qualia by learning from books. I first sketched out this argument at the session organized by this committee at the 2012 APA Central Division meeting. We have a rule to try to publish papers from committee-organized events. (Also, I have been following a self-imposed rule not to publish my work in this newsletter as long as I am the editor, which I have followed for eight years.) I am making a rare exception to the latter rule and in favor of the former (the review process for this paper has been organized by the guest editor of the following issue). In the following article, Ron Loui discusses how machine learning may be able to preserve the philosophical gist of the inductive method in domain-specific epistemic contexts. The author is against black-box approaches in machine learning and in favor of a more philosophical tradition that attempts to portray structure. He views neural networks as only post-hoc explainable and in contrast with the methods of rational theory formation. Loui follows Kyburg's account of scientific theory-formation in hopes of providing an approach available to machine-learning researchers.

The previous issue contained an important statement by Felmon Davis and D. E. Wittkower encouraging departments to consider online publications on their merit as opposed to means of distribution. We close this issue with a statement by Davis and Wittkower on the use of Massive Open Online Courses in philosophical education. The authors provide a helpful graph that illustrates Gartner Hype Cycle and argue that, after the initial hype, MOOCs are now at the low part of that curve. This encourages conversation that will help us move to the slope of further enlightenment and productivity. In several issues of this newsletter we discussed, and often encouraged, various kinds of online education in philosophy and we welcome further debate.[5]

Last but not least, I am on a year-long sabbatical (currently at the Australian National University). As a result, John Sullins has graciously agreed to guest-edit the next issue (spring 2015). I would appreciate if enquiries pertaining to that issue were addressed directly to him at john.sullins@sonoma.edu.

NOTES

1. Engineering achievements such as the wheel, Guttenberg's printing machine, or modern sewing machine may constitute civilizational transformations, but not scientific ones.
2. This point comes from a recent lecture by Kyle Stanford (University of California, Irvine) on "History, Unconceived Alternatives, and a Scientific Realism Dispute Worth Having" (Australian National University, May 8, 2014). The author focused on "our repeated failure to exhaust the space of theoretical alternatives well-

confirmed by the evidence available to us at a given time." This failure bears "significant consequences for how we should actually go about conducting the scientific enterprise." Stanford argues that the peer-review system in science, especially as applied by the funding agencies, favors funding "safe" projects that are unlikely to bring about the science that is transformative or paradigm-changing. I think this applies to peer-review in many journals.

3. Steven Thaler, "The Creativity Machine Paradigm: Withstanding the Argument from Consciousness," *APA Newsletter on Philosophy and Computers* 11, no. 2 (Spring 2012): 19–30.
4. James Moor, "Taking the Intentional Stance Towards Robot Ethics," *APA Newsletter on Philosophy and Computers* 6, no. 2 (Spring 2007): 14–17.
5. For instance in volume 11, number 1 (fall 2011) we had six articles on it, including Ron Barnette, "Reflecting Back Twenty Years," 20–22; Frank McCluskey, "Reflections from Teaching Philosophy Online," 22–24; Terry Weldin-Frisch, "A Comparison of Four Distance Education Models," 24–27; Kristen Zbikowski, "An Invitation for Reflection: Teaching Philosophy Online," 27–29; and Thomas Urban, "Distance Learning and Philosophy," 29–33.

FROM THE CHAIR

Models for Machine Ethics

Thomas M. Powers
UNIVERSITY OF DELAWARE

At this early stage in the development of a philosophical machine ethics, already a wide range of theoretical positions have been staked out. There are machine ethics utilitarians, and of course proponents of approaches following Kant, Ross, and various virtue theorists, too. Further categorization yields top-down, bottom-up, and hybrid approaches. A consideration of the logic to be employed in machine ethics finds proponents of deontic logic, default logic, decision trees, and optimization heuristics, to name just a few. All in all, there has been a healthy dose of philosophizing about machine ethics since 1942. (That's not an arbitrary date—Asimov's "Three Laws of Robotics" in the short story "Runaround" certainly inspired the philosophy of machine ethics, if only implicitly). And we've had a relative explosion of machine ethics in the last dozen or so years.

So it is in this context that I'd like to propose that machine ethics has reached a maturity at which it might be profitably extended, and perhaps re-directed, by means of models. I have a specific sense of "model" in mind. I'm not thinking of a representation or analogue of the way the world is, but rather a test scenario for various configurations of hardware, software, and other infrastructure, with humans co-mingled with machines. A machine ethics model would demonstrate the ways in which a specific configuration might behave. The philosophical theories will be deployed in machines in the model. With these models we can test our philosophical theories while we test hardware and software in a variety of environments. Such models will test the machines in much the same sense that a pharmaceutical company uses clinical trials to test the efficacy and safety of new drugs—in the beginning, with surrogates for humans to provide for some assurance of safety. The models could test the results of a variety of machine ethics approaches, and for this the contributing philosopher will need collaborators in the fields of computer science and engineering and robotics. As philosophers, we should be ready to admit our limitations; we simply don't know (yet) whether our favorite assemblage of ethics and logic will work well once encoded and embodied.

I suggest this operational extension to the exclusively philosophical (armchair) approach to machine ethics because there are two hard problems here. First, we do not and likely will not all agree on what behavior in a machine is best, in a particular environment in which it acts. We have a better chance, however, of agreeing which behaviors are better and which are worse when we see them in context. Second, there may be some appealing philosophical approaches that—for whatever reason—are very hard to implement. Take, for example, a putative "utilitarian" machine that lacks any reliable way to get information on human preferences, or even on the feelings or desires that would be consequent to its actions. Here, the project of a utilitarian machine would be dead in the water, it would seem, and further theorizing might be for naught. Both "hard" problems are at least addressed by an operational approach. With this approach, we have an experimental setting in which to judge our attempts at machine ethics.

Consider the following archetype of the kind of model I've been discussing.[1] Let's say that we want to test a robot that rolls along busy city sidewalks, empties trash cans, and in general keeps the sidewalk and gutter clean and tidy. The model system—simplified here to serve as a quick example—can have the following modules to gather and process information. 1) **mission**—its designed goals and ancillary objectives; 2) **physical parameters**—measures speed, end-effector pressure, etc.; 3) **law**—considers regulations that might apply to its behavior and the behavior of humans; 4) **ethics**—the collaborating philosopher's best effort at normative constraints; 5) **network**—access to relevant information about environment and other agents. Now we are in a position to see how such a robot might consider an ethically relevant problem that could arise in a particular environment. Should a robot pull a human pedestrian back onto the sidewalk, in the event that she has stepped off into the street?

A properly constructed model would show us how the robot answers the question, and how it would act in this situation. In order for it to vary from its mission (cleaning sidewalks), it would have to have an affordance for that behavior. It might also be designed to correct for problems that it caused, such as if it had moved into the pedestrian and forced her into the street. In this way, its mission could allow for modification to objectives to account for complicity in harms, or even to prevent harms opportunistically. Information about physical parameters could tell the robot whether it can move quickly enough to save the pedestrian from harm, and also how tightly its end-effector (claw) would have to grip to pull back the pedestrian. A crushed hand would not be the proper remedy for a low risk of being hit by a car. Concerning the law, the robot could process information related to jaywalking, and would have to be programmed to either impede, ignore, or maybe even report it. Indeed, the robot might be programmed to take a jaywalking pedestrian

into custody! The legal module intersects, of course, with its machine ethics. A relatively "libertarian" ethic would allow for jaywalking—or maybe even an apparent suicide attempt—while a more "paternalistic" approach would lead the robot to intercede in both cases. Finally, if the robot sends and receives networked information, it could have non-local information about impending threats on the nearby road or even on the sidewalk.

This brief description is not supposed to stand in for an actual test model. Indeed, one upshot of my example is that coding a machine ethic, even with a minimal system like the one sketched above, will be quite complicated. But it has been too tempting for our theorizing to get ahead of experimentation, and now it is time to get serious about plausible machine ethics alternatives.

A goal of machine ethics is to produce an autonomous machine that will behave in ways that humans will find generally acceptable from a moral point of view. As I postulated above, this presents a doubly hard problem, and I think we'll only make further progress when we see the outcomes of our theorizing in concrete terms. Philosophers need to partner with major research-capable entities, corporations, and universities to build and test models. In the future, if not already, the mixing of humans and autonomous machines will produce a complex "interaction," with humans behaving differently towards one another and towards machines than they would have without machines in the mix. It will be quite helpful if we can have some preview of that future, and models of machine ethics will afford us that much.

NOTES

1. I introduced the following example in a panel talk on "Ethics and Autonomous Agents" at the 2014 Computer Ethics and Philosophical Enquiry (CEPE) conference in Paris. I would like to thank the organizers—Jean-Gabriel Ganascia and Grégory Bonnet—as well as the other panelists and audience members for their comments.

ARTICLES

Ethical Trust in the Context of Robot Assisted Surgery

John P. Sullins
SONOMA STATE UNIVERSITY

Abstract. Robot surgery began as a general practice in the United States in 2000 when the robotic da Vinci Surgical System was approved for use in hospitals by the FDA. Recently there have been some questions regarding the safety record of these machines and their benefits relative to their expense. Should we as patients and caregivers place more or less trust in the technology of robotic surgery? To answer this question we will look at the possibility that robotic surgery may be an example of reverse adaptation, where the technology drives the social contract between the doctor and patient. Additionally, we will look at the impacts robotic surgery will have on traditional aspects of the medical ethics of surgery, such as informed consent, autonomy of surgeons and patients, corporate marketing, and the duty to provide the best available care, as well as increasing the asymmetry of the trust relationship between surgeon and patient. It will be argued that these issues will grow in importance as future robot surgery systems gain more autonomy in making or suggesting surgical strategies.

1. INTRODUCTION

In 2000 the U.S. FDA (the United States Food and Drug Administration) approved the use of the da Vinci Surgical System for use in performing robotically assisted surgery. Use of this system has continued to grow at an impressive rate. Since its initial beginnings as a technological curiosity in 2000 to its wide acceptance as a cutting-edge medical technology in 2013, it has been reported that 367,000 surgical procedures were facilitated by this system in the United States and 1.5 million worldwide.[1] For certain kinds of surgeries, this system is rapidly becoming the minimally invasive technology of choice with surgeons. The makers of the da Vinci Surgical System, Intuitive Surgical of Sunnyvale California, claim that as of 2013 its da Vinci system is used in 27 percent of all hysterectomies for benign conditions, and a staggering 87 percent of prostatectomies.[2] This has helped lower the number of more risky open surgery procedures where large incisions must be done and the body of the patient opened up to allow traditional surgical techniques to take place.

The da Vinci system enjoys a near monopoly in the world of surgical robotics, but there are more systems in development and competition is inevitable. Yet this system still sets the standard for robotic surgery as of the writing of this paper. It makers claim that

> The da Vinci Surgical System is a tool that utilizes advanced, robotic technologies to assist your surgeon with your operation. It does not act on its own and its movements are controlled by your surgeon. The da Vinci Surgical System has a 3D high definition (3D-HD) vision system, special instruments and computer software that allow your surgeon to operate with enhanced vision, precision, dexterity and control. The 3D-HD image can be magnified up to 10 times so your surgeon has a close-up view of the area he or she is operating on. The da Vinci instruments have mechanical wrists that bend and rotate to mimic the movements of the human wrist—allowing your surgeon to make small, precise movements inside your body. And, da Vinci software can minimize the effects of a surgeon's hand tremors on instrument movements.[3]

This sounds like a wonderful addition to the tools of surgery, but recently some safety concerns have been raised regarding this system. There have been questions of safety and cost. Some studies argue that there are situations where the system is not any more risky than an alternative type of surgery but it has the problem of being more expensive.[4] Others have reported that robotic surgery is not only more expensive but more dangerous to the patient.[5] There is also a report that has made many news headlines claiming

that hospitals regularly underreported injuries suffered by patients undergoing robotically assisted surgeries.[6] Shortly afterwards, there was an increase in the number of safety issues reported by those hospitals using robotic surgery, and this has prompted a new study by the FDA, the results of which are as yet unreleased.[7]

These questions have deep legal and financial implications. Therefore, it is best that we leave it up to large and careful institutions such as the FDA to fully analyze the claims made by both sides of the debate. We are not going to resolve the safety and cost debate here, but we can add to the discussion by looking at the ethical issues raised by the use of robotic surgery systems. It is always best when law or policy is reflected by good ethical thought. We will consider whether or not there has been an ethical breach of trust, which is somewhat different from a claim of legal fraud. Some have reported that the claims made by those selling this new and expensive form of surgery are motivated by commercial interests and oversold to the patients in need of hysterectomies[8] and robot assisted surgeries in general.[9] Even if it turns out to not be true that robotic surgery is more dangerous, we still need to know if this technology has changed the patient to doctor relationship and vice versa.

We are left with the important question of what is the appropriate level of trust that patients, surgeons, and hospital administrators should place in this new device. If we are to trust this technology, then we have to be sure that this is not just another instance of an overhyped product. But there is another concern that is also important: Is robotic surgery another example of the reverse adaptation to a new technology? Reverse adaptation occurs when we alter important social norms and situations to suit the limitations of a new technology, rather than forcing designers to make the technology more suitable to our lives before we begin mass use of the technology in question.

In our exploration of this topic we will need to review the special ethical concerns raised by critics of this type of surgery.[10] Of particular interest from the literature is the question of informed consent when it comes to letting patients know that there is currently a complex debate going on as to the safety and cost of robotic surgery. Since it is very hard for anyone to fully foresee the impacts of new information technologies, it is very common in the information technology world to release products that are not quite finished; once the users start complaining, the devices are quickly upgraded. This strategy does not seem ethical in the medical context, and we may be letting the wow factor of high-tech robotic surgeries influence our decisions.

We also need to know if the special kind of professionalism that has emerged in the surgery profession is in jeopardy of eroding due to possible deskilling by robotic surgery. Also, the rapid raise of robotic surgery seems to be side-lining the development of other, more traditional, surgical techniques. Since these technologies are costly for hospitals to buy, they will not be available in most of the developing countries in the world that need to rely on more traditional surgical methods. Up until now there could be a fruitful exchange of ideas between surgeons in all parts of the world. We need to know if robotic surgery will increase the divide in technical knowledge and know-how to the detriment of developing nations. Even if it turns out that robotic surgery is safe and cost effective, the ethical points raised above will still stand and a thorough ethical analysis of this technology is warranted.[11]

2. SURGERY ETHICS

Since robot assisted surgery is part of the vast territory of modern medical care, there will be much overlap of the ethical concerns questioned in the existing literature of medical ethics. In order to focus our discussion, I want to pay close attention to how the specific technologies of robot assisted surgery affect the ethical decisions made in the practice of surgery. To achieve that end, I will list the questions in surgery ethics that I think are most affected by the technology in question, explain them briefly, and then move on to how these concerns might be impacted by the growth of robot assisted surgery.

In this paper I am primarily interested in the question of ethical trust. There is a significant asymmetry in what the patient knows about the medical situation she finds herself in and what the surgeon believes would be the best treatment to mitigate the medical problem. The surgical specialists that the patent hires to provide the medical care knows much more than the patient, and the patient has to trust that they will work in her best interest. A great deal of energy is spent by medical professionals in maintaining their status as trustworthy members of the community that one can be confident in handing themselves over to for care. Surgery is a special case in medical care that demands even more trust from the patient since the patient undergoing a serious operation places her life almost literally in the hands of the surgeon.[12]

When it comes to surgical procedures, there are some specific questions of trust that the patient has the ethical, and in most cases also a legal, right to be confident of before proceeding with surgery recommended by a specialist.[13]

While surgery is a very ancient practice, surgery ethics is a surprisingly new discipline. Peter Angelos has said, "while twenty years ago, the concept of surgical ethics was thought by many to be an oxymoron, today there is increasing exploration of ethical issues central to surgery in surgical journals."[14]

In their article "Ethics in Surgery," Anji Wall et al. argue that although much of bio and medical ethics applies to surgery, it does so in ways that are different from other medical practices: "in clinical medical ethics, although principles such as autonomy and non-maleficence are distinctly prominent for both medical ethics and surgical ethics, they do not map on to one another exactly."[15] We will address some of those differences now.

The major difference, as was mentioned earlier, is that there is a great deal of physical intimacy involved in the surgeon-patient relationship that is quite unique given that the surgeon operates literally inside the body of

the patient, placing the patient in a radically vulnerable position.[16] The philosopher Rom Harré, in his book "Physical Being," maintains that the human body "is an object whose value is routinely recognized in the ways it is accorded the protection that pertains to those entities a culture takes to be sacred."[17] Not to suggest that modern surgeons make too much of this, but the former statement might suggest that there is at least the hint of something of the sacred in the relationship between the surgeon and her patient. Since people are not things, the surgeon has to perform a special kind of mental ritual in which the patient goes from fellow person to the anesthetized, unconscious, body-as-mechanism in the operating theater, back to a fully functioning person post operation.

One more philosophical point that will underlie my arguments rests on my commitment to a form of embodied morality. This might best be described by Lakoff and Johnson, who state very succinctly that "our concepts of what is moral, like all our other concepts, originate from the specific nature of human embodied experience."[18]

Under this view, morality emerges from our embodiment and the possible interactions with each other and our world that our body affords. It follows that surgical altering of the body is a fundamentally, ethically charged action that carries with it a very high standard of personal responsibility for the surgeon in ensuring a good outcome of the operation.[19] It is astounding that this fact was not seen in earlier ethical thought, but it is now and it must be taken seriously.

Anji Wall et al. describe a number of ethical issues that are specific to surgery such as surgical informed consent, industry relationships, and outcomes reporting. They also inform us that there is a distinct difference in the way trust must develop between surgeons and patients. The relationship, while much more intimate than that with a primary care physician, can also be much shorter, so there is a rushed timeframe in which the surgeon must gain the trust and informed consent of the patient.[20]

Industry relations are another tricky ethical point that will be particularly telling in the discussions later in this paper. While it can be successfully argued that there should be a separation of industry concerns from medical care in general practice, this is more difficult to maintain in surgery. Surgery is a technology, and therefore its innovations are technically mediated. Surgeons need industry to finance new surgical technologies, and since surgeons are the only ones that have the authority to use these new technologies on live humans, industry needs surgeons to propose new surgical innovations and test them when they are built. "An ethics is needed surrounding the collaborative effort between surgeons and industry, which recognizes the necessity of this relationship as well as the potential for significant COI [conflict of interest]."[21] Outcomes reporting is also easy to do with surgical procedures since the causes and effects of actions taken in surgery are far less mysterious than they are with drug reactions, for example. As we saw in the introduction, this ease of outcomes reporting has been a double-edged sword in the case of robotic surgery, as both sides, pro and con, argue over the meaning of the outcomes that have been reported in various forms of robotic surgery.

Nada Gligorov, working with surgery faculty and surgery clerkship interns at Mount Sinai School of Medicine in New York, completed a two-year project to identify ethical dilemmas particular to surgery.[22] They came up with a wide variety of specific cases, such as if a surgeon discovers a dangerous tumor while he is operating for some other condition, is it ethical to remove the tumor without the patient's consent? From the various cases, they concluded that there are three broad categories that the majority of the cases they found can be fit into: (1) the scope of informed consent, (2) truth telling with respect to the occurrence and disclosure of medical errors or the role of learners, and (3) decisional capacity.[23]

All of these make some background assumptions in ethical theory that Peter Angelos describes as "the four widely known ethical principles of respect for autonomy, beneficence, non-maleficence, and justice as the prism through which to consider surgical complications."[24]

Informed consent includes specific ethical concerns such as paternalism, respect for the autonomy of the patient, surrogacy in patient advocacy, and the beneficence of proposed surgical procedures. Since the patient is often not conscious during surgery, the patient must trust that the decisions made on her behalf during the procedure will be done with her best interests in mind.[25]

Truth telling mostly focused on the necessity to report medical errors such as "iatrogenic injury and errors in diagnosis, erroneous administration of drugs and other medications, technical errors during surgical procedures, and errors in interpretation of laboratory findings."[26] Some of the specific ethical questions that they found in informed consent, such as autonomy and beneficence, reappear here and non-malfeasance could be added as well. There is also an interesting new ethical mandate: fiduciary responsibility. Fiduciary responsibility is the ethical mandate that the surgeon place the interests of her patient before her own. Gligorov et al. argue that since the patient has placed so much trust and control of her body in the surgeon's hands, the surgeon has a moral obligation to honor that trust; thus, it would be a gross breach of that trust if a surgeon were to use that power to only further the interests of himself, or the institution he works for. "Thus, a physician may not conceal or refrain from disclosing a medical error in hopes of avoiding a lawsuit or unpleasant emotions and embarrassment."[27]

Decisional capacity is closely linked to the concept of informed consent but is a bit more specific in that it insists that the ability or "capacity" of the patient to make independent decisions about his own health care is of a sufficiently high level to warrant it an autonomous decision made in the patient's self-interest. An interesting wrinkle here is that Gligorov et al. remind us that patients might display inadequacy in their own ethical decision making, for instance, selfishly demanding that only the best surgeon be allowed to work on them, and that these decisions do not have to be honored.[28]

One additional aspect of surgery ethics that I will want to make further reference to is the professionalism and the demand that all surgeons display excellence in the skills necessary for their profession. Peter Angelos has argued that "perhaps nowhere in medicine more than surgery is the influence of role models more important."[29] The apprenticeship model is important to remember when thinking about surgery ethics because it means that concepts from virtue ethics and the ethics of individual character will be of particular value in the rest of our discussion. I also bring up the point of professionalism in surgery specifically because there has been some discussion that technologies such as robotics are diminishing the ability of new medical professionals to develop the skills necessary for their job; for instance, will robot surgery devices eventually replace humans or otherwise diminish the role that human surgeons now play?[30]

The above discussion is not meant to be an exhaustive look at surgery ethics as others have already accomplished that task.[31] There are many other ethical issues that relate to the professional life of a surgeon that we will not need to cover here since they are not directly impacted by robotic surgery technologies. But now that we have reminded ourselves of the most philosophically interesting questions in surgical ethics that have a bearing on robotic surgery, let us now look at the ways in which robotic technologies have, or may, change how we reason about surgical ethics.

3. INNOVATIONS IN SURGERY AND ROBOETHICS

Surgery is a medical practice that very closely links technologies, sciences, and social practices. In the literature of the philosophy of technology and science studies, this kind of complex system is often referred to as a *technoscience*. With the reader's indulgence, I will use this term, but by doing so I am making only the modest claim that technology, science, and society are interrelated in mutually supporting ways, not the stronger claim that science and technology are determined only by culture, which is a view one might read in the works of Bruno Latour, where the word "technoscience" was coined. I am hijacking the word here, as I do on many occasions, for my own purposes. In a technoscience, innovations can be initiated from scientific discovery, technological change, and/or societal needs.[32] The innovations found in robotic surgery find their origins in complex mixtures of these three catalysts for change.

Even though innovations can be influenced by these largely impersonal forces, there is a distinct role for personal ethical reasoning when surgeons attempt to place controls on adopting new innovations. Paulo Palma and Tomas Rosenbaum begin their article, "The Ethical Challenge of Surgical Innovation," with an important quote: "there is no control on surgical innovation outside of the realm of the surgeon's own ethical and moral compass"—but these ethical deliberations are never simple.[33]

The literature on the ethical impacts of surgery innovation is somewhat sparse but there has been some quality thought put into what has been written, and the following factors have been identified as important considerations for surgeons who are considering the use of innovative surgical techniques and technologies.

> Foremost, it is important to remember that:
>
> Ethical dilemmas always appear when new devices developed in the laboratory are transferred to the bedside. Regulatory agencies such as the Food and Drug Administration in the United States are more flexible with new devices than with drugs.[34]

This means that surgical innovations are approved for use relatively easily—sometimes with no clinical trial at all.[35] Therefore, surgeons cannot just rely on governmental agencies, or even academic or institutional review boards, to always make the right ethical decisions before a new innovation is introduced into surgical practice.[36] This tendency grows out of the special history of surgical innovation where many techniques simply occurred to the surgeon while an operation was underway and then were communicated to her peers for discussion later. This tolerance for individual innovation has greatly advanced the technology of surgical procedures. Interestingly enough, since robotic surgery is an innovation that comes from outside the surgical establishment, the introduction of robotics to surgery has had to deal with more regulations than other surgical innovations and has not been marked by this wide open experimental process.[37] We will look at the implications of this fact in a little more detail in the next section of the paper.

In addition to the ethical concerns raised by surgery in general that we looked at in the last section of the paper, there are a number of ethical qualifications that must be addressed when looking at the surgical innovations involved in surgery robots.

Innovation in surgery has always been motivated by the wish to decrease morbidity and mortality in the patient outcomes of surgical procedures. But in modern times this is not the only motivator; sometimes innovations in robotic surgery are for cosmetic reasons. For instance, a traditional thyroidectomy has a high chance of success and a high, twenty-year survival rate but leaves the patient with a large scar at the bottom of her neck. Robotic assisted transaxillary thyroidectomy is a high-tech innovation that allows the surgeon to perform the operation through an opening further down the chest, leaving a scar that is easier for the patient to live with; however, it is hard to tell if that procedure is ethically better than the traditional surgery. Although the twenty-year survival rate is believed to be unaffected, this is still unknown since twenty years have not yet passed from the time the first procedure was done. Consequently, there may be small increases in serious and life-threatening complications from mistakes made in this procedure.[38] The reason for this innovation is not simply that surgeons thought it was better than the traditional procedure, since in fact it might be marginally worse in terms of complications and increased chance of mortality. Instead, this innovation is more likely motivated by technological possibilities afforded by advancements in robotics, manufacturers looking for a way to market

their robotics technologies, and patients wanting a more minimally invasive procedure with a beneficial cosmetic outcome.

Informed consent is also exacerbated when it comes to innovations in robotic surgery since both the patient and the surgeon will be susceptible to the natural human inclination to equate the newness of a technology with the idea that it is also a significant improvement over the old procedures. This can lead to an optimism bias of the surgeon, who may have been involved in the design of the new system.[39]

There is often a steep learning curve involved in becoming proficient in new surgical techniques,[40] and this has been particularly true of surgery robotics.[41] This can lead to a surgeon's inability to fully inform the patient of potential risks since her inexperience with the technology would cause her not to know what they are, and the novelty of the machine might mean the makers of the system don't even know all of the potential risks.

New innovations are expensive and surgery robotics is no exception. For a hospital to buy a robot-assisted surgery device like the da Vinci Surgical System it would cost well beyond a million U.S. dollars, and that is just to get it in the door; it will also require costs related to continued maintenance and technology support. Also, innovations may increase the time a surgery takes in the operating room, though this might also be offset by faster patient recovery times for the more minimally invasive surgeries afforded by the long arms and small actuators, lights, and cameras that robots have which can fit into very small incisions in the patient's body.

As we saw earlier, conflict of interest is a problem here since the robotics industry cannot design these robots without input from surgeons, but this may make these surgeons less capable of properly critiquing the innovations they are financially or academically benefiting from.[42]

Pete Angelos has suggested that it will be difficult to regulate this from the outside, and this is painfully evident in the recent debate over the da Vinci Surgical System where the FDA is having a difficult time accurately determining if there is even a problem for them to investigate or not.[43] This means that the primary locus for ethical debate must be led by surgeons themselves with significant input from ethicists, professional societies, patients, and patient advocates.

Before we end our discussion in this section, we should look briefly at roboethics. Roboethics is a new discipline that seeks to discover the most ethical ways to add robotics and agent-based technologies into various aspects of our lives. Since robots are designed to do things that humans otherwise would have, this technology is particularly ethically charged. Of course, defining what a robot is, that other technologies are not, is a technically and philosophically challenging task,[44] but let's just be pragmatic here and simply define a surgical robot as a medical device designed to perform an action that traditional surgical tools needed a human (either a surgeon or an assistant) to perform in the past. So a scalpel is not a robot, but a machine that manipulates a scalpel blade either on its own, or under the transduced movements of a surgeon, is a robot.[45]

Even though the concept of robot is a shifting term that has evolved to include much more than the imaginary humanoid machines it was originally coined to refer to, there are some concrete ethical concerns that robots of any shape give rise to.

The first is what I and others have referred to as "distancing," which is the tendency for human operators to experience changes in their ethical commitments to other humans based on the technological mediation of the interaction provided by the robot. We will see that this may play a role in altering the surgeon-patient relationship as it becomes mediated through the surgical robot. The second is a kind of ethical confusion wherein the users of a system mistakenly purport more ethical agency to the system than it actually has. And the third is reverse adaptation, in which the social system that the machine is inserted into alters to fit the needs of the machine rather than the other way around. Finally, robotic systems are attempts to make technology more autonomous and proactive in solving our problems. Thus, they always have at least a little AI programming, and we are likely to see more and more added to these systems as advances in AI are forthcoming. This means that eventually a line may be crossed where the machine has a startlingly high level of autonomy and agency, in which case roboethics will need to help describe how to program these machines to make ethical choices. The moral of the story is that robots either influence human ethical decisions or, in more advanced cases, begin making those decisions themselves.

When it comes to robots used in surgery, even now we have machines all along the spectrum of autonomy. Camarillo et al., in their historical overview of surgery robots, show that as of now we have some machines such as CT scanners that are almost fully autonomous, whereas telesurgical devices such as the da Vinci Surgical System are not very autonomous at all. Conversely, the CT scanner has little direct contact with the patient, whereas the da Vinci is very active through the surgical process.[46] Thus, we see that when an action is safe and routine, it can be more readily automated, but if the surgical action is risky and requires a lot of cognitive skill to perform, the machine must be far less automated. Machines in the middle of this spread are those like AESOP, which the surgeon controls with voice commands while the machine autonomously provides imaging that the surgeon can use during the surgery.[47] Another example is RoboDoc, which the surgeon can program to mill bone somewhat autonomously during an operation while she attends to other things; this works since the reaction of bone is predictable and relatively free of complications.[48] In the future, we might see machines that are both autonomous and directly active in making surgical decisions, but we are some years away from that right now.

4. ETHICAL TRUST BETWEEN PATIENT AND ROBOT SURGEON

We now have enough background to approach the question that I started this paper with: What is the ethical status of the trust we as patients and caregivers should place in the technology of robotic surgery? That answer will be found in how much robotic technoscience has modified the traditional relationship between the surgeon and her patient.

The first issue is distancing. Aimee van Wynsberhe and Chris Gastmans have addressed this in their ethical appraisal of telesurgery in the context of an ethics of care.[49] In their analysis, it is most important that the system does not reduce the patient to simply an object being operated on. They also find that there have been some significant benefits in the use of telesurgery for both patients and surgeons, given that the systems allow patients to have the benefits of minimally invasive surgery, while the surgeons get a system that is more ergonomically comfortable for them to perform the procedure with and:

> The robotic interface is subsumed in the caring work of the physician and re-integrates the element of attentiveness of the surgeon. Thus, the phases of care-giving and care-taking are enhanced.[50]

At this time there are few telesurgeries that are done with the surgeon at a great distance from the surgery. In the da Vinci system, the surgeon is seated at a console just a few feet from the surgery, but there is no reason that this distance can't be increased to thousands of miles to be used on distant battlefields, the deep ocean, or even in space; in fact, NASA and others have been working on this capability for some time.[51] There are also many forms of telesurgery, each one a little different—a remote surgeon could consult with a local surgeon through information technology, or the remote surgeon might mentor the local surgeon through a novel process, to the remote surgeon assisting in the operation or even taking over and doing the entire procedure remotely. Each of these has a different ethical dimension; the further the surgeon is from the procedure she is doing, the less likely that the surgeon can develop a human connection to the patient. But as long as there is a human locally present, the local surgeon can take on that role and we can then tolerate the dehumanizing nature of the remote surgery.[52] But we also have to remember that if it is a choice between remote surgery and nothing at all, such as in a battlefield situation, then the comparative luxury of ethical human care will have to be lost in order to gain the good of saving a life.

A technical issue that can lead to ethical concerns is that the system will experience latency in its actions at greater levels the further one gets from the site of the surgery.[53] This would mean that it might be risky to trust such a system to be operating in one's best interest, but again this might have to be tolerated if it is the only option and the patient needs an appendix removed on her way to Mars.

We also must acknowledge that ethical distancing, in the case of surgery, might actually be beneficial. A surgeon actually does her job better when she can compartmentalize her view of the human she is working on as a fellow person during the surgery. In normal situations in most cultures, it is considered quite rude to carve on another person with a knife; therefore, we need to have a situation where the personhood of the patient is set aside just a bit while the operation is underway but returned quickly when it is over.[54] Consequently, a mixed system, in which humans are tasked with caring for the patient and treating her like a person after the surgery, will help make distant telesurgery with robotic systems tolerably ethical and maximize informed consent and ethical trust between the surgeons and their patients.

When it comes to surgery robots, we have a different situation than what is found in care robots commonly. Care robots are designed to make the patient feel that they care for them just by looking at the friendly visage and gestures of the machine. I am thinking of machines like Paro, the companion robot.[55] These machines can make the patient feel that there is a real caring agent there behind the robot's eyes when there most likely isn't. When it comes to surgical robots, they all tend to look like futuristic torture devices and may actually elicit just the opposite reaction from patients who might find them frightful or more dangerous than they actually are. The design of these machines is dictated by function and little effort is placed in making them look friendly. The only personal experience I have had with medical robots was a CT scan, and that was a little emotionally disturbing due to the confined space and loud noises the machine made. I only kept my sanity by listening to the human voice of the technician who talked me thorough the episode. We can see that the design of the machine affects the conditions under which informed consent is sought and on the loss of autonomy that these systems require from the patient.

Finally, we need to address the charge of reverse adaptation raised above. If it is true that robotic surgery devices are being pushed by industry in ways that are unwelcome by surgeons and their patients, then there are few good reasons to trust this technology. More subtly, this technology may slowly alter the technoscience of surgery, where incremental changes add up over time to a system that is not the most ethical system one would hope for.

As we saw in the introduction, some surgeons question the rapid growth of the use of the da Vinci surgical system, claiming that robotic surgery is more expensive, takes more time, and produces no significant benefits.[56] Yet others argue that injuries and complications from the system are under reported,[57] or that the systems are over hyped.[58] But the system also has some vocal defenders who claim that while it is too early to tell, it has promise for some procedures,[59] and others that claim the system is clearly beneficial.[60]

The extent of this debate leads me to conclude that for now, one should think very carefully about allowing one's surgeon to use a robotic surgery device. It does not look like there is a conspiracy to hide the truth about injuries and complications created by using robots in surgery, but it does look as though there is some confusion in exactly

what the correct injury rates are. These systems are still new, which means that if you consent to the use of robotic surgery, you consent to be part of the testing of these new procedures. There are some procedures that clearly benefit from the use of robotics but others that do not. The patient needs to make sure she is a good self-advocate and does not allow the optimism bias of her surgeon to overly influence her decision. It is just an unavoidable problem that informed consent is going to be a more difficult decision with this technology.

It is also clear that a lot of marketing is going into the sale of these machines, and there will be a bit of over prescription and reverse adaptation due to hospitals having these expensive machines and therefore wanting to use them. That in no way implies that there are not many legitimate cases in which they are the right choice, but it does mean that it is possible to be misled if we are not careful. There does not seem to be a conspiratorial ethical breach of trust but, nevertheless, when it comes to robot surgery—trust but verify.

5. ROBOT SURGERY AND PROFESSIONALISM

Our discussion of reverse adaptation leads naturally to the question of how these technologies will change the profession of surgery.

Mark Coeckelbergh has written about the effect information technologies has had on the health-care profession; some of his findings apply to our topic at hand. His starting point is an idea taken from virtue ethics, where good work is seen as something more than just technical efficiency; it's also quality driven and ethical.[61] The ethical situation created by good work is experienced both by the agent and the patient. The patient receives good, humane care, and the agent is made a better, more virtuous person through the process. The agent develops a practical wisdom from doing well so that through the experience of doing quality work she becomes more able to make correct ethical judgments in future situations.

> Ethics should not be understood as something external that is or should be imposed on the practice; this usually does not work and is rightly resisted by professionals of all sorts. The marriage of moral and professional excellence is an internal matter: developing moral and professional skills is internal to developing oneself as a (care) worker.[62]

That is what we would hope to achieve, but it is possible to design technoscientific systems that impede this process. For instance, we might create an ethically dubious system by deploying a robotic surgery system where a remote surgeon works on patients in a distant location, one after another, with no mitigations present for her to ever see the results of the beneficial actions her work. Or, if the robotic surgery systems became more autonomous and caused surgeons to become deskilled, then professionalism and excellence would be diminished along with the ethical value of the work done.

One might make the counter-point that who cares if the surgeon is denied a chance to become excellent at surgery if the machine that replaced her provided a better outcome? Coeckelbergh argues that arguments like this are missing the point. Better means ethical, so one would have to show that the technologies that replace the surgeon produce not only good technical results but also better social and ethical outcomes as well. If this can't be done, then the technical system is not really better.[63]

Coeckelbergh is not arguing that all technological advances are by necessity deskilling and corrosive of professionalism. He contends that it would be possible that while the machine might take some of the necessity for learning technical skills from the caregiver, it will then give that person the time she needs to further develop her interpersonal care-giving skills.[64] The job of future surgery professionals then might not be to be as active as they are now in the surgery, but as human guides that help the patient navigate the somewhat impersonal and technological process of robotic surgery. It is only another human that will experience the appropriate amount of worry and care that goes into contemplating something as drastic as surgery. The robot does not have the capacity to care if the patient lives or dies; only the human operators can bear this burden.

> It is not uncommon for a surgeon to stay up late thinking about a difficult upcoming operation. In fact, when most people think about the risks and burdens of surgery, they tend to focus solely on the patient. However, the toll of operating in complex cases where the risk of complications is great should not be underestimated.[65]

Even with great advances in the autonomy of robotic surgery, until machines become conscious there will always be a role for the human care professional in surgery. As was mentioned earlier, surgery is a profession where role models play a vital role.[66] This means that future surgeons will need to focus more and more on providing good role models for making moral choices and advocating for patients, since this may become their primary duty in a future of robotic surgery.

One last point is that while technologically advanced nations put a lot of effort and money into robotic surgery, we should note one very large undelivered promise. Telesurgery was initially proposed as a way to give better care to the far flung locations of this world where people do not have access to quality surgical care. Unfortunately, the lack of standardization in communications technologies between nations and the digital divide between the technological capabilities of the industrial world with those who still lag behind means those telesurgical technologies are not of any use to the developing world.[67] Surgical innovations that rely on high technology do not transfer well to the rest of the world, and our selfish drive to create this technology for our own use can be seen as unethical when we could be putting that energy into developing more traditional techniques that can be easily adopted in other locations.[68]

6. CONCLUSIONS AND FUTURE CONCERNS

After this discussion we are left with a cautious optimism for the technoscience of robotic surgery. We can be hopeful

because robotic surgery holds a great deal of promise for beneficial outcomes for patients. While operation times have increased with the use of robotics, the recovery times for patients have decreased. If cost of care can decrease and safety increase, and we find a way to cross the digital divide that separates the use of robotic surgery in the developing world, then robotic surgery will be an ethical addition to the technoscience of surgery.

The prognosis for surgery professionals is also a mixed bag. Some surgeons will receive many financial and academic rewards for their innovative work in bringing more autonomous machines to the operating theater, but that success may mean that the profession of surgery is not open to as many new practitioners as it was in the past. If we see them only as skilled technicians, then future human surgeons may lose their status as valued professionals.

> [T]he surgeon must be driven by altruistic motives rather than self-interest. In order for surgeons to maintain their position as professionals in society, they must not allow the lure of the new and the potential for financial benefit to influence their assessment of whether an innovative procedure truly benefits the patient . . . the future of surgical innovation is fraught with ethical concerns.[69]

Some future ethical concerns that surgeons will have to face in the slightly more distant future will be even more challenging. One is the surgical implantation of robotic devices into humans. Robotic prostheses are already on the design board but these are typically wearable items; soon enough there will be robotic items that will be permanently added to the body. Arguably this has already happened with the first artificial heart in the mid-sixties. Many of the ethical concerns we raised above will come into play here, but a new one will be: How ethical is it to enhance the human body? How will a surgeon decide whether or not to remove a perfectly good appendage and replace it with a robotic one just because the patient wants the increased capabilities of the artificial limb? And we haven't even touched on the problems with implanting cognitive upgrades in a human brain.

Another amusing, but real, problem might occur when a skilled autonomous robotic surgeon with a high level of cognitive skill petitions to join a prestigious medical professional society. Hutan Ashrafian et al. have proposed that the answer to that might be in subjecting the machine to a modified form of the Turing Test where it would be tested against humans for its skill in diagnosis with other human doctors. They state that

> the application of diagnostic accuracy meta-analytical capability in the context of the modified Turing test leads to two core issues: (a) what are the ethical implications of developing medical diagnostic systems to meet the Turing test and (b) does a patient have a right to know whether or not he/she is consulting a machine or a human practitioner?[70]

These and other issues will have to continue to be monitored.

ACKNOWLEDGEMENTS

A version of this paper was presented at AISB50 April 1-4, 2014—the Society for the Study of Artificial Intelligence and the Simulation of Behaviour (the AISB). I would like to thank the anonymous reviewers for their helpful comments that made this a stronger paper.

NOTES

1. Robert Lowes, "Complications of Robotic Surgery Underreported, Study Say"; The da Vinci® Surgery Experience Intuitive Robotics Fact Sheet.
2. The da Vinci® Surgery Experience Intuitive Robotics Fact Sheet.
3. Ibid.
4. Will Boggs, "Robotic Colectomy No Better, but More Expensive."
5. John Carreyrou, "Surgical Robot Examined in Injuries."
6. Lowes, "Complications of Robotic Surgery Underreported, Study Say"; Carreyrou, "Surgical Robot Examined in Injuries"; Michol A. Cooper, Andrew Ibrahim, Heather Lyu, and Martin A. Makary, "Underreporting of Robotic Surgery Complications."
7. Robert Lowes, "FDA Investigates Robotic Surgery System After Adverse Event Spike."
8. M. B. Schiavone, E. C. Kuo, R. W. Naumann, W. M. Burke, S. N. Lewin, A. I. Neugut, D. L. Hershman, T. J. Herzog, and J. D. Wright, "The Commercialization of Robotic Surgery: Unsubstantiated Marketing of Gynaecologic Surgery By Hospitals."
9. Linda X. Jin, Andrew M. Ibrahim, Naeem A. Newman, Danil V. Makarov, Peter. J. Pronovost, and Martin A. Makary, "Robotic Surgery Claims on United States Hospital Websites."
10. Peter Angelos, "The Ethical Challenges of Surgical Innovation for Patient Care"; Angelos, "Orlo Clark and the Rise of Surgical Ethics"; Nada Gligorov, Pippa Newell, Jason Altilio, Mike Collins, Amanda Favia, Leah Rosenberg, and Rosamond Rhodes, "Dilemmas in Surgery: Medical Ethics Education in Surgery Rotation"; Laurence B. McCullough, James W. Jones. "Unravelling Ethical Challenges in Surgery."
11. GinaKolata, "Results Unproven, Robotic Surgery Wins Converts"; Carol Peckham and Joseph Colella, "Is Robotic Surgery Worth Its Price? An Interview With Dr. Joseph Colella."
12. Angelos, "Orlo Clark and the Rise of Surgical Ethics."
13. In everything that follows it will be assumed that the patient is a fully rational adult. I acknowledge that ethical reasoning can get much more complicated when the patient is not an adult or is an adult but suffers from some form of mental deficit.
14. Angelos, "Orlo Clark and the Rise of Surgical Ethics."
15. Anji Wall, Peter Angelos, Douglas Brown, Ira J. Kodner, and Jason D. Keune, "Ethics in Surgery."
16. Angelos, "Orlo Clark and the Rise of Surgical Ethics."
17. H. Rom Harré, *Physical Being: A Theory for Corporeal Psychology*.
18. George Lakoff and Mark Johnson, *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*.
19. Wall et al., "Ethics in Surgery."
20. Ibid.
21. Ibid.
22. Gligorov et al., "Dilemmas in Surgery: Medical Ethics Education in Surgery Rotation."
23. Ibid.
24. Peter Angelos, "Complications, Errors, and Surgical Ethics."
25. Wall et al., "Ethics in Surgery."
26. Gligorov et al., "Dilemmas in Surgery: Medical Ethics Education in Surgery Rotation."

27. Ibid.
28. Ibid.
29. Angelos, "Orlo Clark and the Rise of Surgical Ethics."
30. Mark Coeckelbergh, "E-care As Craftsmanship: Virtuous Work, Skilled Engagement, and Information Technology In Health Care."
31. Wall et al., "Ethics in Surgery."
32. Note that by society I mean not only the public outside of professional surgeons but the social groups, both professional and informal, that surgeons themselves form.
33. Quotation from Reiter-Theil S, GJ Agich (2006) Research on Clinical Ethics and Consultation. Introduction to the theme. Med Health Care Plilos 1:3–5. See also P. Palma and T. Rosenbaum, "The Ethical Challenge of Surgical Innovation."
34. Palma and Rosenbaum, "The Ethical Challenge of Surgical Innovation."
35. Ibid.
36. Ibid.; Peter Angelos, "Ethics and Surgical Innovation: Challenges to the Professionalism of Surgeons."
37. David B. Camarillo, Thomas M. Krummel, and J. Kenneth Salisbury, "Robotic Technology in Surgery: Past, Present, and Future."
38. Angelos, "The Ethical Challenges of Surgical Innovation for Patient Care"; Angelos, "Ethics and Surgical Innovation: Challenges to the Professionalism of Surgeons."
39. Angelos, "Ethics and Surgical Innovation: Challenges to the Professionalism of Surgeons."
40. Angelos, "The Ethical Challenges of Surgical Innovation for Patient Care"; Palma and Rosenbaum, "The Ethical Challenge of Surgical Innovation"; Angelos, "Ethics and Surgical Innovation: Challenges to the Professionalism of Surgeons."
41. Camarillo et al., "Robotic Technology in Surgery: Past, Present, and Future."
42. Wall et al., "Ethics in Surgery."
43. Lowes, "FDA Investigates Robotic Surgery System After Adverse Event Spike."
44. Camarillo et al., "Robotic Technology in Surgery: Past, Present, and Future."
45. I realize that that commits me to some possibly strange extrapolations, such as having to say that an aircraft is flown entirely by wire or remote control is a robot, but I am fine with that charge if you want to level it on me.
46. Camarillo et al., "Robotic Technology in Surgery: Past, Present, and Future."
47. Ibid.
48. Ibid.
49. A. van Wynsberghe and C. Gastmans, "Telesurgery: An Ethical Appraisal."
50. Ibid.
51. T. Haidegger, J. Sándor, and Z. Benyó, "Surgery in Space: The Future of Robotic Telesurgery."
52. van Wynsberghe and Gastmans, "Telesurgery: An Ethical Appraisal."
53. Haidegger et al., "Surgery in Space: The Future of Robotic Telesurgery."
54. Harré, *Physical Being: A Theory for Corporeal Psychology*.
55. http://www.parorobots.com/
56. Boggs, "Robotic Colectomy No Better, but More Expensive"; C. M. Kang, D. H. Kim, W. J. Lee., and H. S. Chi, "Conventional Laparoscopic and Robot-Assisted Spleen-Preserving Pancreatectomy: Does da Vinci Have Clinical Advantages?"
57. Lowes, "Complications of Robotic Surgery Underreported, Study Say"; Carreyrou, "Surgical Robot Examined in Injuries"; Cooper et al., "Underreporting of Robotic Surgery Complications."
58. Jin et al., "Robotic Surgery Claims on United States Hospital Websites."
59. H. Wykypiel, J. Bodner, G. Wetscher, and T. Schmid, "Robot-Assisted Versus Conventional Laparoscopic Fundoplication: Short-Term Outcome of a Pilot Randomized Controlled Study."
60. Peckham and Colella, "Is Robotic Surgery Worth Its Price? An Interview With Dr. Joseph Colella"; S. Y. Park, G. S. Choi, J. S. Park, H. J. Kim, and J. P. Ryuk, "Short-Term Clinical Outcome of Robot-Assisted Intersphincteric Resection for Low Rectal Cancer: A Retrospective Comparison with Conventional Laparoscopy."
61. Coeckelbergh, "E-care As Craftsmanship: Virtuous Work, Skilled Engagement, and Information Technology In Health Care."
62. Ibid.
63. Ibid.
64. Ibid.
65. Peter Angelos, "Complications, Errors, and Surgical Ethics."
66. Angelos, "Orlo Clark and the Rise of Surgical Ethics."
67. van Wynsberghe and Gastmans, "Telesurgery: An Ethical Appraisal."
68. Angelos, "The Ethical Challenges of Surgical Innovation for Patient Care"; Angelos, "Ethics and Surgical Innovation: Challenges to the Professionalism of Surgeons."
69. Angelos, "Ethics and Surgical Innovation: Challenges to the Professionalism of Surgeons."
70. Hutan Ashrafian, Ara Darzi, and Thanos Athanasiou, "A Novel Modification of the Turing Test for Artificial Intelligence and Robotics in Healthcare."

BIBLIOGRAPHY

Angelos, Peter. "Complications, Errors, and Surgical Ethics." *World Journal of Surgery* 33, no. 4 (2009): 609–11.

———. "Ethics and Surgical Innovation: Challenges to the Professionalism of Surgeons." *International Journal of Surgery* 11, Suppl. 1 (2013): S2–S5.

———. "Orlo Clark and the Rise of Surgical Ethics." *World Journal of Surgery* 33, no. 3 (2009): 372–74.

———. "The Ethical Challenges of Surgical Innovation for Patient Care." *Lancet* 376, no 9746 (2010): 1046-47.

Ashrafian, Hutan, Ara Darzi, and Thanos Athanasiou. "A Novel Modification of the Turing Test for Artificial Intelligence and Robotics in Healthcare." *The International Journal of Medical Robotics and Computer Assisted Surgery*. Early view, January 20, 2014. doi:10.1002/rcs.1570.

Boggs, Will. "Robotic Colectomy No Better, but More Expensive," Reuters Health, reported on *Medscape News*. December 26, 2013. Accessed December 30, 2013. http://www.medscape.com/viewarticle/818302.

Camarillo, David B., Thomas M. Krummel, and J. Kenneth Salisbury. "Robotic Technology in Surgery: Past, Present, and Future." *The American Journal of Surgery* 188, no. 4, Suppl. 1 (2004): 2–15.

Carreyrou, John. "Surgical Robot Examined in Injuries." *Wall Street Journal*, May 4, 2010.

Coeckelbergh, Mark. "E-care As Craftsmanship: Virtuous Work, Skilled Engagement, and Information Technology In Health Care." *Medicine, Health Care, and Philosophy* 16, no. 4 (2013): 807–16.

Cooper, Michol A., Andrew Ibrahim, Heather Lyu, and Martin A. Makary. "Underreporting of Robotic Surgery Complications." *Journal for Healthcare Quality*. Early view, August 27, 2013. doi:10.1111/jhq.12036.

Gligorov, Nada, Pippa Newell, Jason Altilio, Mike Collins, Amanda Favia, Leah Rosenberg, and Rosamond Rhodes. "Dilemmas in Surgery: Medical Ethics Education in Surgery Rotation." *Mount Sinai Journal of Medicine* 76 (2009): 297–302.

Haidegger, T., J. Sándor, and Z. Benyó. "Surgery in Space: The Future of Robotic Telesurgery." *Surgical Endoscopy* 25, no. 3 (2011): 681–90.

Harré, H. Rom. *Physical Being: A Theory for Corporeal Psychology*. Blackwell Publishers, 1991.

Jin, Linda X., Andrew M. Ibrahim, Naeem A. Newman, Danil V. Makarov, Peter. J. Pronovost, and Martin A. Makary. "Robotic Surgery Claims on United States Hospital Websites." *Journal for Healthcare Quality* 33, no. 6 (2011): 48–52.

Kang, C. M., D. H. Kim, W. J. Lee., and H. S. Chi. "Conventional Laparoscopic and Robot-Assisted Spleen-Preserving Pancreatectomy: Does da Vinci Have Clinical Advantages?" *Surgical Endoscopy* 25 (2011): 2004–09.

Kolata, Gina. "Results Unproven, Robotic Surgery Wins Converts." *The New York Times*, February 13, 2010.

Lowes, Robert. "Complications of Robotic Surgery Underreported, Study Say." *Medscape Medical News*, September 5, 2013. Accessed December 30, 2012. http://www.medscape.com/viewarticle/810490.

———. "FDA Investigates Robotic Surgery System After Adverse Event Spike." *Medscape Medical News*. April 30, 2013. Accessed December 30, 2013. http://www.medscape.com/viewarticle/803339.

Lakoff, George, and Mark Johnson. *Philosophy in the Flesh: The Embodied Mind and Its Challenge to Western Thought*. Basic Books, 1999.

McCullough, Laurence B., James W. Jones. "Unravelling Ethical Challenges in Surgery." *Lancet* 374, no. 9695 (2009): 1058-59.

Palma, P., and T. Rosenbaum. "The Ethical Challenge of Surgical Innovation." *International Urogynecology Journal* 20, no. 4 2009): 375-76.

Park, S. Y., G. S. Choi, J. S. Park, H. J. Kim, and J. P. Ryuk. "Short-Term Clinical Outcome of Robot-Assisted Intersphincteric Resection for Low Rectal Cancer: A Retrospective Comparison with Conventional Laparoscopy." *Surgical Endoscopy* 27, no. 1 (2013): 48–55.

Peckham, Carol, and Joseph Colella. "Is Robotic Surgery Worth Its Price? An Interview With Dr. Joseph Colella." *Medscape General Surgery*. June 20, 2013. http://www.medscape.com/viewarticle/806484.

Schiavone, M. B., E. C. Kuo, R. W. Naumann, W. M. Burke, S. N. Lewin, A. I. Neugut, D. L. Hershman, T. J. Herzog, and J. D. Wright. "The Commercialization of Robotic Surgery: Unsubstantiated Marketing of Gynaecologic Surgery By Hospitals." *American Journal of Obstetrics and Gynecology* 207, no. 3 (2012): 174.e1–e17.

The da Vinci® Surgery Experience. Intuitive Robotics Fact Sheet. Accessed December 30, 2013. http://www.davincisurgery.com/assets/docs/da-vinci-surgery-fact-sheet-en-1005195.pdf.

van Wynsberghe, A., and C. Gastmans. "Telesurgery: An Ethical Appraisal." *Journal of Medical Ethics* 34, no. 10 (2008): e22.

Wall, Anji, Peter Angelos, Douglas Brown, Ira J. Kodner, and Jason D. Keune. "Ethics in Surgery." *Current Problems in Surgery* 50, no. 3 (2013): 99–134.

Wykypiel, H., J. Bodner, G. Wetscher, and T. Schmid. "Robot-Assisted Versus Conventional Laparoscopic Fundoplication: Short-Term Outcome of a Pilot Randomized Controlled Study." *Surgical Endoscopy* 22, no. 5 (2008): 1407.

Yang, Y., F. Wang, P. Zhang, C. Shi, Y. Zou, H. Qin, and Y. Ma. "Robot-Assisted Versus Conventional Laparoscopic Surgery for Colorectal Disease, Focusing on Rectal Cancer: A Meta-analysis." *Annals of Surgical Oncology* 19 (2012): 3727–36.

Information Warfare: The Ontological and Regulatory Gap

Mariarosaria Taddeo
UNIVERSITY OF WARWICK, UNITED KINGDOM,
M.TADDEO@WARWICK.AC.UK

INTRODUCTION

In his 1985 paper "What is Computer Ethics?" Moor discussed the changes that the information revolution was prompting as well as the relevance and the need for conceptual analysis addressing such changes. In his words, "although a problem . . . may seem clear initially, a little reflection reveals a *conceptual muddle*. What is needed in such cases is an analysis which provides a coherent *conceptual framework within which to formulate a policy for action*."[1]

Almost three decades later, with contemporary societies turning into information societies, the policy vacuum and the conceptual muddle underpinning it have become not just evident but pressing issues to be solved. Understanding and regulating privacy and anonymity, as well as security and well-being, in the information age have become crucial to the existence and functioning of our societies and the well-being of their citizens. Information warfare (IW) is one of the most compelling cases to be addressed.

Historically, technological breakthroughs determine changes affecting the structure of both civil society and military organizations. As described by Toffler and Toffler (1997), this was the case with the Neolithic revolution, when human beings first made weapons out of wood and rocks, and with the industrial revolution, which provided the means for industrialized warfare and for the dissemination of weapons of mass destruction. The information revolution is the latest example. It has changed our activities in several ways and at several levels.[2] Information and communication technologies (ICTs) have reshaped social interactions; they provide new tools for the management of information and bureaucracy; and when considered with respect to warfare, ICTs determine the latest revolution in military affairs, making IW the warfare of the information age.

Information warfare raises a number of ethical and regulatory problems, all of which rest on a key feature, namely its transversality.[3] Information warfare may arise and target physical as well as non-physical objects, it may go from non-violent to highly violent, and it also prompts an increased blurring of the distinction between military and civilian, as it no longer reflects the distinction between combatant and non-combatant. The transversality of IW, coupled with the growing dependency of contemporary societies on ICTs, unveils the potential for IW to become a new form of *total war*. For the scope of the mobilization, targets, and resources increasingly overlaps with resources, agents, and infrastructures of contemporary societies.

Regulating IW to ensure its fairness and avoid escalation risk is therefore pivotal. Since the first cyber-attack to Estonian websites in 2008, the debate surrounding the regulation of IW has grown quickly and has accompanied concrete efforts to understand whether and how existing international laws and treaties could be endorsed to regulate it. Such efforts have proven to be quite demanding and were not the exclusive concern of the military; they have also had a bearing on ethicists and policy-makers, since existing ethical theories of war and national and international regulations struggle to address the novelties of this phenomenon.

In the rest of this article I will analyze how some of the most relevant tenets of Just War Theory (JTW), and the international laws and treaties implementing them, are applied to the case of IW. In doing so I will mainly focus on the interpretations of existing laws and regulations given

in the so-called Tallinn Manual.[4] This has been the first and, so far, the most exhaustive work devoted to offer guidance in its application to the case of IW. The manual offers a valuable contribution to the debate over the regulation of IW, for it shows that extant laws and treaties can be stretched to address this phenomenon and that when it comes to the international ground, the cyber-sphere is not a new Wild West. I will argue, however, that it would be a mistake to consider the stretching of existing laws and treaties as the ultimate and perfectly satisfying strategy to regulate IW, for existing laws and treaties struggle to fully address the changes prompted by this phenomenon and the ethical problems that it poses.

While the application of existing laws and treaties to IW is indeed possible, it is at the same time a *stretch*, which will eventually reach its limits and generate a regulatory vacuum. To overcome the latter, a theoretical effort is needed to fully understand the nature of this new phenomenon, its ethical, political, and social implications, and so to design new norms and principles that will allow for regulating IW not by stretching an old blanket but by properly and adequately addressing the novelty of this new phenomenon. I shall begin this analysis by offering a definition of IW to clear the ground of any possible misunderstandings.

DISSOLVING THE MIST OF INFORMATION WARFARE

The expression "information warfare" has already been used in the extant literature to refer solely to the uses of ICTs devoted to breaching the opponent's informational infrastructure in order to either disrupt it or acquire relevant data and information about the opponent's resources, military strategies and so on.[5]

Distributed denials of service (DDoS) attacks, like the ones launched in Burma during the 2010 elections, the injection of Stuxnet in the Iranian nuclear facilities of Bushehr, and "Red October" (discovered in 2013) are all famous examples of how ICTs can be used to conduct cyber-attacks.[6] Nonetheless, such attacks are only one instance of IW. In what follows I will use a definition of IW that I provided in an earlier work and refer to IW to indicate a wide spectrum of phenomena, encompassing cyber-attacks as well as the deployment of robotic weapons and ICT-based communication protocols.[7]

IW is thus defined as follows:

> Information Warfare is the use of ICTs within an offensive or defensive military strategy endorsed by a [political authority] and aimed at the immediate disruption or control of the enemy's resources, and which is waged within the informational environment, with agents and targets ranging across the physical and non-physical domains and whose level of violence may vary upon circumstances.[8]

The informational nature and transversality of IW can be properly appreciated once they are considered within the framework of the so-called information revolution.[9] The information revolution has a wide impact on many of our daily practices—from our social and professional lives to our interactions with the environment that surrounds us. With the information revolution we have witnessed a shift, that has brought the *non-physical domain* to the fore and made it as important and valuable as the physical one.[10]

Information warfare is one of the most compelling instances of such a shift. It shows that there is a new environment, where physical and non-physical entities coexist and are equally valuable, and in which states have to prove their authority and new modes of warfare are being specifically developed for this purpose.[11] The shift towards the non-physical domain provides the ground for the transversality of IW. This is a complex aspect, and it can be better understood when IW is compared with traditional forms of warfare.

Traditionally, war entails the use of a state's *violence* through the state *military forces* to determine the conditions of governance over a determined territory.[12] It is a necessarily violent phenomenon that implies the sacrifice of human lives and damage to both military and civilian infrastructures. Here, the state faces the problem of how to minimize damage and losses while ensuring the enemy is overpowered.

Information warfare is different from traditional warfare in several respects, mainly because it is not a necessarily violent and destructive phenomenon.[13] For example, IW may involve a computer virus capable of disrupting or denying access to the enemy's database, and in so doing it may cause severe damage to the opponent without exerting *physical* force or violence. In the same way, IW does not necessarily involve human beings. In this context, an autonomous artificial agent can conduct an action of war, such as, for example, in the cases of EADS Barracuda, and the Northrop Grumman X-47B, or in the case of autonomous cruising computer viruses targeting other artificial agents or informational infrastructures, like a database or a website.[14] Information warfare can be waged exclusively in a digital context without ever involving physical targets; nevertheless, it may escalate to more violent forms.[15]

As remarked above, the transversality of IW is the key feature of this phenomenon; it is the aspect that most differentiates it from traditional warfare. Transversality is also the feature that engenders the ethical problems posed by IW. The potential bloodless and non-destructive nature of IW makes it desirable from both an ethical and a political perspective, since at first glance it seems to avoid bloodshed and it liberates political authority from the burden of justifying military actions to the public.[16] However, the disruptive outcomes of IW can inflict serious damage to contemporary information societies and at the same time may potentially lead to highly violent and destructive consequences—dangerous for both military forces and civil society. Consider, for example, the data diffused for GridExII.[17] This is a simulation that was conducted in the United States in November 2013. More than two hundred utility companies collaborated with the U.S. government to simulate a massive cyber-attack on the United States's basic infrastructure. Had the attack been real, estimates

mention hundreds of injuries and tens of deaths, while millions of U.S. citizens would have been left in darkness.

The need for strict regulations for declaring and waging fair IW is now compelling. To this end, an analysis that discloses the ethical issues related to IW while pointing in the direction of their solution is a preliminary and necessary step. This will be the task of the next section.

REGULATING INFORMATION WARFARE: *JUS AD BELLUM*

I will now focus on the application of *jus ad bellum* to cases of IW. Part of the problem relating to *jus ad bellum* concerns the so-called attribution problem and the difficulties of tracing back the author of a cyber-attack. As this seems to be more a technically related problem than a conceptual one, I shall not focus on it here. Also, I shall not focus on the problems related to *jus ad bellum*, as they have been extensively analyzed elsewhere.[18] Rather, I will devote my attention to the definition of what counts as a use of force in IW and what, as such, can trigger the waging of a war or a conflict.

I shall first consider some of the most common definitions of cyber-attacks, for they underpin the application of existing tenets of *jus ad bellum* to the case of IW. In this respect it is quite useful to compare two definitions, the one provided by the National Research Council in its 2009 report on cyber-attack capabilities (*Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities* 2014), and the one offered in the Tallinn Manual. In the former, a cyber-attack is defined as "the use if deliberate actions—perhaps over an extended period of time—to alter, disrupt deceive, degrade or destroy adversary computer systems or networks or the information and/or programs resident in or transiting these systems or networks."[19]

The Tallinn Manual defines cyber-attacks as "a cyber-operation, whether offensive or defensive, that is reasonably expected to cause injury or death to persons or damage or destruction to object."[20] The National Research Council's definition offers a more specific characterization of cyber-attacks, including non-physical damages as well as physical ones, while the scope of the definition offered by the Tallinn Manual remains undecided, for it depends on the definition of "objects." If these are understood as *physical* objects, then the manual is by default considering as attacks only kinetic uses of cyber-technologies. This seems actually to be the case if one considers the focus of the definition on physical damages and the absence of any reference to damages to intangible objects (e.g., data, information, informational infrastructure).

The consequences of such an approach are extremely relevant for they affect the application of *jus ad bellum* as well as of *jus in bello*. For example, rule ten of the Tallinn Manual stresses that under *jus ad bellum* a cyber-attack is unlawful if it constitutes a threat or *use of force* against a state. Rule eleven refines rule ten by stressing that a cyber-attack amounts to a use of force if its scale and effects are similar to those of non-cyber-operations. In this sense, the Tallinn Manual follows Hoisington, according to whom cyber-attacks that are intended to cause physical damage should be categorized as uses of force.[21]

Criteria based on the magnitude and effects of a cyber-attack have been proposed to assess if the former amounts to a use of force or to an armed attack, like the one described in rule eleven of the Tallinn Manual. This compares the effects of a cyber-attack with the scale of effects of a conventional attack in order. To support such an assessment, several tests can be run, the most famous one being the Pictet's test, which focuses on the scope, duration, and intensity of the attack.[22] Heyes and Kesan report that there are three models that can be used to apply the test to the case of cyber-attacks:

> [I]nstrument-based models look at whether the damage caused was of the kind that previously would have required a kinetic attack, such as shutting down a power grid. Effects-based models focus on the overall effect on the victim state, such as an information attack on financial institutions that causes significant damage to the economic well-being of the victim state. Finally, a strict liability model would consider any cyber-attack directed at critical infrastructure to be an armed attack.[23]

All of this is quite uncontroversial, for a cyber-attack that has the same or similar effects to a conventional attack should be treated as a kinetic attack in the eye of the law. In this case it is true what Schmitt states: "a thick web of international law norms suffuses cyber-space. These norms both outlaw many malevolent cyber-operations and allow states to mount robust responses."[24]

However, while very interesting and important, this approach inevitably finds its own limit as it overlooks the conceptual roots (i.e., JWT) on which laws regulating IW rest. In doing so, it misses the possibility of truly expanding the scope of existing laws by reshaping their conceptual framework. The consequence is that the approach fails to consider and to account for the conceptual changes prompted by IW and risks confusing an *ad hoc* remedy with the long-term solution, and, in the long run, risks imposing conceptual limitations on the laws and regulation for IW.

A good example in this respect concerns the application of the principle of just cause to IW. As Barrett noted,

> Since damage to property may constitute a just cause, can temporary losses of computer functionality also qualify as a *casus belli*? Like kinetic weapons, cyber-weapons can physically destroy or damage computers. But offensive computer operations, because of their potential to be transitory or reversible, can also merely compromise functionality. While permanent loss of functionality create the same effect as physical destruction, temporary functionality losses are unique to cyber-operations and require additional analysis.[25]

The issue is not whether the case of IW can be considered in such a way as to fit the parameters of kinetic warfare and hence to fall within the domain of JWT, as we know it. This result is easily achieved if the focus is restricted to physical damage and tangible objects. Rather, the problem lies at a deeper level and questions the very conceptual framework on which JWT rests and its ability to *satisfactory* and *fairly* accommodate the changes brought to the fore by the information revolution. The time has come to consider in more detail such changes; this will be the task of the next section.

THE ONTOLOGICAL GAP

Just War Theory mainly focuses on the use of force in international contexts and surmises sanguinary and violent warfare occurring in the physical domain. As the cyber-domain is virtual and IW mainly involves abstract entities, the application of JWT becomes less direct and intuitive. The struggle encountered when applying JWT to the cases of IW becomes even more evident if one considers how pivotal concepts such as harm, target, and attack have been reshaped by the dissemination of IW. The very notion of harm, for example, which is at the basis of JWT, struggles to apply to the case of IW. This is a problem that has been already highlighted in the extant literature; see, for example, Dipert, who argues that any moral analysis of this kind of warfare needs to be able to account for a notion of harm "[focusing] away from strictly injury to human beings and physical objects toward a notion of the (mal-) functioning of information systems, and the other systems (economic, communication, industrial production) that depend on them."[26]

The transversality of the ontological status of the entities involved in IW is particularly relevant as we try to shed some light on IW's novelty. Traditional warfare concerns human beings and physical objects, while IW involves artificial and non-physical entities alongside human beings and physical objects. Therefore, there is a *hiatus* between the ontology of the entities involved in traditional warfare and those involved in IW. Such a hiatus affects the ethical analysis, for JWT rests on an anthropocentric ontology—i.e., moral discourse is solely concerned with respect for human rights and disregards all non-human entities—and for this reason it does not provide sufficient means for addressing the case of IW (more details on this aspect presently).

The gap between the ontology assumed by JWT and the ontology of IW has also been described by Dipert, who stresses that

> [s]ince cyber-warfare is by its very nature information warfare, an ontology of cyber-warfare would necessarily include [a] way of specifying *information objects . . ., the disruption and the corruption of data and the nature and the properties of malware*. This would be in addition to what would be required of a domain-neutral upper-level ontology, which addresses this type of characteristics of the most basic categories of entity that are used virtually in sciences and domain: material entity, event, quality of an object, physical object. A cyber-warfare ontology would also go beyond . . . a military ontology, such as agents, intentional actions, unintended effects, organizations, artefacts, commands, attacks and so on."[27]

The case of the autonomous cruising computer virus will help to clarify the problems at stake.[28] These viruses are able to navigate through the web and identify autonomously their targets and attack them without requiring any supervision. The targets are chosen on the basis of parameters that the designers encode in the virus, so there is a boundary to the autonomy of these agents. Still, once the target has been identified, the virus attacks without having to receive "authorization" from the designer or any human agent.

In considering the moral scenario in which the virus is launched, three main questions arise. The first question revolves around the identification of the moral agents, for it is unclear whether the virus itself should be considered the moral agent, or whether this role should be attributed to the designer or to the agency that deployed the virus, or even to the person who actually launched it. The second question focuses on moral patients. The issue arises as to whether the attacked computer system itself should be considered the moral receiver of the action, or whether the computer system and its users should be considered the moral patients. Finally, the third question concerns the rights that should be defended in the case of a cyber-attack. In this case, the problem is whether any rights should be attributed to the informational infrastructures or to the system compounded by the informational infrastructure and the users.

As noted by Dipert, IW includes informational infrastructures, computer systems, and databases.[29] In doing so, it brings new objects, some of which are intangible, into the moral discourse. The first step towards an ethical analysis of IW is to determine the moral status of such (informational) objects and their rights. Help in this respect is provided by information ethics, which will be introduced in section five.

INFORMATION ETHICS

Information ethics is a macro-ethics which is concerned with the whole realm of reality and provides an analysis of ethical issues by endorsing an informational perspective. Such an approach rests on the consideration that "ICTs, by radically changing the informational context in which moral issues arise, not only add interesting new dimensions to old problems, but lead us to rethink, methodologically, the very grounds on which our ethical positions are based."[30]

In just one sentence information ethics is defined as a *patient-oriented*, *ontocentric*, and *ecological* macro-ethics. It is patient-oriented because it considers the morality of an action with respect to its effects on the receiver of that action. It is ontocentric for it endorses a non-anthropocentric approach for the ethical analysis. It attributes a moral value to all existing entities (both physical and non-physical) by applying the principle of ontological equality: "This ontological equality principle means that any form of reality . . ., simply for the fact of being what it is, enjoys a minimal, initial, *overrideable*, equal right to exist and develop in a way which is appropriate to its nature."[31] The principle of

ontological equality is grounded on an information-based ontology, according to which all existing things can be considered from an informational standpoint and are understood as informational entities, all sharing the same informational nature.[32]

The principle of ontological equality shifts the standpoint for the assessment of the moral value of entities, including technological artefacts. At first glance, an artefact, a computer, a book, or the Colosseum seems to enjoy only an instrumental value. This is because one endorses an anthropocentric Levels of Abstraction (LoA);[33] in other words, one considers these objects as a user, a reader, a tourist. In each case, the moral value of the observed entity depends on the agent interacting with it and on her purpose in doing so.

The claim put forward by information ethics is that these LoAs are not adequate to support an effective analysis of the moral scenario in which the artefacts may be involved. The anthropocentric, or even the biocentric, LoA prevents us from properly considering the nature and the role of such artefacts in the reality in which we live. The argument is that all existing things have an informational nature, which is shared across the entire spectrum—from abstract to physical and tangible entities, from rocks and books to robots and human beings. Further, all entities enjoy some minimal initial moral value *qua informational* entities.

Information ethics argues that universal moral analyses can be developed by focusing on the common nature of all existing things and by defining good and evil with respect to such a nature. The focus of ethical analysis is thereby shifted, since the initial moral value of an entity does not depend on the observer, but is defined in absolute terms and depends on the (informational) nature of the entities. Following the principle of ontological equality, minimal and overrideable rights to exist and flourish pertain to all existing things and not just to human or living things. The Colosseum, Jane Austin's writings, a human being, and computer software all share *initial* rights to exist and flourish, as they are all informational entities.[34]

A clarification is now necessary. Information ethics endorses a minimalist approach. It considers informational nature as the minimal common denominator among all existing things. However, this minimalist approach should not be mistaken for reductionism, as information ethics does not claim that the informational approach is the unique LoA from which moral discourse is addressed. Rather, it maintains that the informational LoA provides a *minimal starting point*, which can then be enriched by considering other moral perspectives.

Lest the reader be misled, it is worth emphasizing that the principle of ontological equality does not imply that all entities have the same moral value. The rights attributed to the entities are *initial*; they can be overridden whenever they conflict with the rights of other (more morally valuable) entities. Furthermore, the moral value of an entity is determined by its potential contribution to the enrichment and the flourishing of the informational environment. Such an environment, the *Infosphere*, includes all existing things, be they digital or analogue, physical or non-physical, and the relations occurring among them and also between them and the environment.[35] The blooming of the Infosphere is the ultimate good, while its corruption, or destruction, is the ultimate evil.

In particular, any form of corruption, depletion, or destruction of informational entities or of the Infosphere is referred to as *entropy*. In this case, entropy refers to "any kind of *destruction* or *corruption* of informational objects (mind, not of information), that is, any form of impoverishment of *being*, including *nothingness*, to phrase it more metaphysically" and has nothing to do with the concept developed in physics or in information theory.[36]

Information ethics considers the duty of any moral agent with respect to its contribution to the informational environment, and considers any action that affects the environment by corrupting or damaging it, or by damaging the informational objects existing in it, as an occurrence of entropy, and therefore as an instance of evil.[37] On the basis of this approach, information ethics provides four principles to identify right and wrong and the moral duties of an agent:

0) entropy ought not to be caused in the infosphere (null law);

1) entropy ought to be prevented in the infosphere;

2) entropy ought to be removed from the infosphere;

3) the flourishing of informational entities as well as of the whole infosphere ought to be promoted by preserving, cultivating, and enriching their properties.

These four principles together with the theoretical framework of information ethics will provide the ground to proceed further in our analysis, and define the principles for a just IW.

JUST INFORMATION WARFARE

The first step towards a definition of the principles for a just IW is to understand the moral scenario determined by this phenomenon. The framework provided by information ethics proves to be useful in this regard, for it allows for answering questions concerning the moral stance of the receivers of the actions performed in IW scenarios. The principle of ontological equality is quite useful in this respect, for it states that all (informational) entities enjoy some minimal initial rights to exist and flourish in the Infosphere, and therefore every entity deserves some minimal respect, in the sense of a "disinterested, appreciative and careful attention."[38]

When applied to IW, this principle enables us to consider all entities that may be affected by an action of war as moral patients. A human being, who gains some benefits from the consequences of a cyber-attack, and an informational infrastructure, which is disrupted by a cyber-attack, are both to be held as moral patients, as they are both the receivers of the moral action. Following information ethics, the moral

value of such an action is to be assessed on the basis of its effects on the patients' rights to exist and flourish, and ultimately on the flourishing of the Infosphere.

The issue then arises concerning which and whose rights should be preserved in the case of IW. The answer to this question follows from the rationale of information ethics, according to which an entity may lose its right to exist and flourish when it comes into conflict (causes entropy) with the rights of other entities or with the well-being of the Infosphere. It is a moral duty of the other inhabitants of the Infosphere to *remove* such a malicious entity from the environment, or at least to impede it from perpetrating more evil.

This framework lays the groundwork for the first principle for just IW since it prescribes the condition under which the decision to resort to IW is morally justified.

I) IW ought to be waged only against those entities that endanger or disrupt the well-being of the Infosphere.

Two more principles regulate just IW.

II) IW ought to be waged to preserve the well-being of the Infosphere.

III) IW ought not to be waged to promote the well-being of the Infosphere.

The second principle limits the task of IW to restoring the *status quo* in the Infosphere before the malicious entity begins increasing entropy within it. Information warfare is just so long as its goal is to *repair* the Infosphere from the damage caused by the malicious entity.

The second principle can be described using an analogy, namely, IW should fulfill the same role as police forces in a democratic state. It should act only when a crime has been, or is about to be, perpetrated. Police forces do not act in order to ameliorate the aesthetics of cities or the fairness of a state's laws; they only focus on reducing or preventing crimes from being committed. Likewise, IW ought to be endorsed as an *active* measure in response to the increasing of evil, and not as a proactive strategy to foster the flourishing of the Infosphere. Indeed, this is explicitly forbidden by the third principle, which prescribes the promotion of the well-being of the Infosphere as an activity that falls beyond the scope of a just IW.

These three principles rest on the identification of the moral good with the flourishing of the Infosphere and the moral evil with the increasing of entropy in it. They endorse an informational ontology, which allows for including in the moral discourse both non-living and non-physical entities. The principles also prescribe respect for the (minimal and overrideable) rights of such entities along with those of human beings and other living things, and respect for the rights of the Infosphere as the most fundamental requirement for declaring and waging a just IW.

In doing so, the three principles overcome the ontological hiatus described in section three and provide the framework for applying JWT to the case of IW. As such they point towards the direction for defining a new regulation for IW, which would be able to take into account the nature of the agents, targets, and environment involved in this phenomenon.

CONCLUSION

The goals of this article have been to fill the conceptual vacuum surrounding IW and to provide the ethical principles for a just IW, which can help in filling the vacuum. I have argued that JWT provides the necessary but not sufficient tools for this purpose. For, although its ideal of just warfare grounded on respect for basic human rights in the theatre of war holds also in the case of IW, it does not take into account the moral stance of non-human and non-physical entities which are involved and mainly affected by IW. This is the ontological hiatus, which I identified as the nexus of the ethical problems encountered by IW.

I also stressed that in order to be applicable to the case for IW, JWT must extend the scope of the moral scenario to include non-physical and non-human agents and patients. Information ethics has been introduced as a suitable ethical framework capable of considering human and artificial, physical and non-physical entities in the moral discourse. It has been argued that the ethical analysis of IW is possible when JWT is merged with information ethics. In other words, JWT *per se* is too large a sieve to filter the issues posed by IW. Yet, when combined with information ethics, JWT acquires the necessary granularity for addressing the issues posed by this form of warfare.

It would be misleading to consider the problems described in this article as reasons for dismissing JWT when analyzing IW, or for discarding altogether existing laws and regulations of warfare. Instead these problems point to the need to consider more carefully the case of IW, and to take into account its peculiarities, so that an adequate conceptual framework will be developed to properly take into account "contemporary values" while developing laws to regulate IW.

NOTES

1. James H. Moor, "What Is Computer Ethics?" Emphasis added.
2. Luciano Floridi, *The Fourth Revolution, How the Infosphere Is Reshaping Human Reality*.
3. Mariarosaria Taddeo, "Information Warfare: A Philosophical Perspective."
4. NATO Cooperative Cyber Defence Centre of Excellence. *Tallinn Manual on the International Law Applicable to Cyber Warfare: Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence*.
5. See, for example, Martin Libicki, *What Is Information Warfare?*; Edward Waltz, *Information Warfare: Principles and Operations*; Winn Schwartau, *Information Warfare: Chaos on the Electronic Superhighway*.
6. http://www.bbc.co.uk/news/technology-11693214; http://news.bbc.co.uk/2/hi/europe/6665145.stm; http://www.cbsnews.com/stories/2010/11/29/world/main7100197.shtml. For an annotated time line of cyber attacks, see NATO's website: http://www.nato.int/docu/review/2013/Cyber/timeline/EN/index.htm.

7. The reader may refer to Mariarosaria Taddeo, "Information Warfare: A Philosophical Perspective," and "Just Information Warfare," for a more detailed analysis of the reasons supporting such a definition.
8. Mariarosaria Taddeo, "Just Information Warfare," 3.
9. Floridi, *The Fourth Revolution, How the Infosphere Is Reshaping Human Reality.*
10. Mariarosaria Taddeo, "Information Warfare: A Philosophical Perspective."
11. The United States only spent $400 million in developing technologies for cyber conflicts: http://www.wired.com/dangerroom/2010/05/cyberwar-cassandras-get-400-million-in-conflict-cash/. The United Kingdom devoted £650 million to the same purpose: http://www.theinquirer.net/inquirer/news/1896098/british-military-spend-gbp650-million-cyber-warfare.
12. Michael Gelven, *War and Existence: a Philosophical Inquiry*.
13. John Arquilla, "Can Information Warfare Ever Be Just?"; Randall Dipert, "The Ethics of Cyberwarfare"; Edward T. Barrett, "Warfare in a New Domain: The Ethics of Military Cyber-Operations."
14. Note that MQ-1 Predators and EADS Barracuda, and the Northrop Grumman X-47B are Unmanned Combat Aerial Vehicles used for combat actions and they are different from Unmanned Air Vehicles, like, for example, Northrop Grumman MQ-8 Fire Scout, which are used for patrolling and recognition purposes only. Abiola Abimbola, José Munoz, and William J. Buchanan, "Analysis and Detection of Cruising Computer Viruses."
15. John Arquilla, "Can Information Warfare Ever Be Just?"; Richard A. Clarke, *Cyber War: The Next Threat to National Security and What to Do About It*; Joel Brenner, *America the Vulnerable: New Technology and the Next Threat to National Security*; Mark Bowden, *Worm: The First Digital World War*.
16. Dorothy E. Denning, "The Ethics of Cyber Conflict"; John Arquilla, "Twenty Years of Cyberwar."
17. http://www.nytimes.com/2013/11/15/us/coast-to-coast-simulating-onslaught-against-power-grid.html.
18. Mariarosario Taddeo, "Just Information Warfare"; Luciano Floridi and Taddeo, *The Ethics of Information Warfare*.
19. National Research Council, *Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities*, p. 80
20. NATO Cooperative Cyber Defence Centre of Excellence. *Tallinn Manual on the International Law Applicable to Cyber Warfare: Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence*, 106.
21. Matthew Hoisington, *Cyberwarfare and the Use of Force Giving Rise to the Right of Self-Defense*.
22. Jean Pictet was a Swiss jurist and the general editor for the Geneva Conventions of August 12, 1949.
23. Carol M. Hayes and Jay P. Kesan, *Law of Cyber Warfare*, 10.
24. Michael Schmitt, "Cyberspace and International Law: The Penumbral Mist of Uncertainty," 177.
25. Edward T. Barrett, "Warfare in a New Domain: The Ethics of Military Cyber-Operations," 6.
26. The need to define concepts such as those of harm, target, and violence is stressed both by scholars who argue in favor of the ontological difference of the cyber warfare (Randall Dipert, "The Essential Features of an Ontology for Cyberwarfare") and exploit this point to claim that JWT is not an adequate framework to address IW, and by those who actually maintain that JWT provides sufficient element to address the case of IW (George R. Lucas, Jr., "Just War and Cyber Conflict 'Can There Be an "Ethical" Cyber War?'"). See also Randall Dipert, "The Ethics of Cyberwarfare," 386.
27. Randall Dipert, "The Essential Features of an Ontology for Cyberwarfare," 36. Emphasis added.
28. Abiola Abimbola, José Munoz, and William J. Buchanan, "Analysis and Detection of Cruising Computer Viruses."
29. Randall Dipert, "The Essential Features of an Ontology for Cyberwarfare"
30. Luciano Floridi, "Information Ethics, Its Nature and Scope," 23.
31. Luciano Floridi, *Ethics of Information*.
32. The reader may recall the informational LoA mentioned in section 2. Information Ethics endorses an informational LoA, as such it focuses on the informational nature as a common ground of all existing things.
33. A LoA is a finite but non-empty set of observables accompanied by a statement of what feature of the system is under consideration. A collection of LoAs constitutes an interface. An interface is used when analysing a system from various points of view, that is, at varying LoAs. For example, a glass of wine observed at a chemical LoA consists of the observables of the chemical processes on going in liquid, while the same glass of wine being observed at the LoA of drinker might be identified by the observables that represent its taste and bouquet. A single LoA does not reduce a glass of wine to merely its on-going chemical processes or to its taste and bouquet. Rather, it is a tool that makes explicit the observation perspective and restricts it to only those elements that are relevant in a given observation. LoAs are hierarchically organized; a high LoA enables a general perspective and allows for a general analysis of the observed system. A low LoA provides a less general perspective and allows for a more detailed analysis. Luciano Floridi, "The Method of Levels of Abstraction."
34. For more details on the information-based ontology, see Luciano Floridi, "On the Intrinsic Value of Information Objects and the Infosphere." The reader interested in the debate on the informational ontology and the principles of Information Ethics may wish to see Luciano Floridi, "Understanding Information Ethics."
35. Luciano Floridi, *Ethics of Information*.
36. Ibid.; Luciano Floridi, "Understanding Information Ethics."
37. Luciano Floridi and J. W. Sanders, "Artificial Evil and the Foundation of Computer Ethics."
38. Ronald W. Hepburn, *"Wonder" and Other Essays: Eight Studies in Aesthetics and Neighbouring Fields*; Luciano Floridi, *Ethics of Information*.

BIBLIOGRAPHY

Abimbola, Abiola, José Munoz, and William J. Buchanan. "Analysis and Detection of Cruising Computer Viruses." 3rd International Conference Electronic Warfare and Security (EIWC), 2004.

Arquilla, John. "Can Information Warfare Ever Be Just?" *Ethics and Information Technology* 1, no. 3 (1998): 203–12.

———. "Twenty Years of Cyberwar." *Journal of Military Ethics* 12, no. 1 (2013): 80–87. doi:10.1080/15027570.2013.782632.

Barrett, Edward T. 2013. "Warfare in a New Domain: The Ethics of Military Cyber-Operations." *Journal of Military Ethics* 12, no. 1 (2013): 4–17. doi:10.1080/15027570.2013.782633.

Bowden, Mark. *Worm: The First Digital World War*. New York: Atlantic Monthly Press, 2011.

Brenner, Joel. *America the Vulnerable: New Technology and the Next Threat to National Security*. New York: Penguin Press, 2011.

Clarke, Richard A. *Cyber War: The Next Threat to National Security and What to Do About It*. 1st Ecco pbk. ed. New York: Ecco, 2012.

Denning, Dorothy E. "The Ethics of Cyber Conflict." In *Information and Computer Ethics*. Hoboken, NJ: Wiley, 2007.

Dipert, Randall. 2010. "The Ethics of Cyberwarfare." *Journal of Military Ethics* 9, no.4 (2010): 384–410.

———. "The Essential Features of an Ontology for Cyberwarfare." In *Conflict and Cooperation in Cyberspace*, edited by Panayotis Yannakogeorgos and Adam Lowther, 35–48. Taylor & Francis, 2013. http://www.crcnetbase.com/doi/abs/10.1201/b15253-7.

Floridi, Luciano. *Ethics of Information*. Oxford, New York: Oxford University Press, 2013.

———. "Information Ethics, Its Nature and Scope." *SIGCAS Comput. Soc.* 36, no. 3 (2006): 21–36. doi:10.1145/1195716.1195719.

———. "On the Intrinsic Value of Information Objects and the Infosphere." *Ethics and Information Technology* 4, no. 4 (2002): 287–304.

———. *The Fourth Revolution, How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press, 2014.

———. "The Method of Levels of Abstraction." *Minds and Machines* 18, no. 3 (2008): 303–29. doi:10.1007/s11023-008-9113-7.

———. "Understanding Information Ethics." *APA Newsletter on Philosophy and Computers* 7, no. 1 (2007): 3–12.

Floridi, Luciano, and J. W. Sanders. "Artificial Evil and the Foundation of Computer Ethics." *Ethics and Information Technology* 3, no. 1 (2001): 55–66.

Floridi, Luciano, and Mariarosaria Taddeo. *The Ethics of Information Warfare*. New York: Springer, 2014.

Gelven, Michael. *War and Existence: a Philosophical Inquiry*. University Park, PA: Pennsylvania State University Press, 1994.

Hayes, Carol M., and Jay P. Kesan. *Law of Cyber Warfare*. SSRN Scholarly Paper ID 2396078. Rochester, NY: Social Science Research Network, 2014. http://papers.ssrn.com/abstract=2396078.

Hepburn, Ronald W. *"Wonder" and Other Essays: Eight Studies in Aesthetics and Neighbouring Fields*. Edinburgh: University Press, 1984.

Hoisington, Matthew. *Cyberwarfare and the Use of Force Giving Rise to the Right of Self-Defense*. SSRN Scholarly Paper ID 1542223. Rochester, NY: Social Science Research Network, 2009. http://papers.ssrn.com/abstract=1542223.

Libicki, Martin. *What Is Information Warfare?* Washington, D.C.: National Defense University Press, 1996.

Lucas, Jr., George R. "Just War and Cyber Conflict 'Can There Be an "Ethical" Cyber War?'" 2012. Presented at the Naval Academy Class of 2014.

Moor, James H. "What Is Computer Ethics?" *Metaphilosophy* 16, no. 4 (1985): 266–75. doi:10.1111/j.1467-9973.1985.tb00173.x.

NATO Cooperative Cyber Defence Centre of Excellence. *Tallinn Manual on the International Law Applicable to Cyber Warfare: Prepared by the International Group of Experts at the Invitation of the NATO Cooperative Cyber Defence Centre of Excellence*. Cambridge; New York: Cambridge University Press, 2013.

Schmitt, Michael. "Cyberspace and International Law: The Penumbral Mist of Uncertainty." *Harvard Law Review* 126 (2013): 176–80.

Schwartau, Winn. *Information Warfare: Chaos on the Electronic Superhighway*, 1st ed. New York: Emeryville, CA: Thunder's Mouth Press; Distributed by Publishers Group West, 1994.

Taddeo, Mariarosaria. "Information Warfare: A Philosophical Perspective." *Philosophy and Technology* 25, no. 1 (2012): 105–20.

———. "Just Information Warfare." *Topoi* (April 2014): 1–12. doi:10.1007/s11245-014-9245-8.

Owens, William, Kenneth Dam, and Herbert Lin, eds. *Technology, Policy, Law, and Ethics Regarding U.S. Acquisition and Use of Cyberattack Capabilities*. Washington, D.C.: National Academies Press, 2009. Accessed July 31, 2014. http://www.nap.edu/catalog.php?record_id=12651.

Toffler, Alvin, and Anna Toffler. "Foreword: The New Intangibles." In *In Athena's Camp Preparing for Conflict in the Information Age*, edited by John Arquilla and David F. Ronfeldt. Santa Monica, CA: Rand Corporation, 1997.

Waltz, Edward. *Information Warfare: Principles and Operations*. Boston: Artech House, 1998.

Meaningful Reality: A Metalogue with Floridi's Information Ethics

Pompeu Casanovas

INSTITUTE OF LAW AND TECHNOLOGY, AUTONOMOUS UNIVERSITY OF BARCELONA, POMPEU.CASANOVAS@UAB.CAT /
CENTRE FOR APPLIED SOCIAL RESEARCH, ROYAL MELBOURNE INSTITUTE OF TECHNOLOGY, POMPEU.CASANOVAS@RMIT.EDU.AUS

Abstract. This is a comment on some aspects of the *The Ethics of the Information* by Luciano Floridi. This paper explores some of the notions advanced in the book, its methodology, and its practical and ontological turn. In the end, some suggestions are made about the relationship between Information Ethics (IE), policy, and law.

1. INTRODUCTION

The Ethics of the Information is a freely written book by a free thinker.[1] This statement intends to be more than a rhetorical one. Freedom means liberty. Liberty to think aloof. Liberty to dive into the philosophical tradition with a fresh gaze. And liberty to quote freely those past and present thinkers that Luciano Floridi believes to be quotable according to the subject he is dealing with, and the argument he is fleshing out. Academic writing puts forward some tacit rules—do avoid conveying personal feelings, do not cite incompatible schools, remain stuck to a single way of thinking; especially, never use the pronoun *I*. He breaches them all. To breach rules and keep reading and referring to Quine, Church, and Moore as often as to Deleuze, Cassirer, and Lacan without falling into syncretism is the privilege of an independent mind. It is a rare quality.

Let's put it differently. If the author of this book were asked "Do you believe in God?," most likely he would reply as Einstein did: "I believe in Spinoza's God." Do not laugh, do not weep, do not wax indignant. Understand. He invites his fellow readers to *understand* what is happening through a literary, pervasive, and sometimes irritating *I* that creates a complete series of neologisms to express his interrelationship and interface within the informational world—*infosphere, inforgs, conceptual design, hyperhistorical predicament, ITentities, re-ontologization, ontological friction, onlife experience, nested telepresence, foreward and backwards presence, metaphysical entropy, artificial evil, ecumenical axiology, homo poieticus, ecology of the self, ecopoiesis, informational privacy, informational structural realism, environmental ethics, distributed morality*, etc.

Let's try to situate their standpoint. If *The Ethics of the Information* were a mere invitation to dialogue, these terms would constitute metaphors the author would live by. A kind of holistic semantic network. But Floridi is carefully operating at each step through what he calls *levels of abstraction* (LoA), a methodological approach that allows a cartographic perspective to grade the conceptual map over the epistemic territory: he defines at each step the meaning of the terms, but he *shows* first the correlate references and co-references they intend to denote. This is what *structural realism* consists of. A sort of low and upgraded

phenomenology of the living web, to be discussed with and within the self-reflecting entities of the living web.

To use his own words when he faces identity, he *individuates* before *identifying* as informational entities the objects he is referring to. And he displays them through an articulated language, comprising the chosen level of abstraction (LoA), the informational and cognitive structure of those objects, and the type of complex representation that links them. So, not only answers but research questions are equally introduced to be discussed alongside with the reader.

2. THE ETHICS OF THE INFORMATION AS A METALOGUE

There is a semantic and pragmatic threshold then that the reader is gently required to cross over. Better than a dialogue, this game can be more adequately defined as a *metalogue*, an interactive and multi-leveled philosophical conversation about what thinking, writing, reading, and talking mean through the new technological conditions of contemporary life. It is worth noticing that this is a discourse (i) focused on the phenomenological *existence* of clusters, entities of information, (ii) and on the structure of its conditions: "To be present is to be the value of a typed variable of LoA."[2] In other words, to be present is to be digitized at a certain level.

This assumption made, the book enters into a self-referential but interactive analytical space in which the author's *I* dissolves to expand the relational reality—*hybrid*, a mixture of physical and artificial—he is proposing to explore and disclose gradually *in between*. For "to be is to be *interactable*."[3] This is not assuming an inner and outer dimension of the infosphere, but the infosphere as a primordial *Umwelt*, as the place in which we all already are supposed to live, sense, feel, think and communicate as producers and consumers (*prosumers*) of content. *Umwelt* points at the inescapable ecological environment in which informational units are *present*, including our encapsulated identity. "The self, and mental life in general," he states, "is located in the brain but not present in the brain. Thus the locus of the self is the brain but the self is not present in the brain."[4]

This largely resonates with the cybernetic perspective that Norbert Wiener and Gregory Bateson set up at the beginning of our information age, after World War II.

> Sometimes we [Bateson/Wiener] used to discuss whether a computer can think. The answer is "no." What thinks is a complete circuit that might include a computer, a man and the environment. Similarly, we can ask whether the brain can think, and again the answer is "no." What thinks is a brain that is inside of a man, who is part of a system comprising a room. Drawing a boundary line between a part (which does most of the computation to a larger system) and the larger system of which this is part means creating a mythological component commonly called I or self.[5]

Almost paradoxically, this mythical *I* is the only place where we can personalize the delocalized interface from which we are gathering, assembling, storing, and processing information. There is something interesting in the main idea that knowledge is always socially and collectively embedded or, coming back to Floridi's formulation, in the idea that we all constitute units of self-organized information capable to "semanticize" our ecological niche, the *infosphere*.

The rejection of anthropocentrism is not new. Bateson's "ecology of mind" (1972), Herbert Simon's and Allen Newell's "artificial intelligence" (1969), Minsky's "society of mind" (1976), and Arne Naess's idea of "deep ecology" (2005) were all grounded on some structured representation of a shared and common knowledge too, rooted onto a complex environment and able to be computed in an independent way. Their modeling was "constructionist" as well: they were building up conceptual models to grasp the nature of ecology or of computational science—"sciences of design"—as complex systems. However, to do so, they didn't need to equate their conceptual universe with a sharp idea of Being. Ontology, in the classical philosophical sense I will expose in the last section, was never a real issue for them. They had instead a strong sense of the sacred, of the boundaries of human knowledge.

Floridi, on the contrary, situates ontology at the center of his formulation: "Being and the Infosphere are co-referential. [. . .] The Infosphere is the *totality of Being* [emphasis added], hence the environment constituted by the totality of informational entities, including all agents, along with their processes, properties, and mutual relations."[6] In a nutshell:

> IE [Information Ethics] is an ecological ethics that replaces *biocentrism* with *ontocentrism*, and then interprets Being in informational terms. It suggests that there is something more elemental than life, namely Being, the existence and flourishing of all entities and their global environment, and something more fundamental than suffering, namely, *nothingness*. It then interprets Being and nothingness at an informational level of abstraction, as *infosphere* and *entropy*, on the basis on an informational structured realism as articulated in Floridi (2011, chs. 14 and 15). In short, it is an environmental ethics based on the phenomena and corresponding concepts of information/infosphere/entropy rather than life/ecosystem/pain.[7]

3. THE PRACTICAL TURN

Why is the author choosing this formulation? Why is he remaining in the philosophical language of Being (*Sein*), Presence (*Dasein*), Care (*Sorge*), and, most surprisingly, Nothingness (*Nichts*)? After the harsh Neo-positivist logical attack on the Heideggerian *Das Nichts nichtet*, it would be difficult to imagine for an analytical philosopher of the twentieth century to consistently keep such clear references to ostensive phenomenology in carrying out his general project on information ethics.

A possible answer is that Floridi is not a philosopher of the twentieth century, but a thinker past and beyond the linguistic turn, less worried with his own language than realistically committed to the description of what he perceives as a radical new way of living brought about by the sudden explosion of the digital world. If I could develop further the metalogical game I started above, I would rather imagine him in the Baroque Age, in good company with the giants of rationalism, enjoying and taking into account the content of the Bible, the Gospel, Greek and Latin philosophy, medieval and renaissance arts and crafts, and turning the results of science and mathematics into philosophical concepts that would stand by their own, *more geometrico*.

For instance, how to avoid *evil*—what Floridi calls now *artificial evil*, a hybrid between natural and moral evil this time—was one of the main obsessions not only of Spinoza, but of Hobbes and Leibniz as well.[8] All three thinkers are easily retrievable from Floridi's writing. Especially Spinoza, whose concepts of *substance* (i.e., information) and *conatus* (i.e., maintenance of Being), are directly and indirectly quoted in the book.

Let's reproduce the four principles of Information Ethics: (i) entropy ought not to be caused in the infosphere, (ii) entropy ought to be prevented in the infosphere, (iii) entropy ought to be removed from the infosphere, (iv) the flourishing of informational entities as well of the whole infosphere ought to be promoted by preserving, cultivating, and enriching their well-being.

Classical works by Cassirer, Hazard, and Skinner, to quote three different historical schools and languages, come easily to mind. The inner connection between the growing protection of rights (life, goods, property) and the evolution from security to happiness in the seventeenth and eighteenth centuries is a well-established fact in intellectual history.[9] In this sense, IE seems to be firmly rooted on the British Enlightenment as much as on the Neo-Kantian synthesis. Do notice the preventive and negative character of the three first principles (*not causing*, *preventing*, *removing*), and the positive but prudential attitude shown in the last one (*cultivating*, *enriching*, *but preserving*). Metaphysical entropy is the kind of loss, destruction, or damage caused on informational entities. The other way around, enriched, enhanced, or *augmented ethics* is intended to monitor the ecology of the infosphere, balancing the decreasing of *ontological friction*, and thus promoting the expansion and well-being of these entities in it.[10] Let us explain what this term is referring to, taking a short detour by what I will call the practical turn of IE.

The Internet and its semantic counterpart, Web 2.0 (social web) and 3.0 (the Internet of Things, the Semantic Web), should not be confused with the infosphere, in Floridi's usage. The former can be understood as technical descriptive concepts. The latter incorporates a *normative* dimension that is user-centered, focused on the creative and social behavior of agents (be they individual or collective), and it can be paraphrased as the ultimate boundaries of their *onlife* experience. And "hyperhistory [the kind of history inaugurated by the use of computational devices] happens onlife."[11] From this point of view, the infosphere is the natural housing of contemporary subjects or, better, of the bundle of information that constitutes them as such. Protecting the infosphere entails therefore to protect the *identity* of its more complex, self-reflecting, and conscious entities: "Selves are the final stage in the development of informational structures, for they are the semantically *structuring structures* [emphasis added] conscious of themselves."[12]

It means that, as macro-ethics, IE embraces innovative foundations (as opposed to deontologist, consequentalist, or contractualist ethics) and possesses a regulatory side. This is the practical turn. It only requires a minimal moral axiology and some procedural rules to be put in place. Good and evil will depend upon specific criteria along a gradual chain of moral value in a sort of "axiological ecumenism."[13] But, and this constitutes the IE turn-off from liberal enlightenment, Floridi resolutely embraces the non-standard, *allocentric* approach, already taken in bioethics, medical, and environmental ethics that "seek to develop a *patient-oriented ethics* in which the receiver of the moral action may not be only a human being, but also any form of life."[14] In a similar way, IE is centered on the informational entity that receives the action, rather than on its relation or relevance to the agent.

This patient-approach reveals particularly fruitful to draw the structure of moral agency, as moral agents are defined at an informational LoA of reality as "any interactive, autonomous, and adaptable *transition systems* that can perform moral justifiable acts."[15] But human individuals, artificial artifacts, institutions, and, especially, Multi-Agent Systems (MAS) are *accountable* (not always responsible) for the events and acts performed over other patient agents that receive the effects of their behavior.[16] IE makes a choice, then, in favor of the victims, the holders of rights, rather than fostering agency qua moral agency. I think this is particularly important to understand the kind of ontological turn that the author is proposing to ground ITC tools on the defense of rights and the correct and effective deployment and evolution of the infosphere. Thus, enhancing plurality is a better strategy than harmonizing such a deployment from a single or monotonic point of view.

The "tragedy of the Good Will"—the lack of balance between power and information—constitutes one example of such a perspective opened up by IE. Nowadays it is perfectly possible, for instance, as already happened in 2004 with the tsunamis caused by the Sumatra-Andaman earthquake, witnessing in real time their devastating effects without having the effective ICT tools to prevent them. Another privileged example is *informational privacy*, which is defined as follows: "informational privacy is a function of the ontological friction in the infosphere."[17] *Ontological friction* refers to "the forces that oppose the information flow within the infosphere, and hence (as a coefficient) the amount of work and effort required for some kind of agent to obtain, filter, and/or block information about other agents in a given environment."[18] According to Floridi, classic ITC tools used to decrease the degree of ontological friction among agents, and therefore decreased their level of privacy as well. However, new generation of protections—

through Privacy Enhancement Technologies (PETs) or the recent Privacy by Design Technologies (PbD)—are able to reverse the situation, balancing the level of protection with the increasing information flow. This certainly constitutes one of the next challenges in the development of the infosphere.

4. THE ONTOLOGICAL TURN

I think the author succeeds in convincing the reader of the interest of developing IE as a foundation for computer ethics, business ethics, and the kind of meta-theory that is needed to link moral codes and regulations to computer science. But the quest for a meaningful reality should not conceal some difficulties in this endeavor.

Perhaps the first one comes from the redefinition of some notions whose functional meaning has been already well established in artificial intelligence and engineering. This is the case, for example, with *Multi-Agent Systems* (MAS). The term usually refers to software agents in computerized systems composed of multiple interacting intelligent agents within a given environment. It is certainly true that MAS can be constituted by a combination of software and human agents. However, the author's use of the term seems to cover all sorts of social groups—institutions, government agencies, companies, NGOs, among other organizations. Although the intentional meaning of such a broad use is clear—mainly not to confuse moral agents with individuals, stressing its social component—the use of the term could be more consistently defined, as it has recently proposed relating to institutions and norms (the so-called nMAS).[19]

The more confusing notion is precisely *ontology*. In philosophy, this term does not belong to the rich Greek tradition. The term *ontology* goes back to the beginning of the seventeenth century in the Netherlands. Johann Clauberg, Rudolf Göckel (Goclenius), and Juan Caramuel de Lobkowitz used the term before Christian Wolff (*Philosophia prima sive ontologia methodo scientifica pertractata, qua omnes cognitionis humanae principia continentur*, 1730). According to recent research, it seems that it was first coined by Jakob Lorhard (*Ogdoas Scholastica*, 1606), and based at its turn on Clemens Timpler's work *Metaphysicae Systema Methodicum* (1604). *Ontology* was a term specifically born in the Protestant philosophical ambience to counterbalance the use of the term *metaphysics* (referred to the being) established by the formidable defenders of the Catholic Counter-Reformation, among them Francisco Suárez (*Disputationes Metaphysicae*, published in Mainz in 1606).[20] *Ontologia* meant the intelligible dimension of being and the organization of knowledge, thus stressing its human side. All of this is well-known and do not intend to lecture the reader.

I am referring to it because these philosophical origins are compatible—but not identical—with the use of the term in contemporary computer science.[21] *Ontologies*, in plural, are formal vocabularies plotting the machine scalable and reusable conceptual structures shared by a community of users to solve problems such as semantic interoperability and transportability.[22]

One of the pillars of Floridi's book is the thesis of *re-ontologization* of ICTs referred to "a very radical form of engineering, one that not only designs, constructs or structures a system (e.g. a company, a machine, or some artifact) anew, but one that also fundamentally transforms its intrinsic nature, that is, its *ontology or essence* [emphasis added]"[23]—e.g., the transition from analogue to digital data, the convergence between digital resources and digital tools, etc.

It seems to me that such a statement involves the description of social processes in a way that departs from the regular philosophical or computational use of the term.

To Floridi, practical use of ontologies in a globalized world deals more with communication and shared common references than with the technical possibility to interconnect formally semantic languages: "I am using 'ontology' to cover the outcome of a variety of processes that allow an agent to appropriate (be successful embedded in), semanticize (give meaning to and make sense of), and conceptualize (order, understand, and explain) her environment, through a wealth of levels of abstraction. In simplified terms, one's ontology is one's world: that is, the world as it appears to, is experienced by and interacted with, the agent in question."[24] This use reminds the use of Wittgensteinian concepts in hermeneutic sociology, as famously put forward, e.g., in the late fifties by Peter Winch (1958): cross-cultural communication would entail partaking a shared knowledge of the same world. But this interpretation has little to do with the technical means that make possible the cross-communication between natural and formal languages.

Both to explain or to design web services, platforms, and mobile applications in an iterative and feasible way, semantic web programming languages and *anchoring* the ontological level through editors such Protégé or Kaon on socially constructed contexts are needed. Of course this is not the purpose of the book: "Understanding philosophy as conceptual design means giving up not on its foundationalist vocation, but rather on the possibility of outsourcing its task to any combination of logico-mathematical and empirical approaches."[25]

The problem is that even accepting it, even if the philosophical analysis consists of an independent level by its own, some connection with its empirical assumptions and with the formal models related to empirical data (NLP, graph modeling, data mining, semantic statistical results, deontic and non-monotonic logical models, etc.) is still needed. How conceptual design relate to the social data that trigger and enable its construction? What kind of "structural coupling" could be put in place between the philosophical architecture of IE and their fields of application? Would it not be possible to connect some of the alleged concepts with social (ethical, legal, political) indicators?

As a matter of fact, Floridi's LoA seems to seek the connection among theoretical concepts, the philosophical layer, and empirical knowledge. To my view, Whitehead's fallacy of misplaced concreteness is accurately avoided, but if this inner connection cannot be made explicit with real

use cases, it is easy to fall into the fallacy of composition, especially because the relationship between micro and macro-ethics is not directly focused in the book, and the author shifts from individual to collective agency.

Stemming from the two last chapters—on distributed morality and global information ethics—there is still room to figure out a reasonable answer to such questions. The degree of ontological friction, or the degree of resilience or tolerance, appear especially apt to be measured. The same for the "moral enablers" (trust, transparence) on the sociological layer that the author calls *infra-ethics*. This layer could be coupled with some complementary ideas coming from the "science of the web" and on "web social machines."[26]

I will end my comments with a final statement on the digital divide and inequality. Reading these last chapters I have had the impression that some of the assertions on global information ethics were putting aside or shelving some facts that can show the present deep differences in accessibility to resources and potential pitfalls of the web. The infosphere is not the same for everybody. Let's look at some examples. According to the numbers that Cambridge mathematician Timothy Gowers made public on scientific research, the profit margins of the major commercial STM publishers, such as Elsevier, Springer, Wiley Blackwell, and Informa, are in the order of 35 percent.[27]

There are then severe restrictions to accessing knowledge for those that are not in the position to pay such high costs. Gathering and constructing reliable data on the web is also difficult, painful, and time consuming. Some years ago, chief scientist Kimberly Claffy, from UCSD-CAIDA Super-computation Center, wrote a seminal, sound, and well-informed article about the *economic and legal* obstacles she had to overcome to mapping the web: "Our scientific knowledge about the Internet is weak, and the obstacles to progress are primarily issues of economics, ownership, and trust, rather than technical."[28] I don't think the situation is much different now with the raising of big data. National governments, big companies, national citizens, and *digital neighbors*—people belonging to the *digital neighborhood* of crisis mappers, *NGO's* volunteers, etc.)—might not share the same interests and have different conflicting values. Since the Justinian Code, traditional regulations—laws, statutes, and rights—in Western culture are based on power, discrimination, and inequality (i.e., on the making of conceptual boundaries and differences among subjects, groups, elites, and classes grounded on and backed by the force of arms as last resort).

The author of the book could consider these barriers when balancing facts and values to apply information ethics principles. More changes are required related to the structure and composition of regulations, and what they mean to our culture, to face the challenges he is pointing out in the volume. Some kind of legal imagination should be at stake to cope with digital rights, rules, and norms. The very concept of law (and consequently the Rule of Law) cannot remain unchanged either, especially if the author is going to postulate a fundamental "*ontic trust* binding agents and patients," "a primeval, entirely hypothetical *pact*, logically predating the social contract."[29]

Why should this pact be binding? And how? And to whom? To some extent, Floridi is changing the rules of the game. The methodology of LoA and previous meta-theorical IE and agency schemes do not hold here. This kind of "pre-logical" and "hypothetical" explanatory constructs—such as the Kelsenian *Grundnorm*—were already postulated and discussed in similar terms by neo-Kantian and phenomenologist legal philosophers in the Weimar Republic, following the legacy of the German nineteenth-century dogmatic *Konstruktion*. The architecture and rational structure of the state and the difference between legality and legitimacy were one of the main topics in their updated discussions on the Leviathan. I do not think the author is really willing to start again with this kind of discourses to link IE to policies and legal issues. At a certain level, I think he will unavoidably have to face the problem of power within the infosphere. But this task is by far too complex to tackle without the aid of legal, policy, and economic analyses. I would encourage him to expand the scope of IE in this direction. Actually, a closer look at recent publications would reveal some steps.[30] With *The Ethics of the Information* he has already done a very good job. He is not in need of fictions. For the time coming, better to not wake up the sleeping dogs.

ACKNOWLEDGEMENTS

DER2012-39492-C02-01. *Crowdsourcing*. DGICYT (Spanish Ministry of Innovation and Competitivity); SINTELNET FP7-ICT-2009-C-286380 (7FP, EU Commission); CAPER, Grant Agreement 261712 (7FP, EU Commission).

NOTES

1. Luciano Floridi, *The Ethics of the Information* (Oxford: Oxford University Press, 2013).
2. Ibid., 42.
3. Ibid., 13.
4. Ibid., 222.
5. See Gregory Bateson, "The Birth of a Matrix," 39–64.
6. Floridi,*The Ethics of the Information*, 65.
7. Ibid., 98.
8. See Spinoza's Letters on Evil (correspondence with Blayenbergh, written between December 1654 and June 1656). Cfr. Gilles Deleuze, chapter 3, *Spinoza: Practical Philosophy*, 30–43. See also the hidden Leibnizian interest in Spinoza's Ethics in Matthew Stewart, *The Courtier and the Heretic*.
9. See, for all, Ernst Cassirer, *Die Philosophie der Aufklärung*.
10. Floridi, *The Ethics of the Information*, 204 and 160.
11. Ibid., 8.
12. Ibid., 227.
13. Ibid., 123.
14. Ibid., 62.
15. Ibid., 134.
16. Ibid., 158.
17. Ibid., 232.
18. Ibid.
19. G. Andrighetto et al., *Normative Multi-Agent Systems, Schloss Dagstuhl - Leibniz-Zentrum für Informatik GmbH*.

20. Cfr. the recent works of Peter Øhrstrøm and his team about it. E.g., Øhrstrøm et al., "Jacob Lorhard's Ontology," 74–87.
21. Nicola Guarino and Pierdaniele Giaretta, "Ontologies and Knowledge Bases. Towards a Terminological Clarification."
22. "A specification of a representational vocabulary for a shared domain of discourse—definitions of classes, relations, functions, and other objects—is called an "ontology." Thomas R. Gruber, "A Translation Approach to Portable Ontology Specifications," 199. For updated methodology on ontology building, see Mari Carmen Suárez-Figueroa et al., "The NeOn Methodology for Ontology Engineering," 9–34.
23. Floridi, *The Ethics of the Information*, 6.
24. Ibid., 298.
25. Ibid., 2.
26. Tim Berners-Lee et al., "A Framework for Web Science"; Jim Hendler and Tim Berners-Lee, "From the Semantic Web to Social Machines: A Research Challenge for AI on the World Wide Web."
27. Elsevier's profit was £826 million in 2013. Twenty Russell Group university libraries in the UK now pay Elsevier alone nearly £16 million per annum. Oxford University subsequently revealed it spends nearly £1 million a year with Elsevier (C. Steele, "Who Owns Scholarly Knowledge").
28. Kimberly Claffy, *Top Ten Things Lawyers Should Know about the Internet*, 2.
29. Floridi,*The Ethics of the Information*, 301.
30. Ibid.; Mariarosario Taddeo and Luciano Floridi, *The Ethics of Information Warfare*.

BIBLIOGRAPHY

Andrighetto, Giulia, Guido Governatori, Pablo Noriega, and Leendert W. N. van der Torre, eds., *Normative Multi-Agent Systems*. Schloss Dagstuhl - Leibniz-Zentrum für Informatik GmbH, Saarbrücken/Wadern, Germany: Dagstuhl Publishing, 2013. Available at http://www.dagstuhl.de/dagpub/978-3-939897-51-4.

Bateson, Gregory. *Steps to an Ecology of Mind: Collected Essays in Anthropology, Psychiatry, Evolution, and Epistemology*. Chicago: University of Chicago Press, 1972.

Bateson, Gregory. "The Birth of a Matrix or Double Bind and Epistemology." In *Beyond the Double Bind. Communication and Family Systems, Theories, and Techniques with Schizofrenics*, edited by Milton M. Berger, 39–64. New York: Brunner/Mazel, 1978.

Berners-Lee, Tim, Wendy Hall, James A. Hendler, Kieron O'Hara, Nigel Shadbolt, and Daniel J. Weitzner. "A Framework for Web Science." *Foundations and Trends in Web Science* 1, no. 1 (2006): 1–130.

Cassirer, Ernst. *Die Philosophie der Aufklärung* (1932). Hamburg: Meiner Verlag., 1998.

Claffy, Kimberly. *Top Ten Things Lawyers Should Know about the Internet. The COMMONS Initiative: Cooperative Measurement and Modeling of Open Networked Systems*. CAIDA: Cooperative Association for Internet Data Analysis, based at the San Diego Supercomputer Center at UCSD, 2008. Available at http://www.caida.org/publications/papers/2008/lawyers_top_ten/.

Deleuze, Gilles. *Spinoza: Practical Philosophy* (1970). San Francisco: City Light Books, 1988, pp. 30–43.

Floridi, Luciano. *The Philosophy of Information*. Oxford: Oxford University Press, 2011.

Floridi, Luciano. *The Ethics of the Information*, Oxford: Oxford University Press, 2013.

Floridi, Luciano. *The Fourth Revolution. How the Infosphere Is Reshaping Human Reality*. Oxford: Oxford University Press, 2014.

Gruber, Thomas R. "A Translation Approach to Portable Ontology Specifications." *Knowledge Acquisition* 5, no. 2 (1993): 199–220.

Guarino, Nicola, and Pierdaniele Giaretta. "Ontologies and Knowledge Bases. Towards a Terminological Clarification." In *Towards Very Large Knowledge Bases*, 25–32. Amsterdam: IOS Press, 1995.

Hendler, Jim, and Tim Berners-Lee. "From the Semantic Web to Social Machines: A Research Challenge for AI on the World Wide Web." *Artificial Intelligence* 174 (2010): 156–61.

Minsky, Marvin. *The Society of Mind*. New York: Simon & Schuster, 1976.

Naess, Arne. *Deep Ecology of Wisdom: Explorations in Unities of Nature and Cultures. Selected Papers*. In *The Selected Works of Arne Naess*. Vol. 10, edited by H. Glaser. Dordrecht, Heidelberg: Springer Verlag, 2005.

Øhrstrøm, Peter, Henrik Schärfe, and Sara L. Uckelman. "Jacob Lorhard's Ontology: A 17th Century Hypertext on the Reality and Temporality of the World of Intelligibles." In *Conceptual Structures: Knowledge Visualization and Reasoning: 16th International Conference on Conceptual Structures, Proceedings, ICCS 2008, Toulouse, France, July 7–11,2008, Proceedings*, edited by P. Eklund and O. Haemmerlé, 74–87. LNAI 5113. Heidelberg: Springer, 2008.

Simon, Herbert A. *The Sciences of the Artificial* (1969). Cambridge, MA: The MIT Press, 1981, 1996.

Steele, Colin. "Who Owns Scholarly Knowledge?" July 28, 2014. http://campusmorningmail.com.au/open-access-special/.

Stewart, Matthew. *The Courtier and the Heretic. Leibniz, Spinoza, and the Fate of God in the Modern World*. New York, London: W. W. Norton & Company, 2006.

Suárez-Figueroa, Mari Carmen, Asunción Gómez-Pérez, and Mariano Fernández-López. "The NeOn Methodology for Ontology Engineering." In *Ontology Engineering in a Networked World*, 9–34. Dordrecht: Springer Verlag, 2012.

Taddeo, Mariarosario, and Luciano Floridi, eds. *The Ethics of Information Warfare*. Law, Governance, and Technology Series no. 14. Dordrecht, Heidelberg: Springer Verlag, 2014.

Winch, Peter G. *The Idea of a Social Science and Its Relation to Philosophy*. London: Routledge & Kegan Paul, 1958.

Mary's Acquaintance

Peter Boltuc

UNIVERSITY OF ILLINOIS AT SPRINGFIELD AND AUSTRALIAN NATIONAL UNIVERSITY

This paper is largely about the Knowledge Argument. I argue that not all physical knowledge is easily accessible to human beings as knowledge by description. Human cognitive architecture is the cause of this lack, not limitations on physicalism.

The paper is also about the philosophical method. Computer science and neuroscience provide the methods that traditional philosophical approach, related primarily to semantics and philosophy of language, lacks and has a dire need for. This approach goes beyond computationalism, for many reasons;[1] the present article relies primarily on biologically inspired cognitive architectures (BICA).[2]

The article may seem like an argument in favor of reductive physicalism, but this is not the case. I only show that the knowledge argument against physicalism does not work. In order to clarify my broader position, briefly, I end with a surprise note in favor of non-reductive approaches. The non-reductive argument, sketched out here, is not focused on qualia but more directly on first-person consciousness.

This paper, and especially the postscript, is a bit of a pre-print, an invitation for discussion before the final formulation of my position emerges. Comments are very welcome.[3]

1. MARY'S ACQUAINTANCE WITH REDNESS

Assume the following:

1) Mary knows all there is to know (from black and white books) about colors. But she has never seen any colors.

2) She is let out of her black and white room and sees her first ripe red tomato. Probably she goes: *Oh, that's what red objects look like!*

3) Does Mary learn anything by seeing her first red tomato?

4) If she does, the argument goes, physicalism is false since not all knowledge is "physical knowledge."[4]

Discussions abound. First, it is argued that, although Mary knows everything there is to know about color from books, she still does not have all the propositional knowledge about color. In particular, Mary does not know how the concept red functions, since she misses some of the proper uses of this concept. (For instance, she does not know how to use it correctly to recognize colors.) This is Harman's functionalism of concepts, reiterated by others in various versions.[5] This argument questions premise 1: By knowing everything there is to know from books about colors, Mary does not have all the propositional knowledge about colors.

Second, we can have more than one version of physicalism. Conclusion (4) would be true of narrow physicalism, but is false of broad physicalism. I return to this issue in section two.

Third, what Mary learns can be viewed as an ability but not quite knowledge. This is the famous ability argument. I return to this issue in section three.

Finally, in section four, I show how the notions that come from cognitive science, especially from biologically inspired cognitive architectures, help us shed a new light on the whole knowledge argument. I argue that human cognitive architecture is such that we do not have access from propositional knowledge to entirely new cognitive qualities (such as colors).

Here is the gist of the broad message I want to convey: In acquisition of new psycho-motoric skills, such as riding a bicycle, we function more like a neural network (NN) than an artificial intelligence (AI) system. We can learn from training and not quite from a written set of instructions. In acquisition of new phenomenal knowledge we are more like sensors that need to be *initialized* by acquaintance with new properties than like AI or NN. It may or may not be essential whether we account for what we get acquainted with as just wave-lengths, the way a sensor operates, or as first-person phenomenal qualities, the way they present themselves to us. The relevant level of description is that we need to get acquainted with colors, just like some sensors, in order to perform color recognition or more advanced cognitive tasks that require qualia concepts.[6] The sensors that need to get initialized by acquaintance with properties they would later recognize do not seem like non-physical entities, emphatically so. Also, AI and NN cognitive architectures are, quite clearly, not counterexamples to physicalism. This provides partial support for broad physicalism since the above explanation of what Mary learns does not require us to assume any non-physical facts.

2. BROAD PHYSICALISM[7]

According to the standard versions of physicalism, all knowledge is expressed in physical language. What does it mean? It may mean either (A), the narrow theory, or (B), the broad theory:

A) All facts are expressible in propositional language. More specifically they are expressible in the language of science.

B) All facts are either expressible in propositional language (of science) or there is an explanation in propositional language (of present, or not too distant future science) why those remaining facts are not directly expressible in the language of science, and how they are consistent with science.

Thesis (B) includes thesis (A) plus an added on clause (B') that allows the lack of direct expressibility of some of the facts in the language of science. Those facts can be accessed indirectly by providing a scientific explanation of why we are unable to directly express them in propositional language.

One way, a rather minimal one, to satisfy (B') is by employing Stoljar's *missing concept strategy*.[8] The claim is that we do not currently have the concepts to express certain pieces of knowledge propositionally. This is an offshoot of Nagel's future knowledge claim, which assumes that currently our science is unable to explain many problems but should be able to do so in the future. (A sceptical version is McGinn's point that we may be unable to ever arrive at such knowledge).

Another way to satisfy (B') is to provide a scientific explanation why we are unable to cast certain facts in propositional language. Such explanations could be viewed as filling in the blanks in Stoljar and Nagel's claims: The concept, or knowledge, that is missing at time T may no longer be missing at some point T1 after time T, and the T1 may be now. This is my current claim. I think that AI in general, and biologically inspired cognitive architectures in particular, provide us with a good explanation of why human beings do not learn entirely new phenomenal concepts propositionally. This explanation has nothing to do with negating physicalism. Instead, it explains how and why human cognitive architecture does not include a link from propositions to qualia.

3. NON-PHYSICAL BIKERS

In this section I focus on the old argument that Mary does not obtain new knowledge upon her release since she only acquires a new ability, which is similar to the ability to ride a bicycle.[9] Such ability, the authors claim, is not propositional knowledge. The argument is kind of defunct by now,

criticized successfully on many fronts, so my primary goal in this section is not to demonstrate this point again. I provide what I hope is a somewhat new critique of the knowledge argument, inspired by AI and NN, for two reasons. A minor reason for presenting it is that the new way seems cool enough, but the main reason is broader. My discussion of the ability argument allows me to initiate a more general argument that shows how BICA makes broad physicalism a plausible solution of the knowledge argument.

As Lewis and Nemirow pointed out, riding a bicycle is not something one is likely to learn from a book. While propositional knowledge acquired from books, or by talking to somebody, may help you grasp the basics of riding a bicycle, it is very unlikely that one would grasp such an ability without practice. Human beings are not wired in a way that enables them to translate propositional knowledge directly into muscular coordination. Lewis and Nemirow do not view such ability as knowledge; they see all knowledge as propositional in a relatively narrow sense in which propositional knowledge is what you can write down in terms of propositional language. The claim seems to be that the ability to ride a bike is not a piece of knowledge just because it is not acquired propositionally. I think this presumption is false and the following case may help understand why:

> CASE 1. We can program a robot, Janis, to ride a bicycle. A standard artificial intelligence (AI) program can be viewed as a fully quantifiable algorithm (Pinkus). In any case, nobody of a sane mind would argue that a robot needs some non-physical information (whatever that would mean) to ride a bike.

If a robot acquires an ability to ride a bicycle on the basis of new information (a program to ride a bike), why wouldn't Mary? Defenders of the ability claim have a few responses to choose from. One is that the "knowledge" acquired by the robot is "non-propositional" in some sense. Yet, remember that the gist of the argument is not whether all knowledge is propositional, but whether it is "physical knowledge." If the ability argument shows that to ride a bicycle requires a physical (programmable) knowledge but that such knowledge is non-propositional, in some sense of the term, this conclusion seems outside of the scope of the present debate.[10]

The second potential response is this: If a human version of Janis was programmable like a digital robot, then definitely she would have found the right piece of programming in her black and white library and off she goes biking. (If Mary were like Janis, she would have found a piece of AI program that allows her to link the word red—or some broader propositional knowledge of red things—with a certain phenomenal experience instantiated in the part of her brain responsible for producing such experiences. We discuss this point later on.) Hence, the argument continues, humans are unlike robots in this respect. This is a fair try but it does not work as a defense of the ability thesis. Case two demonstrates why.

> CASE 2. We can program a different robot, Alice, using a different way of programming, the neural networks programming (NN). This kind of programming requires practice; it builds neural connections based on various responses and interactions with the environment. Would anybody argue that neural networks programming is "non-physical"?

While people may worry whether NN programming is efficient, the view that this method of programming is somehow non-physical, or non-material, would make little sense. Yet, without such a claim the knowledge argument may turn out hard to defend. If so, by a close analogy, if neural networks require training as a way of programming, why wouldn't training humans count as a method of programming as well?

The third potential response breaks the link between riding a bicycle and phenomenal knowledge. Riding a bicycle is a programmable skill whereas one could argue that seeing phenomenal qualities of experience, such as redness, is not. This may work as a response to this particular objection. Nevertheless, it is detrimental to the ability argument. The ability argument, in its canonical Lewis-Nemirow form, does not get off the ground without the assumption of the analogy between acquiring phenomenal knowledge and other abilities such as riding a bike or playing tennis. In the next section we examine the knowledge argument using the same strategy as in the cases of Janis and Alice.

4. THE MISSING LINK IN HUMAN COGNITIVE ARCHITECTURE

The knowledge argument assumes that if Mary learned something upon seeing her first colored object then physicalism is false. Yet, not all physical knowledge must be conveyable in propositional language.[11] The example of various cognitive architectures in computer engineering—especially biologically inspired cognitive architectures—is one way to demonstrate this point.

Let me put the points we learned in section three in a broader context. Human cognitive architecture allows us to learn many things directly from propositional attitudes. We learn things from talking to parents, teachers, reading books, and checking things on the net. But there are things that we need to learn by training, or direct acquaintance. We learn advanced psycho-motoric skills, such as the ability to walk, swim, play tennis, or ride a bicycle by training. Learning skills through training is common to humans and our evolutionary ancestors. There is nothing metaphysically unique about learning by doing, of course. It would in principle be possible to describe (in propositional language) how to do those things and then do them correctly. Yet, human beings, or at least 99.9 percent of us, are unable to follow psycho-motoric instructions without trying, often for an extended period of time. One could write an interactive computer program that tells us what to do in relation to specific actions we perform. A program of this kind would include ostensive and indexical terms, such as "now press the pedal more strongly with your left foot." Those terms are propositional but they are not instructions of the kind

that can be learned from a book, in advance. This method of learning is exactly the way neural networks learn—by training, trial and error—the way most animals do.

Learning psycho-motoric skills is not exactly the same thing as getting acquainted with radically new qualities of experience. Human beings learn new secondary qualities, such as colors, sounds, and smells, by acquaintance. Here our cognitive architecture seems even more dependent on direct contact with the object. While in the case of psycho-motoric skills learning them entirely from a textbook but never even practicing would be very difficult, it is not clear whether learning phenomenal qualities of experience from their description would even be possible. Learning a psycho-motoric ability is more like combining already acquainted pieces of knowledge, whereas learning phenomenal qualities of experience is an entirely new kind of knowledge.

If we were to accept that learning entirely new phenomenal qualities by description is possible, it would entail acceptance of a strong version of physicalism. In particular, we would be committed to accepting the claim that "*a priori* knowledge alone, plus knowledge of the physical truths will allow one to know the mental truths," which is what Stoljar calls the fourth notion of reductionism.[12] There are two versions of this account of physicalism:

A) Mental truths are defined functionally. If I know the notion red then I can detect red objects. (*That's all there is to a mental truth.)*

B) The phenomenal definition of mental truths: If I know the notion of red (for instance, from reading a book) then I can get the first-person feel of red. (*This is the non-reductive version of strong physicalism.*)

It is easy to imagine a cognitive system that satisfies the first, functional, version. We can have a robot, or even a simple sensor, that is programmed to recognize green objects *by acquaintance*; it needs to be first "initialized" by sensing *the paradigmatic case(s) of greenness*. Only then would it be able to recognize other objects as being green. Human beings share this element of cognitive architecture with such a sensor. The analogy holds as long as we care solely about the functional aspects of learning.

The phenomenal definition requires a broader picture in which there are *first-person feels* of things. In fact, something or someone could function in such a way that knowledge of the physical truths, plus the *a priori* knowledge, would allow one to know the mental truths in the first-person, phenomenal way. Yet, human cognitive architecture does not allow us to figure out the first-person feel of secondary qualities from their physical description.

Here come the punch-lines of the argument: In version A (functional definition), if the sensor is material, there is no reason to think that human cognitive architecture is not material just because it gains some knowledge from first-person acquaintance. Version B (phenomenal definition) may be analyzed in a similar way. The fact that human beings have no direct connection in the brain between knowledge of color frequency and the phenomenal experience of it provides a sufficient explanation of why Mary was unable to figure out what red tomatoes look like before seeing one. This is not an argument against broad physicalism. It is a fact about human cognitive architecture that we have no direct link between the centers in the brain that acquire "physical knowledge" about qualia and the centers that produce first-person phenomenal experiences. One could surely design such a connection and try to implant it into (or bioengineer it within) human-like brains. The fact that we do not grow such connections has clear evolutionary causes since propositional knowledge is evolutionarily much newer than color, smell, or sound recognition. This fact should not be viewed as lending any support for the critique of physicalism.

We can see the conundrum of how to view the knowledge by acquaintance within physicalist framework as dissolved by the current understanding of biologically inspired cognitive architectures. In fact, Stoljar anticipated a similar argument though he did not relate it to artificial consciousness.[13] Stoljar presents Descartes's version of the conceivability argument:

> We can certainly conceive of a machine so constructed that it utters words, and even utters words which correspond to our bodily actions causing a change in its organs (e.g. if you touch it in one spot, it asks what you want of it, if you touch it in another it cries out that you are hurting it, and so on). But it is not conceivable that such a machine should produce different arrangements of words so as to give an appropriately meaningful answer to whatever is said in its presence, as the dullest of men can do."[14]

This leads Stoljar to conclude that, according to Descartes, "it is inconceivable that a (mere) machine however complicated should speak a language."[15] Then Stoljar asks,

> What explains the fact that a conceivability argument which seemed powerful to Descartes seems to us to be not powerful at all? [. . .] a reasonable hypothesis is that Descartes was operating with what is from our point of view a rather impoverished conception of the physical world. In particular, he did not have access to the ideas of information processing and computational complexity which apply to the physical world, even if in a certain sense they are not part of physics.[16]

The gist of strategy of Stoljar's *missing concept strategy* "is that our contemporary position with respect to experience is parallel to Descartes's position with respect to language." And Stoljar's argument continues, "the reasoning goes wrong because the conceivability claims on which it is based are generated by a conception of the physical world from which crucial concepts are missing."[17] Stoljar points out here, indirectly, to something like Nagel's future knowledge argument which can be interpreted as pertaining to a broad solution to the mind body problem (or, to the hard problem of consciousness). But Stoljar's

response to Descartes's point was appropriately local—our understanding of *information processing and computational complexity* does not solve the issue of dualism, it just makes possession of language less of a conundrum than it was in the 1600s.

Let me restate my argument here. The knowledge of biologically inspired cognitive architectures gives us an avenue to avoid puzzlement over the fact that human beings can learn some things by propositions (the way AI learns) and some other things, such as psycho-motoric skills and radically new perceptual qualities through training and direct acquaintance respectively. We could, in principle, have a cognitive link from propositional knowledge to the parts of the brain that produce visual, tactile, and olfactory phenomena, but we do not have such a link. The lack of this link can be understood based on our evolutionary history, since propositional learning is an evolutionarily new skill and acquaintance with qualities of experience is an ancient one that we share with our distant ancestors. Hence, the knowledge argument has no bite against broad physicalism. It does show, however, that not all human knowledge is knowledge by description, as Flanagan has been claiming for several years. The analogy with biologically inspired cognitive architectures demonstrates how this is the case.

In conclusion, The knowledge argument does not show that physicalism is false. Instead, it demonstrates that human cognitive architecture does not allow us to go from the kind of statements one can read in a book to imagining phenomenal properties (for instance, to visualizing a given color).

This statement may help us understand that we should scratch a little deeper in search of the hard consciousness (like the hard problem of consciousness). We may need to go beyond qualia for doing so. In the postscript I sketch out how we may be able to go beyond qualia, and in fact beyond any functional features of consciousness, in defending non-reductive consciousness.[18]

5. POST-SCRIPT IN FAVOR OF NON-REDUCTIVE CONSCIOUSNESS

I promised, in the introduction, to end this paper with a note in favor of non-reductive approaches to consciousness. This is important since the argument presented in the main body of this article shows that the knowledge argument does not work against reductive physicalism, and it may suggest that my goal is to defend such a view. My goal is to show that a functionalist argument, such as the knowledge argument, is unlikely to work since the hard problem of consciousness can be demonstrated only at the level of generality that goes beyond any functional arguments. The point is difficult to present within the standard Anglo-American tradition, though various versions surfaced over the years.[19] The argument has been made within German philosophy, primarily in Fichte's theory of knowledge but also in Husserl's *Ideas*.

Non-reductive approaches are often based on qualia—more specifically, upon first-person qualities of experience. I think that *internal* objects of experience are still *objects* of experience and should be viewed as such. There are well-known problems with privileged access to one's own qualia. It can be shown that the privileged access problem really boils down to the problem of other minds. Yet, the problem of other minds may be solvable by empirical means. There is nothing contradictory about G. Tononi, or somebody else, providing a machine able to put whatever one sees, or imagines, onto a computer screen. Such a machine would constitute an advancement within what Chalmers calls the easy problem of consciousness since it would, no doubt, operate at the level of neural correlates of consciousness.

If we take seriously the fact that non-reductive consciousness (*pure subject* in Fichte's terminology[20]) is not an object, we should not expect any function-based cases, such as the knowledge argument (or the case of zombies, or the Chinese room) to be successful in establishing non-reductionism. They may be no more than intuition pumps, sort of like Nagel's "What is it like to be a bat?" question. Instead, we should work with the notion of pure subjectivity (Fichte, Globus) viewed as the *locus of consciousness* (Shalom). As in arithmetic, we may move all the objects, including the objects of first-person experience, to one side of the equation—to its objecthood-side. We group there any objects—the objects of thought, of perception, of imagination and recollection and those with yet unclear status to us.[21] This is a bit like the old-style constitution of reality from sense-data[22] but without the assumption that sense-data are the bricks of our epistemic knowledge (perceptions may come in larger units as *Gestalt* psychology has demonstrated). What remains on the other side of the equation is pure consciousness. It needs some relation to an object to manifest itself; only then may we put forth the regular predicative claims about it.

My claim is that the kind of consciousness that is the center of Chalmers's hard problem of consciousness (pure subject) is not an object since it is non-predicative. It can be thought of as potential subjectivity that manifests itself only when it enters into relation with objects. It has to enter into such relation in two ways: First, it needs to be instantiated, or actualized, in the objects such as brains. Second, it requires objects it projects (objects of perception) that it is conscious of. Those can be objects of any ontological status such as dreams, movies, holograms, physical objects, etc. This second requirement demonstrates that pure consciousness is merely *a potential consciousness* as long as it does not enter in relation with objects; in particular with objects of consciousness.

All functionalities of such consciousness may be imitated by AI, which I think follows from physical interpretation of the Church-Turing thesis. So, we may ask Chalmers's famous question: Is there a non-epiphenomenal difference between a world X where everybody is just a zombie and a world Y where there is first-person consciousness? Chalmers poses it as an argument in favor of non-reductive views on consciousness, but in the present context it is viewed primarily as an argument about epiphenomenalism. I think there are two points to be considered. First, physicalism assumes that there is no difference without physical difference so the physical instantiation of those zombies

would have to be different from non-zombies; otherwise, the argument begs the question against physicalism.[23] Hence, there would be differences in cognitive architecture of beings with and without first-person consciousness, and we should be able to see which ones have it and which ones lack it.[24] But this argument is not relevant to the issue whether non-reductive consciousness is epiphenomenal.

I have one indirect argument why it makes a difference whether a being functionally identical with a human being has or does not have first-person consciousness. The argument is indirect since it relates to human values and interests.[25] One is supposed to care about how his or her friend or significant other feels. In the clearest case, imagine a difference between a person A, an actual lover with first-person consciousness, and a "person" B (whom I define elsewhere as a *Church-Turing lover*) who is just like A but lacks first-person consciousness.[26] An important part of friendship, and sexual relationship, at least for many people, involves caring how the other partner feels. Yet, if the other partner feels nothing then this caring could never be satisfied even if the partner behaves exactly the way one would if he or she did have first-person experiences. Dennett and others would likely reject this claim, but this is largely due to their background in verificationist methodology. Those philosophers who are willing to follow inference to the best explanation in the instance of inductive arguments for first-person awareness tend to consider that it is likely that other persons have first-person consciousness since they have it themselves. This move retorts against the problem of other minds; moreover, it helps establish good reasons to believe in non-reductive consciousness.

Whether there is a subjective stream of experiences, and in particular pure consciousness, is important for relating to humans and even pets, since to care about them is oftentimes to care about how things feel for them. If, in such relationships, a person cares only about how friends, lovers, and even pets behave towards her, not what they feel in their first-person consciousness, then she misses at least a part of the meaning of friendship and love. First-person subjectivity is the only level at which such care makes full sense—all kinds of experiences, even internal ones, can be implemented (and their content explained) in functional terms. The one thing that makes them non-reductive is that they are experienced by first-person consciousness.

ACKNOWLEDGEMENTS

The final version of this paper crystalized in July 2014 during the early stage of my fellowship at Australian National University. I am grateful to Dave Chalmers, Frank Jackson, Dilectiss Liu, Dan Stoljar, and Giulio Tononi for conversations, and for comments on its various versions. I am also indebted to the reviewer for this newsletter. The argument was presented at the APA Central Division meeting in Chicago, Illinois, February 17, 2012. I am grateful to other participants, especially Terry Horgan, for the comments. A short version was published in (Boltuc 2013); for my earlier take on the problem see (Boltuc 1998a, 1998b). I want to thank the Hartman-Schewe Endowment at the University of Illinois at Springfield for providing funds for the current project, and the Department of Philosophy at ANU for hosting me when I wrote the final version. I am also thankful to the Department of Philosophy at Poznan University (UAM) for inviting me to offer a Ph.D. seminar and a series of graduate lectures on closely related topics as a part of my Visiting Professorship.

NOTES

1. Consciousness in the sense used by non-reductive theories is not just a computational function but something "inlayed in nature." Nicholas Boltuc and Peter Boltuc, "Replication of the Hard Problem of Consciousness in AI and Bio-AI: An Early Conceptual Framework."
2. There is much philosophically interesting stuff to learn from imagitrons (discovery machines that exhibit independent creativity, Stephen L. Thaler, "Synaptic Perturbation and Consciousness") and other developments in applied science of consciousness, although these topics go beyond this article.
3. My email address is listed on my website: https://sites.google.com/site/peterboltuc/home.
4. Frank Jackson, "Epiphenomenal Qualia"; Jackson, "What Mary Didn't Know."
5. For Harman's functionalism of concepts, see Gilbert Harman, "The Intrinsic Quality of Experience"; and Harman, "Explaining the Explanatory Gap." Other versions include Scott Sturgeon, "The Epistemic View of Subjectivity"; Brian Loar, "Phenomenal States"; Christopher S. Hill, "Imaginability, Conceivabilty, Possibility, and the Mind-Body Problem"; A. Byrne, "Cosmic Hermeneutics"; Tye, "Knowing What It Is Like: The Ability Hypothesis and the Knowledge Argument"; and John Perry, *Knowledge, Possibility, and Consciousness*.
6. I apply a version of Bas van Fraassen's anti-realism; whichever way the properties present themselves is just a subjective/phenomenal story that is a part of our recognition of colors.
7. In previous work I used the term naturalism in a similar context; Marcin Milkowski, the organizer of yearly workshops on naturalism, may have persuaded me to do so. Recently, Dan Stoljar presented me with better reasons to use term physicalism—the main reason is that this term is standard in the current debate (the term is no longer closely linked with reductive materialism).
8. Daniel Stoljar, "Physicalism and Phenomenal Concepts"; Stoljar, *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*.
9. David Lewis, "What Experience Teaches"; Laurence Nemirow, "Physicalism and the Cognitive Role of Acquaintance."
10. The claim that not all "physical" knowledge is propositional (in some sense of propositionality) is relevant to the theory of propositional knowledge but not directly to the debate between reductive and non-reductive theories of consciousness.
11. Owen Flanagan,*Consciousness Reconsidered*; Peter Boltuc, "Reductionism and Qualia"; Torin Alter, "A Limited Defense of the Knowledge Argument."
12. Daniel Stoljar, "Physicalism." It seems impossible, to me, to learn completely new qualities, such as to learn a color for a person who has never seen any. It seems quite doable to use one's imagination in order to blend already known qualities, for instance, to imagine something like orange for a person who knows what yellow and red things look like.
13. Daniel Stoljar, "The Conceivability Argument and Two Conceptions of the Physical."
14. René Descartes, *Selected Philosophical Writings*, 44.
15. Stoljar, *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*, ref. 30.
16. Stoljar, *Ignorance and Imagination*.
17. Ibid.
18. For a broader argument, see Peter Boltuc, "The Philosophical Problem in Machine Consciousness," sec 4; and Boltuc and Boltuc, "Replication of the Hard Problem of Consciousness in AI and Bio-AI."
19. Gordon Globus, "Mind, Structure and Contradiction"; Albert Shalom, *The Body/Mind Conceptual Framework and the Problem of Personal Identity: Some Theories in Philosophy, Psychoanalysis, and Neurology*.
20. The term goes back at least to Kant but has a somewhat different meaning in his philosophy.

21. Sometimes it is unclear whether an object in our mind is a perception of actual tulips, a recollection, a dream, perception of a hologram, etc.

22. H. H. Price, *Perception*.

23. Daniel C. Dennett, *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*.

24. Peter Boltuc, "The Engineering Thesis in Machine Consciousness."

25. There is one more argument in related literature that may be relevant in the present context: Peter Unger's. We have two ideal twins with the same content of consciousness until now—one in room 1 and the other in room 2. If I am one of those persons, it makes sense to care whether person in room 1 or person in room 2 is about to be tortured. This argument, originally aimed against Parfit's stream on consciousness view on personal identity (and the claim that personal identity does not matter), can be used to show why, in one sense, non-reductive consciousness is not epiphenomenal. It is not epiphenomenal, at least in some sense of the term, since it does matter to me whether I am the object with numerically this first-person stream of consciousness or the other one. In Unger's case, it actually does not matter whether the other person has hard consciousness or not at all. My argument differs from Unger's in an important way since it is about caring about what, if anything, happens in/for the other person's stream of consciousness, but in both senses there is a sense in which that first-personal stream of consciousness matters and so is not completely epiphenomenal.

26. Peter Boltuc, "What Is the Difference between Your Friend and a Church Turing Lover?"

BIBLIOGRAPHY

Alter, Torin. "A Limited Defense of the Knowledge Argument." *Philosophical Studies* 90, no. 1 (1998): 35–56.

Boltuc, Nicholas, and Peter Boltuc. "Replication of the Hard Problem of Consciousness in AI and Bio-AI: An Early Conceptual Framework." In *AI and Consciousness: Theoretical Foundations and Current Approaches*, edited by Antonio Chella and Riccardo Manzotti, 24–29. Menlo Park, CA: AAAI Press, 2007. Also available online at http://www.consciousness.it/CAI/online_papers/Boltuc.pdf.

Boltuc, Peter. "Reductionism and Qualia." *Epistemologia* 4 (1998): 111–30.

———. "Qualia, Robots and Complementarity of Subject and Object." World Congress of Philosophy. Boston, 1998. http://www.bu.edu/wcp/Papers/Mind/MindBolt.htm.

———. "The Engineering Thesis in Machine Consciousness." *Techné: Research in Philosophy and Technology* 16, no. 2 (2012): 187–207.

———. "The Philosophical Problem in Machine Consciousness." *International Journal of Machine Consciousness* 1, no. 1 (2009): 155–76.

———. "What Is the Difference between Your Friend and a Church Turing Lover?" In *The Computational Turn: Past, Presents, Futures?*, edited by Charles Ess and Ruth Hagengruber, 37–40. Aarchus University, 2011.

Byrne, Alex. "Cosmic Hermeneutics." *Philosophical Perspectives* 13 (1999): 347–83.

———. "Something about Mary." *Grazer Philosophische Studien* 62 (2002): 123–40.

Conee, Earl. "Phenomenal Knowledge." *Australasian Journal of Philosophy* 72 (1994): 136–50.

Chalmers, David. *The Conscious Mind. In Search of a Fundamental Theory*. Oxford: Oxford University Press, 1996.

Dennett, Daniel C. *Sweet Dreams: Philosophical Obstacles to a Science of Consciousness*. Cambridge, MA: The MIT Press, 2005.

Descartes, René. *Selected Philosophical Writings*, edited. by J. Cottingham. Cambridge: Cambridge University Press, 1988.

Flanagan, Owen. *Consciousness Reconsidered*. Cambridge, MA: The MIT Press, 1992.

Globus, Gordon. "Mind, Structure and Contradiction." In *Consciousness and the Brain: A Scientific and Philosophical Inquiry*, edited by Gordon G. Globus, Grover Maxwell, and Irwin Savodnik, 269–94. Springer Verlag, 1976.

Harman, Gilbert. "The Intrinsic Quality of Experience." *Philosophical Perspectives, Action Theory and Philosophy of Mind* 4 (1990): 31–52.

———. "Can Scinece Understand the Mind?" In *Conceptions of the Human Mind: Essays on Honor of George A. Miller*, 111–21. Lawrence Erlbaum, 1993.

———. "Explaining the Explanatory Gap." *APA Newsletter on Philosophy and Computers* 06, no. 2 (2007): 2-3.

Hill, Christopher S. "Imaginability, Conceivability, Possibility and the Mind-Body Problem." *Philosophical Studies* 87 (1997): 61–85.

Horgan, Terence. "Jackson on Physical Information and Qualia." *Philosophical Quarterly* 34, no. 135 (1984): 147–52.

Jackson, Frank. "Epiphenomenal Qualia." *Philosophical Quarterly* 32 (1982): 127–36.

———. "What Mary Didn't Know." *Journal of Philosophy* 83 (1986): 291–95.

Lewis, David. "What Experience Teaches." *Proceedings of the Russellian Society* 13 (1988): 29–57.

Loar, Brian. "Phenomenal States: Second Version." In *The Nature of Consciousness: Philosophical Debates*, edited by Ned Block, Owen J. Flanagan, and Güven Güzeldere, 597–616. MIT Press, 1997.

Ludlow, Peter, Yujin Nagasawa, and Daniel Stoljar, eds. *There's Something About Mary: Essays on Phenomenal Consciousness and Frank Jackson's Knowledge Argument*. Cambridge, MA: The MIT Press, 2004.

Lycan, William G. *Mind and Cognition*. Oxford: Wiley-Blackwell, 1990.

Nemirow, Laurence. "Physicalism and the Cognitive Role of Acquaintance." In *Mind and Cognition*, edited by William G. Lycan, 490–99. Oxford: Wiley-Blackwell, 1990.

Nida-Rümelin, Martine. "Qualia: The Knowledge Argument." *The Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. Stanford, CA: The Metaphysics Research Lab, 2010.

Perry, John. *Knowledge, Possibility, and Consciousness*. Cambridge, MA: The MIT Press, 2001.

Pinkus, Gadi. "Propositional Logic, Non-Monotonic Reasoning and Symmetric Networks – On Bridging the Gap Between Symbolic and Connectionist Knowledge Representation." In *The Neural Networks for Knowledge Representation and Inference*, edited by Daniel S. Levine and Manuel Aparicio IV, 175–204. Hillsdale, NJ: Lawrence Erlbaum Associates, 1994.

Price, H. H. *Perception*. Methuen and Co., 1933.

Shalom, Albert. *The Body/Mind Conceptual Framework and the Problem of Personal Identity: Some Theories in Philosophy, Psychoanalysis, and Neurology*. Humanities Press, Atlantic Highlands, 1985.

Stoljar, Daniel. "Physicalism." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. First published February 13, 2001; substantive revision September 9, 2009. Accessed October 15, 2014. http://plato.stanford.edu/entries/physicalism/.

———. *Ignorance and Imagination: The Epistemic Origin of the Problem of Consciousness*. Oxford: Oxford University Press, 2006.

———. "Physicalism and Phenomenal Concepts." *Mind and Language* 20, no. 5 (2005): 469–94.

———. "The Conceivability Argument and Two Conceptions of the Physical." *Philosophical Perspectives* 15 (2001): 393–413.

Sturgeon, Scott. "The Epistemic View of Subjectivity." *Journal of Philosophy* 91, no. 5 (1994): 221–35.

Thaler, Stephen L. "Synaptic Perturbation and Consciousness." *International Journal of Machine Consciousness* 6, no. 2 (2014): 75–108.

Tononi, Giulio. *Phi: A Voyage from the Brain to the Soul*. Pantheon Books, 2012.

Tye, Michael. "Knowing What It Is Like: The Ability Hypothesis and the Knowledge Argument." 1999. Accessed October 15, 2014. http://www.nyu.edu/gsas/dept/philo/courses/consciousness97/papers/tye/ability.html.

———. "The Subjective Qualities of Experience." *Mind* 95, no. 377 (1986): 1–17.

Scientific and Legal Theory Formation in an Era of Machine Learning: Remembering Background Rules, Coherence, and Cogency in Induction

Ronald P. Loui
UNIVERSITY OF ILLINOIS, SPRINGFIELD

WE'RE NOT IN VIENNA ANYMORE

Modern machine learning attempts to erase all distinctions between various kinds of induction from data, as if all learning from cases can be expressed as the selection of features followed by the determination of an appropriate set of separating hyperplanes in that feature space, on which to build a statistically effective classifier. In this recent research program that has dominate AI (artificial intelligence), the intellectual question is no more than what is the preferred mathematical transformation of feature space that permits convenient re-representation of data—e.g., What is your favorite kind of nonlinear support vector kernel?

Few things could be more historically or intellectually insulting to the student of theory-formation in the grand tradition of Nagel, Carnap, Hempel, and Popper.[1] There are different purposes for different kinds of induction and in a few cases—notably, the cases of scientific induction and legal induction—the modeling of the nuances has led to great ideas in the history of logic and reasoning. Many philosophers will not know the work of Hart, Raz, MacCormick, and Schauer on legal theory-formation (the "positivists" in law), but the domain-specific complexities are arguably even more interesting there.[2]

Artificial intelligence at one time had more interest in these nuances, for example, in the works of Simon, Lenat, and Gennari-Langley-Fisher on "automated scientific theory formation."[3] Rule-formation and rule-extraction were topics of study, even in applied expert systems. Case-based reasoning was concerned with representing and transforming structure, not just the greedy-algorithmic selection of correlates and eigenvectors. DeJong and Mooney, for example, insisted on an explanation for the relation of feature to forecast: an explanation, not just a correlation.[4] Artificial intelligence and law is a sub-area of AI where modeling the structure of the case, and reasoning from the case with explanations, based on what humans actually appear to do (or say they do), have always been part of the program.

Then came classifiers. The neural network was the beginning of a kind of anti-intellectual, "close your eyes, do a lot of training, and see what results" method of using data to model phenomena. The statisticians were aghast (I witnessed a red-faced senior scholar bolting from an academic colloquium crying, "Call the neural network what it is! It's just one kind of nonlinear regression!" – P. H. Dybvig). Computation would replace the art of selecting a theoretical framework. Fitness would be considered entirely measurable, and only the ones who could compute the best fit would survive.

There had been precedents (perceptrons and flowchart-like decision trees), and competing methods (Bayesian influence graphs), but there was nothing like the change of view brought by neural nets. Researchers even looked for "explanations" among the trained neural classifiers, as an ex-post-facto olive branch in exchange for cutting down the olive tree of understanding. One can see how this is possible in a feed-forward neural net, where sharp sigmoids can be turned into fuzzy binary variables, where other activation functions represent linear mixes, and where strong connectivity equates to joint influence. My own doctoral student, Gadi Pinkas, began the practice of extracting logical constraints from Hopfield networks.[5] But even this is explanation post hoc, not theory formation with an understanding of the phenomenon as a mental guide.

In contrast to medical diagnosticians who sought to understand pathways, and weather forecasters who modeled physical processes, and sports writers who sought to give reasons for a team's success or collapse, a NASA Bayesian, Peter Cheeseman, once told me that he had so much data and so much computing power, it didn't matter where his priors came from.

This reductivist travesty of induction is not actually the fault of the vector representation and computational emphasis of the machine learning field. Computations over mathematically convenient representations have proven to be valuable exercises, especially when they enable easy automation, improve the clarity of discussion, or make genuine discovery of unforeseen covariance or unanticipated irrelevance. The intellectual vacancy we see in machine learning is the result of an unwillingness to study the phenomenon first, before doing the mathematical modeling.

This has also been machine learning's strength. The classifier or predictor derives from the data at hand, or from a principal components analysis, not from the tradition of an entrenched paradigm or mental model. There is no question that a non-specialist, unbiased, algorithmic processing of data has led to methods of knowledge discovery and datamining (KDDM) that have transformed applied statistics and optimization. For one thing, it relieves the theoretician of assuming linear functions or Gaussian distributions, or limiting attention to a small number of separable influences, when those assumptions are made for simplicity and human calculation (or worse, to enable closed-form derivation—a good reason for making sure your functions are all familiar differentiable forms). It also means that the biostatistician needs to know less biology, the wolves on Wall Street need to know less about what corporations do, and the litigator needs to read fewer cases in exchange for statistical behavioral analyses of particular judges.

In this paper, I want to remind the readers what kinds of considerations have been lost in the rush to embrace machine learning and KDDM. My objective here is to give Kyburg's account of scientific theory-formation, the best

account I have seen, in language that would make sense to a machine learning researcher.[6] Then I will give my own account of legal theory formation in similar language.[7]

It is also an opportunity to contrast these two major forms of theory formation: one based on constraint, probability, control of error, and inference; the other based on analogy, defeasibility, accomodation of exception, and coherence. It is a conversation that I began with John Pollock and Henry Kyburg, over two decades ago, about rule-formation for defeasible and non-defeasible rules.

The essential difference with contemporary KDDM is that the target representation is a logical language, not a mathematical function weighing dimensions, or a classifying process applied serially to features. There are many interesting nonlinearities that logical language can introduce. In many ways, the contrast between Cyc and KDDM raises the same debate ("real artificial intelligence" is harder because there is more to intelligence than getting good I/O behaviors: internal structure matters).[8] Inductive logic programming seems to occupy a middle ground here.

Philosophers' theory formation models go beyond even the induction of logical rules. What is most important to me is that the intellectual traditions of theory-formation, which had made such progress in the past half-century in philosophy and legal philosophy, do not disappear.

A BASIC CLASSIFIER AND BASIC CURVE FITTING

The basic collection of observations, *O*, contains feature-dimensions and predicted dimensions. In the simplest, degenerate case, there is one feature, with an integer-valued or real-valued variable, *f* (we won't make a big distinction between the name of the feature and the variable representing its value), and a binary prediction, *p* or *not-p*. If the value *f* is high enough, *p*. Otherwise, *not p*.

The inductive problem is to look at *O*, a collection of prior $\langle f, p \rangle$ values, $O = \{ \langle f(i), p(i) \rangle \mid i=1..n \}$ of *n* observations, and find a threshold *t* which minimizes predictive error. In this degenerately simple case, *t* is your induced theory or predictor. Errors come in two different kinds: *p*-classified-as-*not-p*, and *not-p*-classified-as-*p*, your type I and type II errors. This is like setting the A-/B+ line in a list of numeric grades, or at least predicting where the instructor will set such a line.

In a slightly more interesting case, in two dimensions, the problem is familiar to high school students as finding a separating line between points in a plane: *f* is now a vector in two dimensions, *f* in $R \times R$, e.g., $f(i) = \langle x(i), y(i) \rangle$, and *O* can be plotted as a cloud of points. Some points are red (those that are *p*) and some are black (those that are *not-p*), and we seek a line, or some kind of curve, that separates red and black points with fewest exceptions (Figure 1). This is like deciding who has acceptable body-mass-index, or not, based on height and weight.

If the curve has enough parameters, it is usually possible to do this with zero error. For a line, it is almost always the case that there will be some "errors" in classification. Thus one learns the trade-off between simplicity and predictive power for the first time. Here is where knowing something about the phenomenon in question can help determine how simple a curve is required. Even without meta-knowledge constraining the number of parameters and the family of functions, the student learns immediately that the fewer the parameters, the greater the predictive value of the theory: adopting as a rule that the separating curve has *n* parameters requires at least O(*n*) observations in the future. To reuse the induced classifier in a new "fact circumstance," one first achieves a reasonable fit to the new data.

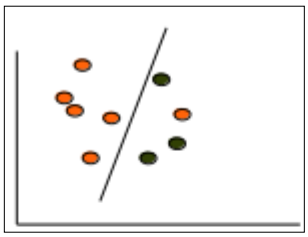

Figure 1. An imperfect linear binary-classifier with two features.

Binary (predictive) classification should not be confused with predicting real-values: the exercise of fitting a real-valued function to real-valued pairs of observations. In the latter case, $O = \{ \langle x(i), y(i) \rangle \mid i=1..n \}$, and each *x*-value acts like a feature, with the *y*-value like a prediction. The line, $t(x)$, that the student fits has output that is a real-valued *p* (Figure 2). Here, error can be measured in the reals, not just counted in the integers, and many like to sum the squares of the deviations between each theoretical $t(x(i))$ and $y(i)$, to produce the error of *t*'s fit to *O*, $Err(t,O) = \sum\{i\} square(t(x(i)) - y(i))$. Note that we have one feature variable and one forecast variable (to accommodate the variety of *p* values, the black/red and *p-not-p* labeling in the prior examples became the *y*-values).

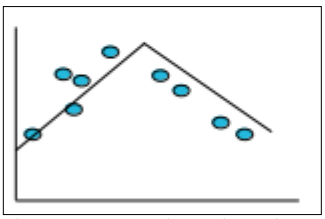

Figure 2. A piecewise linear predictor on one feature.

Why should error be the sum of squares? A good question. This error function has known, desirable mathematical properties. It really is meta-knowledge from the theory of curve fitting that is being applied to the selection of curves. But that is a complication beyond the scope of the present discussion. One should note though that as computation invades induction, traditions like sum-of-squared errors are likely to fade away.

Both curve fitting a (predictive) classifier and curve-fitting a real-valued predictor are easily generalized to higher dimensions. In fact, much of the action takes place in the higher dimensions, so it is an important exercise to consider more than one or two features. For the binary classifier, imagine red and black points in 3-space. A plane, or manifold, must be found that separates the red and black points with minimal type I and type II error (Figure 3). For the real-valued forecast, imagine very thin columns upon a plane. The *xy* plane is the variation of feature values. The *z*-value, or height above the plane, is the observed (dependent) variable. One wants to fit a plane, or manifold, to the tops of the columns, to describe the skyscrapers as a function with minimal sum-of-squares error. In higher dimensional space, the planes are hyper-planes (no longer visualizable with analogy to physical space).

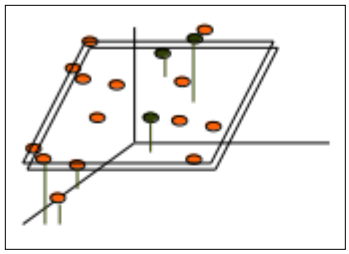

Figure 3. An imperfect planar binary-classifier with three features.

Why should the induced manifold be continuous? Can't one fit step-wise linear functions as predictors instead of a single line or continuous curve? Can't one have islands of black points and pools of red points that do not connect? These are also good questions. Non-linearities are mathematical inconveniences. But discontinuities and absence of simple-path-connectivities are truly inconvenient; usually they require a logical or quasi-logical representation: "If *x* is less than 5, then *g*; else *h*." What we now know is that theory-formation in the form of logical constraints is at least as inconvenient as discontinuous manifolds in regression and machine learning. Why would one permit theorizing by inducing such a discontinuous classifier or predictor? Because sometimes the error rates are too high without this discretization of the planes and hyperplanes. Neural nets and decision trees are perfectly capable of inducing such non-simple representations, so there must be even more complexity that philosophers' theory-formation has involved, or my complaint would end here.

KYBURG'S SCIENTIFIC THEORY INFORMATION

The simplest version of Kyburg's theory-formation takes place with two features, and a we can focus on a single "final" prediction. So imagine a set *O* of points in the plane, with some labeled red and some labeled black. Further, imagine that there is great uncertainty about some of the *y*-values, so some of the points are actually vertical line segments. In fact, if *y*(*i*) is not even observed for some *i*, then that point is not a segment, but a whole vertical line, unbounded on both sides. Kyburg assumes one is working at an implicit acceptance level, say .9, with a background set of rules at that level, *K*(Ur) (the "Ur-corpus," the meaning postulates of language, the theoretical principles at the heart of the web of belief, or the axioms of the domain). At the acceptance level .9, there are statements that are acceptable, *K*(.9), including all statements acceptable at a higher level, e.g., *K*(.95), as well as the logical constraints or rules, *K*(Ur).

Only when *K*(Ur) is non-empty does the Kyburgian account start to get interesting. The complication begins because some of the unobserved *y* values will be predicted, in support of, or collusion with, the red/black prediction.

Suppose *K*(Ur) contains a single rule "*y* is always in the range [.5*x*, 1.5*x*]." That is, instead of "*F=ma*" or "all crows are black," one has adopted the rule that *y*'s value does not deviate from *x* by more than 50 percent. *X* might be the number of strikeouts a pitcher throws in a game, and *y* the number of walks; perhaps for a specific pitcher these numbers tend to be proportional (apparently not a very good pitcher, since most good pitchers fan more than they walk!). The red and black labeling might be whether the pitcher won the game or not.

Suppose we did not record the number of walks perfectly—maybe there is even a dispute whether to include intentional walks in the *y* value (Figure 4). No problem: where *y* is not known, the rule provides a short line segment for *y*; where *y*'s line segment is broad, the *x* value, plus inference, sometimes tightens the *y* segment. It may now be possible to draw a line or slightly wiggly line on the plane that separates black and red points (i.e., separates wins and losses) very well. Perhaps a third-degree polynomial can fit so well as to separate 91 out of 100 of the points in accord with their labels. Because the predictive fit is better than .9, any new observations of *x* without an observation of *y* can make use of the induced classifier to predict a win or loss with high enough probability that the conclusion is acceptable in *K*(.9). This is theory formation.

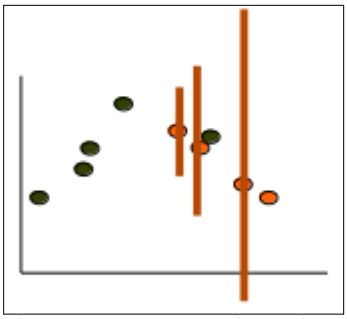

Figure 4. Kyburg's starting point – uncertain y's.

But there are three catches that have to do with the interplay of (1) power and error, (2) power and simplicity, and (3) power and coherence.

First, those tightened *y*-value inferences might actually disagree with the observed *y* values. There might be an *x*=4 and *y*=[0, 1], among the observations. The inference rule in *K*(Ur) says *y* should be in the interval.[9] Since this is a contradiction, one must conclude that if the rule is correct, then the ability to discern *x* and/or *y* well is imperfect. This counts as an error against the probability that an apparent *x* value is correct, and/or the probability that an apparent *y* value is correct. If error rates, based on past observations and past errors, rise so that the acceptability of an *x*-observation is less than .9, then none of the *x* observations is acceptable at *K*(.9)! If a rule in *K*(Ur) is mere wishful thinking, or out of date, or improperly qualified, or an inaccurate portrayal of the universe's true regularities, then it will start generating errors in the "measurements."

Of course, the rule could be weakened, from *y*=[.5*x*, 1.5*x*] to y=[.1*x*, 4*x*], or further, to *y*=[0, 5*x*]. This would always decrease the error rates of observations, but weaken the classifier, because some "points" would be too long as segments to be classified. Obviously the theory can take *x* not simply to become unacceptable at .9, but instead, to have a penumbra of error, ±10%, for example, in order to keep *O*'s <*x*(*i*), *y*(*i*)> observations .9-probably in accordance with *K*(Ur). This is how Kyburg uses probability and error to control theory's predictive power.

The second tradeoff in this inductive scheme is between power and simplicity, mediated by error and coherence. We supposed we had fit a third-degree polynomial, which had a .91 veridicality, or .09 error rate (in Kyburg's world, all probabilities are actually intervals at the presumed level of acceptance, and acceptability depends on the entire interval clearing the acceptance threshold). We could

instead have fit a simpler curve, even a line, in which case the error rate might have risen to, say, .2. At this level of error, none of the predictions would have been acceptable at the working acceptance level (a situation similar to the insisting on keeping the error-generating strikes/balls rule, above, except that here the insistence is on unworkable simplicity).

Or we could instead have fit a fourth-degree polynomial, an even more wiggly separator, violating the usual inductive desire for simplicity. But in those cases where the induced classifier needs to be applied and only a small number of points have been observed, one will be unable to find the parameters of the curve and unable to forecast wins and losses. (The most well-protected fantasies are rendered benign because there are too many theoretical parameters, and the real world fails to give practicable fit.) Remember that if the <*x*(*i*), *y*(*i*)> observations from *O* are being accepted into *K*(.9) with a plus or minus "standard error," so that all points are actually little rectangles in the plane, then it might take more than *n*+1 points to pin down the *n* parameters in a polynomial with sufficient accuracy to permit classification of all but the easiest cases. The highly-parameterized theory would have little predictive power. This is how Kyburg uses power to control simplicity.

Third, since the induced classifier is entered into *K*(Ur), generating inferences at *.9*, the classifier could be kept while rejecting the balls-strikes predictor, if errors start to climb. In most of baseball, a rule that the number of walks don't often deviate much from the number of strikeouts in each game would be hard to maintain even for the most regularly proportional pitchers. One could insist on keeping the rule, but it would have no bite because it would apply to no observations of acceptable epistemic grade, and this would lead to its eventual repeal. One would like to have multiple upstream predictors funneling inferred values into a downstream classifier. But when error rates rise, whichever rule is most incoherent, whichever rule's removal would result in the most acceptable inferences, is in fact removed. This is like training in machine learning except that Kyburg imagines it happening during normal inference as well. Thus, one has a method of revising theories, not just forming them. By preferring that the totality of *K*(.9) information be maximized, Kyburg uses coherence to control power.

It is quite easy to see the difference between Kyburg's starting place and the machine learning paradigms. Observations are partial; if there are dozens of features, then Kyburg presumes some observations report the values for just a few features, perhaps all with penumbrations based on past reliability rates, and some that will simply have to be inferred by a predictor. Theory consists of what was learned on prior data, especially about the interplay of partially observed features. It also consists of any rule induced on the current observations (the training data), applied (after some parameters are determined by data) on future observations de novo.

So Kyburg is talking about the aggregate effect of many predictors and classifiers interacting as a whole. Bayes networks and multi-layer neural networks can also consider such aggregations of multiple predictors. However, those approaches are acyclic (predictors are applied in serial fashion, with no feedback of error to control inference, although there may be a lot of feedback during training), and they do not envision the full combinatorial range of rules as candidates for rule-adoption (their pre-existing connectivity restricts the relations that can be formed or rejected). Pinkas's approach to logical rule extraction from Hopfield networks might actually be very similar to this Kyburgian power-error-simplicity-coherence exchange model of theory formation. Recursive neural networks may also have interesting tradeoffs between power, error, and simplicity, though these considerations may be metaphors derived from structure, or epiphenomena: not exactly *ding-an-sich* modeling that would make Carnap and Nagel sit up and applaud.

LOUI-NORMAN LEGAL THEORY FORMATION

In legal theory formation, the problem begins with finding a classifier for a set of cases. These are not empirical observations, but are decisions made by judicial authorities (as often a bureaucrat as a judge), possibly in different jurisdictions, all of which are supposed to be consistent. Sometimes a judicial decision can simply be overridden as "erroneous," especially in appellate situations, but that is not the normal way that exceptions can be treated in "lower" courts.

Many are unfamiliar with Anglo-American judicial decision making and the use of precedent: the main theory-formation phenomenon is the expression of the new case as a rule (the "rule of the case" or "theory of the case"). One sometimes seeks the expression of a set of cases as a set of rules. This expression is important because it is one way that consistency and fairness are "hardened" in the system. Frequently, legislative guidance is imperfect and envisions the use of cases and courts to determine the boundaries (machine learning people call this "lazy learning," but it is really deliberately vague specification). On occasion, an entire section of legal code is rewritten in order to simplify, clarify, make more practicable, make more precise, or even to change the rights and obligations wholesale. But most of the time, an incremental theory of the case, as an explication of new judicial guidance, suffices.

In scientific theory formation, one seeks a simple expression of the regularities observed in the world. In legal theory formation, one seeks a suitable expression of the norms implied by the judiciary.

Consider a graduate student who once complained, "The housing policy is unfair. When I was a freshman, they said sophomores could have first pick. When I became a sophomore, they were inconsistent; they let some freshmen pick ahead of us." The legal theory formation problem is to find an explanation, presumably faithful to the actual decision rationale, which explains why one case was decided one way and another differently. One does not get to say "higher error rate." One has at least to say that the housing rules changed year-to-year. Most likely, one cannot rely on that chronological feature and must find, or invent language to describe, a new and relevant feature to explain the distinction. "Sophomores who are

majoring in computer science are treated like freshmen for the purposes of housing selection." Or, "freshmen who are children of significant donors pick ahead of sophomores." The reason this explication is important is that it says which future sophomores will be treated like freshmen, and vice versa, and it says, at least for now, that there is nothing preventing preferred selection next year, when the student is a junior. It would be even better if it explained why computer science majors were treated this way (perhaps the rule regarding children of donors is self-evident!). "Black persons can vote, but not if they can't read and write" is a perfectly good expression of a rule, albeit a politically foul one, because it says that when presented with black persons who can read and write, they should not be barred by this rule, and should have *prima facie* ability to vote (under their status as citizens under the 15th Amendment).

We can start with one feature, with a set of cases, O = { $\langle f(i), p(i) \rangle \mid i=1..n$ }, and a labeling on each case, *p* or *not-p*. A threshold t is drawn, so that those f values above t are *p*, and those below t are *not-p*. In legal theory formation, there is nothing wrong with selecting t in such a way that many of the observed cases are inconsistent. That's because the t-threshold is a "defeasible rule." It admits exceptions, and in fact, the number of exceptions may rival the number of subsumptions-sans-exception (consider "by presumption, the signed contract is valid"; "by presumption your tax return is properly filed"; etc.). Theory-formation requires the invention (or selection) of a feature, a second dimension, in which all (or most, or many) of the exceptions can be subsumed by rule: O = { $\langle f(i), y(i), p(i) \rangle \mid i=1..n$ } where a more specific rule can be adopted such that the following disjunction holds:

for all members i of O,

if $f(i) > t$, then $p(i)$ is *p*

or

if $f(i) > t$ and $p(i)$ is *not-p*, then $y(i) > t'$.

We eventually want the theory T to consist of two defeasible rules:

if $f > t$, then defeasibly, *p*

if $f > t$ and $y > t'$, then defeasibly *not-p*.

Those who do not like defeasible rules will say that this is equivalent to two material rules:

if $f > t$ and $y \leq t'$, then *p*

if $f > t$ and $y > t'$, then *not-p*

But the defeasible rules give "a leap to *p*," when the y-value is unknown, or known only within a tolerance interval that straddles t. As students of non-monotonic reasoning in AI can recite, the defeasible theory permits inference when y is unknown or unconsidered; that inference is revisable/corrigible on further knowing, or revisable/defeasible on further considering (the inference is not "fallible" because it is not qualified by a probability grade, and it is not "fuzzy" because it is not qualified by an exemplification grade).

With a set of cases O, one can imagine the following serial algorithm: select a feature randomly, choose a threshold that accounts for much of the *p-not-p* separation. Add that as a rule, *r1*. *E1* is the maximal subset of O whose members are not correctly classified under this rule, *r1*. Choose a feature, randomly, and a threshold that accounts for many of the *not-p*'s in *E1* that are incorrectly classified under {*r1*}. Add that as a more specific (defeating) or higher priority rule, *r2*. *E2* is the maximal subset of O whose members are not correctly classified under {*r1*, *r2*}. Continue until all exceptions have been subsumed under rules, and the next E set is empty.

One can immediately see a problem here. The features should be selected in an order that makes sense with respect to the shifting burdens of conflicting rights or presumptions in the showing of legal concepts. Randomness here is disrespectful to the jurist's prudence.

There is more of a problem for those who understand how to record cases, and this is our contribution. Cases do not appear as feature sets and decisions. They have internal structure: arguments pro, con, riposte, rebuttal, distinction, overriding, undercutting, restriction, extension, meta-reasoning, and hypothetical reasoning. They contain almost-binding precedent, interesting but non-binding precedent, and *dicta*. Most importantly, they contain the essential decision between conflicting features or arguments, which one is persuasive. The dialectical and analogical parts of the case are easily represented as a tree of arguments. At root, there is an argument (which is itself a tree, but which we can flatten into a set of features and a conclusion, here), *A1pro* = < {*f1*, *f2*}, c }. This argument may have been attacked by counterarguments, *A2con* = < {*f1*, *f3*, *f4*, *f5*}, *not*-c >, and *A3con* = < {*f6*, *f7*}, *not-f2* >. And there may be a rebuttal, *A4pro* = < { *f8* }, *not-f6* > (Figure 5).

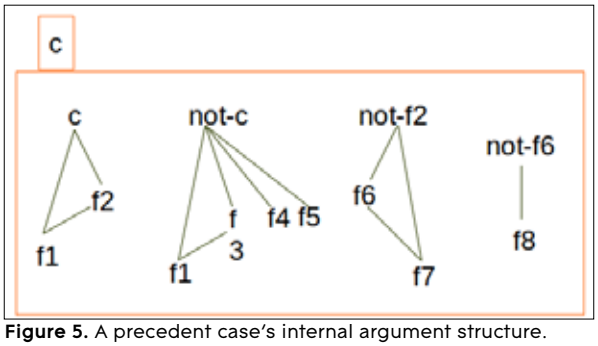

Figure 5. A precedent case's internal argument structure.

We permit some rules of this case, but not others, based on the argumentative structure:

we permit the adoption of

f1, *f2* >== c

f1, *f2*, *f3* >== c

f1, *f2*, *f3*, *f4* >== c

f1, *f2*, *f3*, *f4*, *f5* >== c

f1, *f2*, *f3*, *f4*, *f5*, *f6*, *f7*, *f8* >== c

but not

f1 >== c

because the top level argument required at least *f1* and *f2* (the proposed rule would over-generalize);

also not

f1, *f2*, *f3*, *f4*, *f5*, *f6*, *f7* >== c

because the introduction of *f6* and *f7* appears to require the rebuttal based on *f8* in order to yield a case for *c*.

There may be other rules that can be adopted or that should be barred based on the interaction of the arguments in relation to the background set of rules and cases.

In short, just because a set of dialectical rules can cover a set of cases without error does not mean that the set of rules respects the internal structure of cases, the argumentation pro and con, that led to the decision of the case.

Some might think that the fewer the defeasible rules, or the fewer number of exceptions to the general rules, the simpler the theory. But here again, the peculiarity of the phenomenon trumps mathematical convenience. The selection of features should be cogent because the public should comprehend the rules and recognize the principles on which they are based. The relation between the antecedent features and defeasible consequents should make sense in terms of public good and effective governance. There is no discovery of "hidden" nomology in law; there is only explication of norms going forward. The cogency of rules derives from the source of the rule used in the prior case: (1) judgment from other precedent cases, (2) the legislative history of the rules, or (3) the counterfactual arguments that if *r1* were to be adopted, some other clear cases would be at risk of anomaly or countermand.

RELEVANCE

Kyburg struggled to describe relevance during theory-formation. His theory of the reference class, which lays the foundation for acceptance and error, also requires that candidate features be relevant. At the very least, according to Kyburg, spurious disjunctions (unions of sets) should not be entertained as candidate theoretical concepts. What is the probability of my being elected president of the United States? Well, consider the percentage of successful elections among the reference class that comprises myself and all prior presidents. This is not as relevant as the percentage of successful national elections among persons from Hawaii, or persons from Harvard, or even better, persons with no party nomination and no experience as an elected official. Somehow relevance and irrelevance are often discernible.

The fact of the matter is that relevance comes from theoretical knowledge. By the time the machine learning or statistical regression practitioner has selected candidate features, that practitioner has applied some background knowledge of relevance (even if there are thousands of candidates such as dictionary words or gene expression markers). When the medical-outcomes statistician is handed the spreadsheet to do Principal Components Analysis, the columns have already been pared down to those that make medical sense, those that have some possible biochemical, morphological, or genetic pathway. So perhaps the anti-intellectual, computation-heavy, matrix-oriented, statistician-upsetting methods of machine learning and KDDM today are not completely oblivious to the phenomenon of interest. They just do their magic after the human has had some say, after rules adopted in human minds are willing to pass the baton to the automaton.

My early favorite example of this was when sports broadcasters would announce statistics for a weekend game: the record at home is 50-30, but after two losses, it is 10-2, and the pitcher on the mound today is 15-6 while the opposing pitcher is 10-10. But on weekends, the team is 4-20. Now, why would anyone think that weekend timing is probabilistically causally relevant to rate of victory? It's like those spurious correlates that show up with small samples when local medical offices study their patient data, right? Well, it turned out that this team partied a lot on Friday and Saturday nights, and didn't get enough sleep, so perhaps the "is-a-weekend-game" feature was perfectly relevant and belonged in the regression analysis. Only one with a very good understanding of causal pathways, who is willing to adopt rules regarding the structure of the phenomenon, can say.

CONCLUSION

The beauty of AI's ML/KDDM PCA SVM ANN/RNN ID3 C4.5 SVD and ILP "new math" is that it always works, regardless of what the operator knows about the domain, and regardless of whether the acronyms can be decoded by those who can turn the crank. These have become the Ajax, Kleenex, Clorox, Xerox, Q-Tip, Scotch Tape, Google, and AJAX (Asynchronous Javascript And Xml) brands. They are synonymous with the problem, not merely naming a specific solution, and standing in place of any thought that might be given to alternatives.

But every once in a while, people look back and discover that there was a Tesla to the world's Edison, or a Lampredi behind so many Fiats. To me, this is what has become of the philosophers of science and philosophers of law in theory-formation. So much commercial success has accompanied the "black box," behavioral, input-output methods of machine learning that it has become easy to

forget the older traditions that sought to portray structure. It is easy to ignore structural work on scientific and legal theory-formation even with much recent progress: many philosophers of law still do not know of our argument-based models of the case in AI and law; many philosophers of science do not know Kyburg's final framework on measurement errors and the web of belief. Applied success is not always anti-intellectual, because frequently the former obscures the latter with no special antipathy. But the loss of theoretical understanding, deliberate or not, targeted or not, is something we must resist.

NOTES

1. See Ernest Nagel, *The Structure of Science: Problems in the Logic of Scientific Explanation*.
2. See Frederick F. Schauer, *Playing By the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*.
3. Herbert A. Simon, "Does Scientific Discovery Have a Logic?"; Douglas B. Lenat, "On Automated Scientific Theory Formation: A Case Study Using the AM Program"; and John H. Gennari, Pat Langley, and Doug Fisher, "Models of Incremental Concept Formation."
4. Gerald DeJong and Raymond Mooney, "Explanation-Based Learning: An Alternative View."
5. Gadi Pinkas, "Propositional Non-monotonic Reasoning and Inconsistency in Symmetric Neural Networks"; Pinkas, "Symmetric Neural Networks and Propositional Logic Satisfiability"; Pinkas, "Reasoning, Nonmonotonicity, and Learning in Connectionist Networks That Capture Propositional Knowledge."
6. Henry E. Kyburg, Jr., *Science and Reason*.
7. Ronald P. Loui, "Rationales and Argument Moves"; Loui, "A Modest Proposal for Annotating the Dialectical State of a Dispute."
8. Douglas B. Lenat, "CYC: A Large-Scale Investment in Knowledge Infrastructure."
9. Gennari, Langley, and Fisher, "Models of Incremental Concept Formation"; Lenat, "CYC: A Large-Scale Investment in Knowledge Infrastructure."

BIBLIOGRAPHY

DeJong, Gerald, and Raymond Mooney. "Explanation-Based Learning: An Alternative View." *Machine Learning* 1, no. 2 (1986), 145–76.

Gennari, John H., Pat Langley, and Doug Fisher. "Models of Incremental Concept Formation." *Artificial Intelligence* 40, no. 1 (1989): 11–61.

Kyburg, Jr., Henry E. *Theory and Measurement*. Cambridge University Press, 1984.

Kyburg Jr., Henry E. *Science and Reason*. Oxford University Press, 1990.

Lenat, Douglas B. "On Automated Scientific Theory Formation: A Case Study Using the AM Program." *Machine Intelligence* 9 (1979): 251–86.

Lenat, Douglas B. "CYC: A Large-Scale Investment in Knowledge Infrastructure." *Communications of the ACM* 38, no. 11 (1995): 33–38.

Loui, Ronald P., and Jeff Norman. "Rationales and Argument Moves." *Artificial Intelligence and Law* 3, no. 3 (1995): 159–89.

Loui, Ronald P. "Comment on the Cardozo Conference on Graphic and Visual Representations of Evidence and Inference in Legal Settings." *Law, Probability and Risk* 6 (2007): 319–26.

Loui, Ronald P. "A Modest Proposal for Annotating the Dialectical State of a Dispute." *Script-ed* 5, no. 1 (2008): 176–97.

Nagel, Ernest. *The Structure of Science: Problems in the Logic of Scientific Explanation*. New York: Harcourt, Brace & World, 1961.

Pinkas, Gadi. "Propositional Non-monotonic Reasoning and Inconsistency in Symmetric Neural Networks." In *Proceedings of the 12th International Joint Conference on Artificial Intelligence-Volume 1*, 525–30. Morgan Kaufmann Publishers Inc., 1991.

Pinkas, Gadi. "Symmetric Neural Networks and Propositional Logic Satisfiability." *Neural Computation* 3, no. 2 (1991): 282–91.

Pinkas, Gadi. "Reasoning, Nonmonotonicity, and Learning in Connectionist Networks That Capture Propositional Knowledge." *Artificial Intelligence* 77, no. 2 (1995): 203–47.

Schauer, Frederick F. *Playing By the Rules: A Philosophical Examination of Rule-Based Decision-Making in Law and in Life*. Oxford University Press, 1991.

Simon, Herbert A. "Does Scientific Discovery Have a Logic?" *Philosophy of Science* 40, no. 4 (1973): 471–80.

Statement on Massive Open Online Courses (MOOCs)

Felmon Davis
UNION COLLEGE

D. E. Wittkower
OLD DOMINION UNIVERSITY

The following statement was prepared by Felmon Davis and D. E. Wittkower in consultation with the American Philosophical Association's committee on philosophy and computers.

Since 2012, Massive Open Online Courses, or MOOCs, have generated much discussion as a innovative response to several pressures bearing in on traditional "brick and mortar" pedagogy, including the promise of reaching a wider public, revolutionizing the means of pedagogy, offering more "value" at lower costs, and providing more current information and access to research than traditional education. MOOCs are typically open to the public, can in principle reach thousands of individuals all over the world, and may employ various technologies that encourage participation such as blogging or online chats. There is usually no cost of enrollment except for a fee for students interested in gaining a certificate, if such is offered. Some philosophers have offered MOOCs, among them prominent figures such as Michael Sandel, Walter Sinnott-Armstrong, Tom Beauchamp, and Peter Singer. There is no way around the question whether this particular form of "delivery" of "content" is an apt medium for the essential distinctive features of philosophical activity: If the medium is the message, what message does a MOOC in philosophy convey?

A brief report cannot do justice to the complexities of this issue; instead, we want to set markers for some of the important places where MOOCs offer promise to philosophers and where they set pitfalls. Our hope is to initiate a discussion of "best practices" for philosophical pedagogy using MOOCs.

This effort only has a point if the phenomenon of MOOCs is not ephemeral. The MOOC phenomenon has been touted as "The Most Important Education Technology in 200 Years" [MIT Technology Review] but now we read fatalistic voices decrying MOOCs as "a futile experiment."[1] One has to place one's bets here, and our feeling is that the phenomenon follows the *Gartner Hype Cycle* (Figure 1), where a phenomenon is hyped too much, followed first by waning interest and then by slow and steady subsequent growth.[2] If this is so, it is worth studying the phenomenon

now—perhaps particularly now that skepticism seems to reign—because the present offers a good opportunity to take a stronger hand in shaping the course of the future.

Figure 1. Gartner Hype Cycle.

Let's consider some of the virtues of MOOCs and some qualifications.

WIDE REACH OF MOOCS

While, like traditional broadcast media, MOOCs can in principle reach tens or hundreds of thousands of individuals, the Internet can seep into corners where access to traditional broadcast media is limited. Furthermore, MOOCs can more easily solicit participation than televised or streamed video as participants can blog, post videos, engage in online discussion sessions, stream their own live video sessions, and so on.

On the other hand, access is not necessarily cheap even if the courses are themselves free. Access to material resources such as the necessary broadband, reliable computers, etc., is a problem especially for streaming video, but also audio, which require webcams and microphones at a minimum. A somewhat up-to-date computer is necessary even to access content online. One needs to recall that 75 percent of the world's population has no internet access or lacks the "connectivity" necessary for satisfactory access to cloud-based apps; 81 percent of Internet access in China is via mobile devices, which are a tight fit for online course work.[3]

Moreover, the broad diversity of participants militates against quality and consistency of participation as teenagers mix with older students, as the highly educated mix with the unschooled, and as cultural expectations diverge—for the very diversity which offers unique opportunities of communication and learning can lead to deep problems there as well, particularly in courses which use peer-grading.

The distance format and the large number of participants can make it difficult to take full advantage of the opportunities provided by the mix of cultural attitudes and to integrate them into coursework. The idea of critical engagement with others requires a sense of privilege or an easy egalitarianism not everyone is comfortable with, and even societies that tout their openness are often surprisingly eager to restrict cross-cultural debate; for example, recently the U.S. government has compelled Coursera to ban students from Cuba, Iran, Sudan, and Syria.[4] Cultural diversity imports not only problems of linguistic understanding but also problems of tone, temper, and status, exacerbating the notorious challenges of maintaining civility, tolerance, and nuance in online discussion.

EVALUATION OF WORK; CREDENTIALS

Most courses do not seem to offer credentials or offer credentials, which seem little more valuable than the paper (or PDF) the student receives. And the *London Times* has reported that, when given the option to get course credit for their MOOC (for a fee), none of the thousand or so students who enrolled in a British online class did."[5]

And the drop-out rate from MOOCs is enormous. A study from the University of Pennsylvania found that only 4 percent of registered users finished their courses, and half of the enrolled did not view even a single lecture:

> Emerging data from a University of Pennsylvania Graduate School of Education (Penn GSE) study show that massive open online courses (MOOCs) have relatively few active users, that user "engagement" falls off dramatically—especially after the first 1–2 weeks of a course—and that few users persist to the course end.[6]

It is not clear whether what we are looking at in these cases is a "bug or a feature"; as a recent contributor to *Slashdot* put it:

> In "The Online Education Revolution Drifts Off Course," NPR's Eric Westervelt reports that 2013 might be dubbed the year that online education fell back to earth. Westervelt joins others in citing the higher failure rate of online students as evidence that MOOCs aren't all they're cracked up to be. But viewed another way, the ability to try and fail without dire debt or academic consequences that's afforded by MOOCs could be viewed as a feature and not a bug.[7]

According to a recent study by *Coursera*, many people treat these courses as a resource which they "browse" for interesting lectures or learning something of interest, not to earn credentials or certification.

It seems a clear *prima facie* good to offer to a wide interested public such easy access to university-level courses, at a reasonable price (cost of equipment and bandwidth) and largely independent of time and place, even if the public does not treat it as an opportunity to gain a degree or university discipline. This *prima facie* good may be particularly advantageous to philosophy, a field that interests many but to which most people have limited access outside of university settings.

WOULD PLATO OFFER A MOOC?

Professional philosophers have an interest in attracting a wider public to their work, but there are questions about the suitability of the medium. These questions are particularly delicate for philosophers, who do not always agree about the ends and methods of the discipline, and for whom, indeed, the proper ends and means are part of the subject matter. If you conceive of philosophy as requiring thoughtful dialogue leading towards reflective equilibrium about an issue, significant problems for philosophical instruction and practice are posed by the very massiveness of MOOCs, their lack of intimacy, the discontinuity of

discussion, and absence of mutual acquaintance in both the student-teacher and in the peer relationships in the course. As Alison Byerly points out, professors take it for granted that they should respond to emails from students, provide advice on further study, write recommendations, and perhaps meet with students but in some cases "responding to even one percent of those taking a MOOC could mean interacting with 1,000 students."[8] Moreover, individuals often do not produce work themselves—they may just "attend" lectures—or do not receive pointed feedback and evaluation of their work, virtually assuring their engagement will be "casual."

We may think even of Plato's hostility towards writing in the *Phaedrus*, where Socrates denigrates the value of writing, which he compares to a mere image of speech—unable to explain or defend itself, to adapt and respond to its audience, or to know when to speak or be silent. While we may shake our heads at what seems to many of us today to be a misguided and technologically deterministic dismissal of the value of writing, many of our current pedagogical concerns with MOOCs are not much different, both in content and structure. Surely Plato was right to some extent to worry that philosophical development cannot take place by engaging with dead words on a page which cannot answer back to our questions and critiques, and yet this admission in no way commits us (or Plato, apparently, since he wrote this very dialogue) to the view that written work cannot play a vital role in philosophical development.

Similarly, a conception of philosophy as requiring intimacy of dialogue and interaction does not argue against the utility of MOOCs; instead, it simply points up their limitations and due recognition of their limitations may open up their true promise. A course in which students do nothing but read texts in lonely isolation and take periodic quizzes looks unattractive as pedagogy but a canned MOOC is not much different from that. But just as videos and guest lectures can play a vital role in learning, so can MOOCs when combined with trenchant discussion, serious writing that receives individual assessment, and thoughtful counselling from a teacher.

Even without the living word MOOCs can still have their usefulness; one notes that most MOOCs operate at an undergraduate level—thus, they are not geared to generating and organizing original research (except, perhaps, for the teachers!). They instead introduce amateurs (often in the original sense of the word) to basic concepts and techniques and hopefully entice them to look further. MOOCs can open the world of philosophy to people whose busy lives occupy them elsewhere but who want to participate in the life of the mind.

PITFALLS FOR THE PROFESSION

Aside from these pedagogical concerns, there are also reasons to be concerned about MOOCs and the future of the profession of philosophy—much more significant concerns. In a recent piece in the *Chronicle of Higher Education*, Peter Schmidt reports that the AAUP is wary of the copyright implications for course materials teachers develop:

> With the emergence of MOOCs, however, colleges have begun asserting ownership of the courses their faculty members develop, raising the question of what is keeping such institutions from claiming ownership of other scholarly products covered by copyright, such as books.[9]

It is important that scholars and teachers, and the APA, keep eyes peeled for threats to intellectual property.

There are also significant concerns about the effects on labor and employment in the profession. Sometimes teaching staff is enlisted for offering MOOCs at much lower salaries than regular faculty, especially at less prominent universities. In addition, enrollment caps on online courses are sometimes set higher than in traditional classrooms—with MOOCs, we should expect this trend to either hold or expand. For these reasons, we worry that MOOCs could provide an avenue by which administrators may either cut lines in favor of increased use of contingent faculty, or simply reduce adjunct employment by increasing the number of credit hours served per instructor per class.

This bottom-line thinking might enhance the employment opportunities available within the profession in some ways but jeopardize them in others. The courses most well-suited to MOOCs are those introductory and general education courses which are the primary source of adjunct teaching and employment opportunities for recent graduates and others seeking to land a full-time position. If classroom capacities are higher in MOOCs than in traditional classes, there may be fewer courses available to those seeking tenure-track employment opportunities. Arguably, MOOCs can lead to decreased tuitions and expand employment opportunities for contingent faculty; the danger is that these opportunities might provide these teachers less value if they are deprived of credible teaching evaluations and the rich teaching experience that support applications for regular employment.

And the heavier the emphasis on general education as a potential revenue stream for finance-strapped universities and colleges, the more we should be concerned for the fate of the liberal arts ideal of engaged, Socratic, student-centered learning in a university culture increasingly focused on vocational training and cost-saving measures.

We must also beware of false economies. The conviction that MOOCs will enhance the bottom line for universities with inadequate budgets may be a fantasy. MOOCs "require investing expensive technological and labor resources to create experiments of questionable educational value to be given away," as Jason Mittell writes.[10] Big-name universities, showcasing academic stars—which may incidentally condition the public to favor "intellectual celebrities" over other worthy teachers and courses—are in a far better position to take these risks than public universities with limited budgets.

The idea that MOOCs offer relief for the problems of underfunding of higher education may have broader and unwelcome consequences for the "educational divide" between, on the one hand, prestigious colleges and

universities that can offer their students vivid face-to-face engagement with real teachers and, on the other hand, lesser-funded public schools whose students may become consumers of packaged courses or at best interact with a teacher who is no more than a "glorified teaching assistant."[11] But is it better to have an excellent teacher and researcher such as Michael Sandel in a video presentation or living interaction with faculty of the local university or college? This is a false dilemma so long as we retain the ability to design courses that combine the virtues of both approaches with respect for both modes of teaching.

TREADING WITH CARE

It seems to us that MOOCs offer some promise of opening the gates of philosophy to many people near and abroad who could not otherwise approach it, but many MOOCs are now constituted in a way that limits their pedagogical value to undergraduate coursework or just casual browsing usually without much promise of academic credit. Unless integrated as one component among others of live education with professors who are actively engaged in research and teaching, the medium still seems ill-suited to the practice of philosophy as reflective collaboration and argument. And the broad reach of MOOCs carries its own dangers of intercultural misunderstanding.

Professional philosophers and the APA should work closely with administrators to address concerns of justice in both intellectual property and the remuneration for labor, which should also include consideration of how MOOCs affect the career path of members of the profession, and how MOOCs may put existing faculty lines and departments at risk. The APA should be particularly concerned about the long-term future of the discipline if academic positions are curtailed and promising scholars are barred from pathways to solid entry-level positions. And as citizens we should all resist tendencies that can degrade the quality of education for the broad public.

ACKNOWLEDGEMENTS

We thank the committee and especially Colin Allen, Fritz Allhoff, and John Sullins for valuable criticism. Thanks also to Audrey Hunt (Union College) for research assistance.

NOTES

1. Jonathan Rees, "Anti-MOOC Really Is the New Black," *More or Less Bunk* (blog), August 20, 2013, http://moreorlessbunk.wordpress.com/2013/08/14/anti-mooc-really-is-the-new-black/.
2. Rick Anderson, "MOOCs and the Cycle of Hype," *The Scholarly Kitchen*, October 24, 2013, http://scholarlykitchen.sspnet.org/2013/10/24/moocs-and-the-cycle-of-hype/.
3. Kaylene Hong, "China's Internet Population Hit 618 Million at the End of 2013," *The Next Web*, January 16, 2014, http://thenextweb.com/asia/2014/01/16/chinas-internet-population-numbered-618m-end-2013-81-connecting-via-mobile/#!zsn3K.
4. Joey Ayoub, "U.S. Bans Students from 'Blacklisted' Countries from Getting a Free Education," *Hummus for Thought* (blog), January 29, 2014, http://hummusforthought.com/2014/01/29/us-bans-students-from-blacklisted-countries-from-getting-a-free-education/.
5. Chris Parr, "'Spooky Mooc' Students Fail to Bite Over Credits," February 20, 2014, *Times Higher Education*, http://www.timeshighereducation.co.uk/news/spooky-mooc-students-shun-edge-hill-academic-credit/2011445.article.
6. Kat Stein, "Penn GSE Study Shows MOOCs Have Relatively Few Active Users, With Only a Few Persisting to Course End," December 5, 2013, *University of Pennsylvania Graduate School of Education Press Room*, http://www.gse.upenn.edu/pressroom/press-releases/2013/12/penn-gse-study-shows-moocs-have-relatively-few-active-users-only-few-persisti.
7. Soulskill, "Are High MOOC Failure Rates a Bug Or a Feature?" (undated), *Slashdot*, http://news-beta.slashdot.org/story/14/01/01/2155201/are-high-mooc-failure-rates-a-bug-or-a-feature.
8. Alison Byerly, cited in Jason Mittell, "The Real Digital Change Agent," *The Chronicle of Higher Education*, March 4, 2013, https://chronicle.com/article/The-Real-Digital-Change-Agent/137589/.
9. Peter Schmidt, "AAUP Sees MOOCs as Spawning New Threats to Professors' Intellectual Property," *The Chronicle of Higher Education*, June 12, 2013, http://chronicle.com/article/AAUP-Sees-MOOCs-as-Spawning/139743/.
10. Jason Mittel, "The Real Digital Change Agent," *The Chronicle of Higher Education*, March 4, 2013, https://chronicle.com/article/The-Real-Digital-Change-Agent/137589/.
11. San Jose State University Department of Philosophy, "An Open Letter to Professor Michael Sandel from the Philosophy Department at San Jose State U," *The Chronicle of Higher Education*, April 29, 2013, http://chronicle.com/article/The-Document-an-Open-Letter/138937/.