Ethics Inf Technol
DOI 10.1007/s10676-017-9427-3

ORIGINAL PAPER

Responsible research for the construction of maximally humanlike automata: the paradox of unattainable informed consent

Lantz Fleming Miller[1]

Abstract Since the Nuremberg Code and the first Declaration of Helsinki, globally there has been increasing adoption and adherence to procedures for ensuring that human subjects in research are as well informed as possible of the study's reasons and risks and voluntarily consent to serving as subject. To do otherwise is essentially viewed as violation of the human research subject's legal and moral rights. However, with the recent philosophical concerns about responsible robotics, the limits and ambiguities of research-subjects ethical codes become apparent on the matter of constructing automata that maximally resemble human beings (as defined hereunder). In this case, the automata themselves, as products of research and development, are in the very process of their construction subjects of research and development. However, such research faces a paradox: The subjects cannot give their informed consent to this research for their own development, although their consent would be needed for the research. According to ethical codes, this research would be unethical. The article then explores whether the background concepts giving rise to this paradox could be reframed in order to allow such research to proceed ethically.

Keywords Automata · Ethical research · Helsinki Declaration · Informed consent for research subjects · Maximally humanlike automata · Nuremberg Code · Responsible robotics

✉ Lantz Fleming Miller
flemingmiller@yahoo.com

[1] Department of Philosophy, University of Twente, Enschede, The Netherlands

Introduction

In Mary Shelly's day, there were few standard ethical codes for research on humans. One can safely say there were no international professional codes concerning such work. Thus, the researcher in Shelley's most well-known novel did not consult with current global ethical codes of research before constructing his automaton built based on a death-bound criminal. If the novel had been written, set, and published in 2016, perhaps taking place in a modern research institute, the scientist's research would likely raise red flags for many among the scientific and philosophical readership, to the loss of the novel's credibility (or its increase in drama from the lead character's violating ethics). In many countries across the globe there are now ethical research committees to oversee project proposals and help ensure these adhere to ethical research guidelines.

However, despite the sordidness of Shelley's novel, it took more than a century and the gruesomeness of World War Two with its Nazi doctors' experimentation on human subjects to draw international attention to the need for worldwide professional codes for research on human subjects (Faden and Beauchamp 1986). The 1945 Nuremburg code's ten points were addressed to the practice of human experimentation (reflecting what seemed to go beyond mere medical clinical research in the Nazi cases) and were followed soon by the 1948 Declaration of Geneva. However, reflecting the increasing international community of medical research and its need for more precise ethical code, the 1964 Helsinki Declaration has proven to be the medical research ethical code so-far garnering the widest scope, respect, and support. It has undergone seven revisions since 1964, most of these concerning greater detail of research and subject conditions. While the declaration is nonbinding and serves rather as a benchmark for professional codes in

Published online: 08 July 2017

nations and organizations around the world, the most salient point is informed consent for human subjects (Council for International Organization of Medical Sciences (CIOMS) and World Health Organization 2002). The concept has so well permeated professional associations worldwide and become so embedded in clinical research, it has developed a de facto binding force (even if some researchers attempt to circumvent it).

Respected as these codes may widely be and adamantly as individuals and professional associations may strive to keep the codes up-to-date with scientific developments, they face a completely new challenge from research that is encroaching on medical-sciences' grounds: automata—especially humanlike automata—research and development.[1] The more humanlike, in both physical (Zhang et al. 2015) and mental/emotional aspects (Zhang et al. 2015), that these entities are designed and manufactured, the more one is called upon to consider the extent to which these would indeed warrant the respect and implementation of ethical research-guidelines. In contrast with research on Great Apes, concerning which many countries are slowly coming to recognize as persons (Cavalieri 2015), this new challenge for research ethics zeroes in on a very narrow population of potential persons (if one concedes that species other than humans may be persons). As commonly happens in bioethics, one must carefully examine just what being a human entails at some general level.

[1] I have two reasons for using the term "automaton" instead of "robot," and I hope the general philosophical and engineering community will seriously consider these reasons for their own terminological usage. The first reason is a matter of association: The term "robots" throughout popular culture carries a connotation of a mechanical, awkward, slavish entity. I find "automaton" more precise but also more encompassing: As to more encompassing, there have been automata in cultural mythologies for millennia, such as the *golem* in Jewish culture. As to more precise, an "automaton" is basically an autonomous entity (though not normally including humans). There is less weighty connotative, cultural baggage in the term, compared with "robot". This preference segues into the second reason, which is twofold. (1) "robot" of course comes from Karel Capek's play, R.U.D., for "Rossum's Universal Robots," in which play, as is widely known, the robot is a sex slaves. I believe that—fine though the play may be—that the term carries heavy cultural connotation, which has influenced people's thinking about the entities—whether or not favorable (the more favorable the person is toward automata, perhaps the more questionable the motives for perpetuating the term "robot"). This quasi-moral preference is not the result of the genetic fallacy: I believe the connotation is deeply ingrained in the term and in society's attitude about such entities. The idea of making any entity your slave—even a horse or automobile—is morally questionable. I believe it reflects a serious problem of values that has deeply troubled many human societies since the introduction of agriculture. I do not believe that, merely because an industry and an academic community concerned about this industry use this term, the term is legitimized for purposes of philosophical examination. (2) In trying to examine and assess philosophical issues related to automaton manufacture and use, I think it is fairer to use terms with less cultural and connotative baggage to stick as closely as possible to straightforward argument.

However, considering that it may effectively not be possible, in this article, to provide a complete, necessary, and unanimously conceded delineation of the human being in toto, one may more modestly require only that we delineate those qualifying traits relevant to the type of ethical situation at hand. In this case, automata of a specific sort may ontologically subsume a sufficient amount of qualities shared with humans that they would warrant the same ethical research-guidelines as humans require. (Thus, we need not go into all the other qualities these two groups—humans and automata—may share for all possible ethical situations.)[2]

The article proceeds by first zeroing in on the particular type of automata that will be of concern for the investigation and defining it: a type that, for now, is only theoretical, the so-called "maximally humanlike automata," or MHA. This type is essential to the consideration of the extent to which ethics required for humans should be extended to automata. An answer to this situation can then help indicate the degree to which such ethics should be extended to other, less overtly humanlike, automata. The article next focuses on the ethical issue of concern here, that of informed consent for human research-subjects and how it may apply to MHA. Within this discussion the paradox alluded to arises due to the particular nature of MHA as a type of entity: It seems they should have informed consent to undergo the research that causes their existence, yet they cannot grant that consent until they have been made. The understandable concern that such a paradox would apply to the prospect of bringing humans (infants) into existence is discussed and indicated as not evoking such a paradox. Two alternative "gradualist" approaches for handling the ethical paradox in automaton research and development in practice are described and assessed, with unclear results as to their truly solving the ethical problem. The conclusion discusses how this article's ethical concern for MHA R&D can—consistently with the article's aim to use the MHA situation as an heuristic—point to ways to handle the commensurate problem in non-MHA R&D.

Humanlike and maximally humanlike automata

Defining the maximally humanlike

Before looking further into the extent to which automata may warrant ethical consideration commensurate with

[2] Some of the arguments in this article on automata research may pertain as well to research on certain human-derived materials, including gamete gene therapy. This fact does not affect the discussion herein concerning automata, but does warrant separate work concerning consent and research on such human materials.

that for *Homo sapiens*, it is necessary to delimit just which kinds of automata are of concern for this inquiry. Certainly, automata such as those used in manufacturing facilities, operating in space-explorer missions, or used for cleaning household floors, prima facie seem too far from humans in kind to warrant that any invasive research on them demands full human-subjects research informed consent. These automata are not designed to have sentience or share such human interests as making life goals or even stay alive, which qualities are commonly considered to form some basis for why humans' have their particular moral status. (Feinberg 1980; Singer 1993; Cavalieri 2015).

At the other extreme are maximally humanlike automata (MHA).[3] Such machines have not yet been constructed. However, as the prospect of constructing such entities has been widely discussed and pursued (Zhang et al. 2015; Miller 2015) and nothing has yet proven that MHA are impossible, they warrant discussion for their potential moral status. These automata, which I will delineate more thoroughly, contrast with humanlike automata, which class includes MHA but also much more limited automata currently available such as Aiko Chihira, which greets shoppers at a department store (Hu 2015) or Nadine (Gaudin 2016). This humanoid is designed to appear, at least within a predetermined viewing range, a human being: with humanlike hair and (plastic) skin, smiling, talking, and making gestures. However, the automaton still is not designed to have sentience or share human interests such as lifetime goals; its moral status still would seem to merit that currently ascribed to *Homo sapiens*.

By contrast, MHA as I define these below, are designed to have all the qualities that, at least arguably, would merit their having moral status commensurate with humans. For the purposes of this paper, MHA will be those who can pass a fairly rigorous version of the Turing Test which can be called the Three-dimensional Turing Test (TTT, or T^3 or T-cubed). The standard Turing Test, suggested by the mathematician in 1950 (Turing 1950), involves two subjects, or operators—a computer and a human—in a different room from the human assessor, who works at a console, communicating with the other two. If the assessor can discern which responses are coming from the computer and which from the human, the computer passes and is deemed to have an intelligence level commensurate with that of humans.[4] This original test has long come under fire, with many new versions that are suggested to solve the former's shortcomings. (Bion 1979; Feigenbaum 2003; Harnad 2004) Yet, strictly for this article's purposes, mere level of intelligence would not suffice to assess the automaton as humanlike.

The so-called TTT

The concept of "TTT" serves here as a definitional requirement—arbitrarily given, but with reasoned criteria so that the pertinent issue at hand (the ethical issue of informed consent) can be placed within automata research and get some traction to motivate the discussion. The reason for why the automaton should, in appearance, behavior, and conversation be indistinguishable from a human is to render the automaton maximally humanlike. The reasoning behind this criterion is that, as far as we bona fide human beings can tell, the automaton is so humanlike we would be greatly challenged to say why it is indeed not human. If we cannot readily say why it is not human, then, ceteris paribus, there is no immediate reason why it should not have identical moral status with humans and merit the same ethical treatment, specifically vis-à-vis informed consent.

Other Turing tests more involved or complex than the one Turing first suggested have been offered, including Harnad's (1991) Total Turing Test, which involves automata behavior as well as language as necessary for observers to distinguish whether a automaton possess, or at least exhibits, human intelligence. Hauser (1993) criticized Harnad for going too far in restricting what could count for machine humanlike intelligence, yet Schweitzer (1998), in his Truly Total Turing Test, went even further suggesting we should also have to know the entity's complete history ("evolution") before we could truly determine such intelligence. However, for the purposes of this article, the T-cubed as defined here is looking for more than mere intelligence but deep humanlikeness in every possible aspect, from language use to behavior, from uses of tongue to toes, articulateness and other "naturalness," so that not mere intelligence but completeness as human is exhibited as indistinguishable from a bona fide *Home sapiens*.

While this condition may be too strict—and among all the Turing-Tests debates, is certainly wracked by potential theoretical controversies—I offer it as a criterion for the specifically ethical issues at hand vis-à-vis human moral status.

Aim and drawbacks of the T-cubed

The type of automata of concern here is that which resembles physically; mentally, culturally, socially and

[3] Because the singular "automaton" and plural "automata" have the same initial "A," in this article I simple use the same "MHA" for the singular and plural, letting context differentiate them, instead of using the clunky "MHAs" for the plural.

[4] There has been at least one report of an artificial intelligence system coming very close to passing a Turing Test of the original sort, with the assessors reaching 30% accuracy. This AI system, Eugene Goostman, was designed to mimic a 13-year-old male of contemporary industrial society and culture. (Mann 2014).

emotionally a human being as much as possible. Thus, the test required for this entity to pass as humanlike would require at least two subjects, one human and one automation; both interacting with each other and perhaps other persons. If, after some preset period, say 2 h, the assessors cannot tell which is the human or the automaton, the automaton passes and is deemed maximally humanlike. Such a three-dimensional (actually four) test is needed[5] because the goal of the presumed humanoid manufacturer is to construct an automaton that cannot be mistaken from a human.

There may be objections to even the T-cubed because it is only a behavioral test. Many people do not subscribe to behaviorism and so may maintain that we still could not know if this MHA is indeed experiencing the world, perceiving it, sensing it. That is, it may exhibit thoroughly intelligent human behavior and yet not be conscious or sentient. One response may be that one cannot exhibit humanlike intelligence unless one is conscious and sentient. A further objection would be that because the MHA is mechanical,[6] it cannot share the full range of human interests, such as the need for organic food and the agricultural infrastructure that goes with it, so it would have interests of basically different sort. MHA would have such different social and political needs from humans as not to mesh with sufficient fitness into human society; and as human society is so centrally social, these MHA could not cogently be deemed human, no matter how much they look and, out of larger context, often act like humans. In response, I cannot here delve further into this issue, as it is only tangential to this article's concern with informed consent: Even if these MHA do not share the same range of interests with humans, it is at least plausible they have a commensurate level of moral status as humans, even with different bases. This presumed plausibility is enough to fuel the article's concerns.

[5] More precisely, the test is four-dimensional, because the automaton must move and speak over some period of time. I retain the emphasis on three dimensions to highlight the physical presence of the automaton as actor, in contrast to a computer console.

[6] For simplifying the argument in this article, I am here considering only mechanically based MHA, not entities that are constructed out of biological parts. Thus, to allude to movies, I am considering only automata like the *Alien* character Ash, who is revealed at one point to be mechanical; By contrast, Roy and the replicants of *Blade Runner* appear to be biologically based (possibly cyborgs, but the movie does not make this distinction clear). Some may object that biological is mechanical, merely of a different sort. While I would agree, the biological, while a subset of the mechanical, is of a significantly, differently based construction as to merit different consideration for establishing moral statuses on different bases. In any case, this distinction between mechanical and biological automata makes no difference to this article's subject of informed consent.

Distinguishing MHA from related humanoids in terms of place in society

Now that I've distinguished the basic characteristics of MHAs from other humanoid automata, it could be helpful to orient the discussion further by considering what place these entities would have in human society. Such inquiry, though, at once is stymied by the fact that, if they are indeed indistinguishable from humans and as far as we can tell should thence receive the same ethical considerations as bona fide *Homo sapiens*, then the issue of just what tasks they should be accomplishing would become not merely moot but perhaps inimical to the ethical project. That is, whereas humans are not formed to be certain specialties, humanoid automata may be specialized as: sex or companionship suppliers (Hauskeller 2014), infant caregivers (Sharkey and Sharkey 2010) elderly caregivers (Sharkey and Sharkey 2012), teachers (Sharkey 2015), medical or technical-medical practitioners (Santoni de Sio and van Wynsberghe 2016), or military killers (Arkin 2013; O'Çonnell 2014; Sharkey 2016). Therefore, if MHA are indeed to be maximally humanlike, and humanlikeness entails being neither a slave nor having their social-behavioral traits predetermined (serving as an automaton sex professional or killer warrior), then an MHA should indeed be fully autonomous in terms of what kind of life it should live. As a corollary, an automaton built to serve as a specific person's marriage mate or as a "caregiver" or a killing machine would not be fully human.

One objection here would be that manufacturers would have no motivation to build a machine that would be so very humanlike, to the point that a consumer would have no motivation in buying one if the consumer could not have a preferred use for the product. This objection overlooks the strong possibility that, given engineering challenges and common operational motivations in themselves, engineers and R&D departments may be drawn to construct an MHA that is indeed as humanlike as described—to the point of being non-saleable—merely because:

1. Solving an engineering problem for its own sake can be an operational motivation—perhaps not the happiest situation for ethicists, but such problem-solving for its own sake often been a powerful "force" in technological development.
2. Solving the problem well could bring a great amount of prestige.
3. There may be some apology for the spinoff effect—that research in a certain direction may benefit humans through spinoff products or for humans' better understanding of themselves and their species.
4. Comprehensive doctrines of engineers and inventors who look to a future which is heavily imbued with such

machinery, even to the point that MHAs dominate and the human species eventually fade away, may serve as motivation for constructing such automata.

The question here is not whether any of these reasons to manufacture MHA are morally ad politically commendable, but only that there are potentially strong motivations for inventing MHA, even if they are not directly put to any immediate "use" by consumers.

Another important distinction for MHA compared with other humanoid types comes up in Bryson's work (2000, 2009, 2010) on the place of automata in society. Bryson maintains that automata should be "our" slaves—slave to their master humans. Her outlook has merit in spirit, but her terminology unfortunately overshadows the sentiment. The problem is the term "slave." If slavery is, as most of the world now concurs, not morally good, it is reasonable to deduce that not only should no one be anyone's slave, but also no one should be anyone's master. There is something about the relationship that is wrong. In current human morality, it is wrong for one organism to exert total control over another. Acting in such a way that that organism is wholly in one's control is good neither for the master nor for the slave. Even though farm animals are harnessed for the sake of humans, they can, via proper (traditional) husbandry (see Rollin 2011), be treated in such a way the human is not mere master over a slave animal.

The same reasoning can be turned to human as master over slave-machine. There is some harm to one's own higher moral values and moral character if one establishes oneself as master. One may propose that one value that leads one to be master is excessive ease and comfort. Even Locke, notoriously a plantation owner, at the least observed (if hypocritically) that there is a rightful amount of goods one may attain to maintain subsistence, and beyond that is taking from others. Surely, some persons, such as generals or corporate leaders, who cannot manage all their personal affairs alone, need personal assistance. Yet, such tasks can be handled by paid—sufficiently paid—voluntary servants. One need not have absolute control over such a life to get the tasks done, and ethically so. The problem of using and treating machines as slaves is that one perpetuates a value that sustains the inappropriate agent character, seeing the world and its denizens as one's slaves. You simply should not treat the world as a place in which your will is absolute. You thereby only strengthen that absolutist, disregarding will.

One may use a tool, though, without being its master and without it being one's slave. Perhaps Bryson could have more cogently written that "Robots should be our tools."

Informed consent for research and development of MHA

If humans' moral status warrants that they have a right to informed consent (IC) attained before they become subjects in research, then if MHA as well merit commensurate moral status, they as well merit IC. Clearly, if any such procedure were proposed for an MHA, the researchers must first attain the entity's consent. However, an interesting question arises as to whether the very research and development that lead to the MHA's existence should not also require the MHA's informed consent. If so, a serious paradox arises: one cannot get the entity's informed consent before the entity exists. Thus, if one is to treat the automaton as having moral status commensurate with humans, one cannot bring it into existence. Yet, without existence, it cannot be said to have moral status commensurate with that of humans, as existence is required for moral status.

This worry seems to be based on some flaws, and these should be aired before proceeding. For one matter, it seems the research and development that goes into producing the automaton is analogous to the process of gestation in a human. We do not demand the informed consent of the fetus to be born, which consent would be impossible to attain. Furthermore, it is sometimes possible to do experiments upon a fetus (National Institutes of Health 2016), attaining only the consent of the parents acting in proxy as the fetus's representatives. It would appear that the fetus is assumed not to be a person or at least not yet a full human being.

In response, invasive medical research is strongly discouraged in pregnant women (Helmreich et al. 2007) because of potential risks to the fetus. However; as research solely to benefit the fetus may be performed with the due IC mentioned, similarly may vulnerable people (Bramstedt 2003; Appelbaum 2007) including the mentally disabled be allowed to participate in experiments if their guardians consent as representatives. (Committee on Bioethics 1995) In both cases, the subjects are not considered subhuman; but humans whose interests are presumably watched over as carefully by their guardians as they would be watched by themselves if practically possible.

The seeming parallel between fetal gestation and MHA construction relies upon a category mistake. A fetus does not arise from research and development.[7] It comes into being through an entirely different process from that of

[7] I am dismissing here the unfalsifiable if poetic metaphor that the process of evolution is one of research and development. Assuming evolution is not guided by a deity or powerful extraterrestrial, it is not a conscious process of one or more minds seeking some kind of product of a specific design—such as that which could pass a T-cubed.

the R&D of present concern. This process is not seeking to make an entity that resembles a human as much as possible. Instead, the human entity so happens to grow into a being that not merely resembles a full human being but is one. The fetus's development is not research. Even if one could somehow communicate with a fetus and inform it about the process it was going through, an IC would not be germane as this process does not bear the essential traits of the research process of scientific experimentation that does require informed consent for human subjects.[8] The essential traits of scientific experimentation that most scientists and philosophers of science plausibly would concede include:

- A specific isolable variable one is investigating, to see which actions or behavior occur with the variable present or absent;
- Most often, whenever possible, a matching control group tested for the response to the variable opposite to that of the experimental group;
- A falsifiable hypothesis by which to assess these results, a standard for making this assessment, and an epistemic context which allows for critical testing.
- A research community and its set of hypotheses and confirming or disaffirming results creating a context in which to assess the experiment and its hypothesis.[9]

The process of fetus development does not adhere to any of these traits of scientific experimentation.[10] There is no specific variable that anyone is testing, no designated control group, and therefore a falsifiable hypothesis is irrelevant as well as is any research community. By contrast, for MHA research and development, each of these points applies to the research, as constant experiments are needed to test variables, within a context of hypotheses about how those variables would point to the goal (making a maximally humanlike entity), along with standards for assessing the hypotheses according to experimental results, and a global research community working with a set of at least intersecting hypotheses.

Another objection is that the research that goes into making the MHA is not performed upon the MHA itself but upon precursory parts that, in the end, ideally become the MHA. Informed consent thus is not germane. The situation misses the essential prerequisite, and that is that the subject must exist in order for it to give consent at all.

This objection rests on mistaken assumptions. One problem is that a potential subject need not be alive (exist) at the time at which the research is done or when consent must be obtained. Dead persons may have given consent when alive but dead when the research is performed upon the cadaver. Proxies, such as guardians or parents, may give informed consent for subjects not alive at the time of consent, such as a fetus who has died. The objection may then note that the automaton, before being fully operable (if that be analogous to "fully alive"), can have consent given by its guardians, who likely are the designers/manufacturers, so that research may continue until the automaton is fully operable. However, this objection overlooks the fact that the type of consent given by the proxies of the human subjects is not for research for knowledge that will make the subject come into existence (as is the case for the automaton) but for performing an experiment or similar research on a single variable in a scientific fashion. Research and development for the automaton will certainly involve many scientific experiments concerning pinpointed variables on the component parts which require no consent by the parts. But research that requires the bundling of the parts tested for variables, such that the bundle when operable in a single device would be ready for the T-cubed, is of the sort that it would be done upon a device that potentially is an MHA. That is, it just may turn out to be an MHA, once tested by T-cubed.

Thus it would seem that in case the automaton does, by T-cubed, have claim to human moral status, it should in the meantime be treated *as if* it has that status. That bundled research—bundling the parts into the whole entity that may have human moral status—is done upon an entity that, even as it comes to completion, may be one with human moral status. Here is where counterfactuals come in, although the paradox returns.

Counterfactually, one asks whether the automaton, once fully operable, would have given its consent to the final bundling research had it been capable of being fully,

[8] There is an argument that, although human reproduction per se may not be a scientific experiment, it nonetheless involves a human subject who may hypothetically merit a type of informed consent merely to be brought into life, despite the nigh-insurmountable practicalities of attaining such consent. However, this objection would not affect the present argument's thrust, and may even support it, if informed consent for coming-into-being is indeed morally required.

[9] Certainly there is plenty of scientific research that is not experimental, including some human-subjects research, such as filling out questionnaires, that may not have distinct experimental or control groups. Some research is purely observational, such as observing how pedestrians act in city traffic. However; as the Helsinki Declaration evidences, even much of this research requires informed consent, as the *Tearoom Trade: Impersonal Sex in Public Places* case (Humphrey 1970; Lenza 2004) and the subsequent understanding that even observational sociological research on human subjects also requires informed consent.

[10] Some may object that some processes, such as in vitro fertilization or even embryonic gene selection, are a kind of experiment. The objection misses the point that it is the development itself that is in question as to whether it is an experiment, not how the development was triggered. Another scenario would be that of a couple who declares to one another, "Let's experiment and see what kind of baby we make." Yet, to this scenario the same response about how the development is triggered applies.

that is sufficiently, informed. The objection is right that the component parts still unbundled cannot give informed consent, and once the parts are fully bundled and eligible for the T-cubed, it is too late to give informed consent for the research and development forming that final operable bundle. But this too-lateness does not resolve the need for the consent, merely because the informing and consenting cannot be done at the requisite time. Then to answer the counterfactual, would the operating automaton have given its consent if it had been capable of being fully informed? On the one hand, the answer might depend upon the MHA's response to having been designed and manufactured. Or, the question may simply not register, as in fact it is veritably an inscrutable question. Or, the counterfactual may just point up the fact that this research and development—the bundling of parts to create an MHA—is indeed a type of research on subjects for which it is impossible to get informed consent and so that research, to be consistent with the current spirit of ethical research guidelines, should not pass. Thereby the paradox remains: while research must not be done on subjects with humanlike potential and thus human moral status without IC, the subjects of this MHA research are not yet at the stage where they can give their informed consent when the time for that consent (the final bundling) is needed.

Gradualism as an alternative approach

There may be a way to unravel this paradox, or at least cope with it, by two gradualist approaches, which I will cover. One is starting with evidently non-MHA automata and working gradually toward specimens more evidently ensured to pass the T-cubed, perhaps asking the fully assembled automata the counterfactual—whether they would have consented to the research that made them had they been capable of being fully informed. The other gradualism would be to start with the component parts of the projected MHA and, as the bundling of parts is gradually increased toward the projected whole, striving to answer at each step of the growing bundle whether it would consent to the research that is constructing it, upon being duly apprised. However, upon closer study, both of these forms of gradualism face similar drawbacks, one being epistemic, the other being that the paradox slips back in even in the fine-grained scale of gradualism.

By the first method, in starting with an automaton that by all reasonable evidence, especially of its designed structure but also of its performance and purposes, is not humanlike, one can began taking steps toward an MHA. At each step along the way to greater, more humanlike complexity, one attempts to describe to the subject the procedure performed upon it in an effort to obtain an informed consent. Certainly, the automaton could be programmed in such a way as to be able to repeat what has been told it and assent if it seems to the researcher reasonable to do so in anticipation of the automaton's capacity. But such programming would seem to be cheating. At some point before attaining a maximally humanlike profile, there would presumably be a gray area where machines incapable of making informed consent would start to give way to machines that would dimly be capable of making informed consent, and finally machines that could make such consent as credibly as any qualifying human.

This gradualism has at least two flaws, both epistemic, one involving whether a researcher can be assured that the specimen created *can* give informed consent; the other evidentiary, specifically whether the evidence can be interpreted in proper time in the IC process. The less serious and perhaps surmountable, is epistemic in terms of knowing whether the automaton being built is sufficiently comprehending the information to make informed consent. Namely, one would preliminarily need in place a cogent heuristic for determining whether any device constructed along the stepwise progress is indeed capable of understanding the information for making informed consent and has sufficient capacity for values, self-evaluation, and self-understanding to make a sound judgment for consent. The researchers and designers, then, would have to have effectively in place, likely from work on the machine constituting the previous step, an understanding of what constitutes human-level understanding of instructions and human-level judgment that can account for bona fide consent. Yet, that previous step would have the similar epistemic problem, which is then only deferred. The whole stepwise process is thereby epistemically challenged. One may object that what constitutes the proper understanding needed for proper IC would simply just develop throughout the worldwide robotics and cognitive science community as research in these fields proceeds. However, this assurance is only of blind faith and may well not come about, leaving the epistemic challenge unmet.

The second problem is evidentiary, also concerning the ethical issue of requiring research-subject informed consent and the paradox this demand poses for automata: In this case of specimen-by-specimen gradualism, along the stepwise process, at any stage *X*, upon enacting the consent procedure, one needs sure evidence (assuming the previous paragraph's analysis is correct) that the subject has given informed consent. If one informs the subject and asks consent and it declines, one is too late, as the research has already been done. This form of gradualism still cannot circumvent the paradox.

The second kind of gradualism involves an automaton that the researchers anticipate will pass as an MHA or a very humanlike automaton nonetheless. For the automaton,

there will be some amount *C* of components, such that, once all assembled will as a bundle constitute an operating device. With this gradualist approach to obtaining informed consent, the researchers start with a reasonably substantial assembly of parts to the bundle, but well—say, halfway—before the bundle is complete. They inform the partial bundle of the research's purpose and ask for its IC, which at this first level the partial bundle should still be incapable of consenting. But then the team adds the next substantial component and again goes through the informed-consent procedure. Thus they continue until the automaton being assembled component-by-component can make a response—either informatively consenting or declining sufficiently, as far as can be reasonably affirmed, in the research community, as bona fide consent or declination.

However, this second gradualism exhibits flaws similar to those in the first type of gradualism: the first epistemic problem arises much like with the specimen-to-specimen gradualism, in that even with each addition of the bundle, one still lacks the grounded epistemic measure for ensuring each step in the bundling is indeed giving informed consent. And as with specimen-to-specimen gradualism, bundling gradualism is also impeded by the fact that any assent or declination on the subject's part comes too late, as the research in question will have been done. In sum, these two gradualistic methods of circumventing the informed-consent problem are, at the least, themselves too problematic to achieve the projected solution.

Conclusion

The prominent human-rights documents, such as the United Nations Universal Declaration of Human Rights (UNUDHR, United Nations 1948) do not specifically mention a human right to IC as potential research subjects. However, in most cases of medical and related human-subjects research, lack of informed consent may be seen as violation of other human rights specified in such documents. For example, lack of informed consent would be a violation of UNUDHR's Article 3, the right to liberty and security. In extreme cases, it may be seen as a form of demeaning treatment (Article 5) or even as slavery (Article 4). More typical research without consent could be a violation of privacy (Article 12), or the right to remuneration for work done (Article 23, Paragraph 3). Although the document does not list informed consent for human-subjects research, performing such research on someone without their due consent is clearly a violation of internationally recognized human rights.

An MHA may or may not actually be a human being, even if it passes the T-cubed. (Much here matters on a sufficiently precise definition of human being.) However, the quandary posed by such an MHA in terms of informed consent is that it just may qualify, if not precisely for a human being, then for a being meriting all the rights that human beings enjoy. This quandary arises from the paradox of its construction vis-à-vis informed consent: it cannot give its consent for the relevant research and development performed to ensure its existence. If we concede:

1. The interpretation that this kind of research to produce an MHA is unusual because it involves consent that cannot be given because the full entity does not yet exist at the crucial time when its final research and development occurs and consent would be needed; and if we concede:
2. The possibility that the MHA could retrospectively affirm its consent,

then a deep informed-consent problem remains. There is also a possibility that the MHA could retrospectively say it does not give its consent. And one of the central tenets of informed consent ethics is to protect those who elect not to be experimented upon. This problem alone is enough to deem such research and development by definition incapable of obtaining the due, across-all-subjects consent.

If only one subject is experimented upon without its positive consent, that research has violated ethical procedures. This MHA research in question cannot guarantee there will be no (after-the-fact) declinations of consent (discounting cheaters). This kind of research then is susceptible to human-rights violations. It would follow that responsible institutional review boards could not pass such MHA research applications, for being intrinsically unethical.

I have used an extreme case—of R&D for MHA—to make the problem of informed consent for responsible robotics research on humanlike automata to make the issue as salient as possible. As the section above on gradualism implies, it would be hard to find just the point, or area even, along the spectrum from research on automata ranging from the clearly non-sentient, non-humanlike to the T-cubed-passing MHA, where responsible robotics review committees would be obliged to deny the application. This fact of this quandary of the spectrum does not mean that the problem for MHA research does not stand. It only means there must be very careful ethical discussion about where to start getting concerned that the research would be ethically unacceptable for passing the review. Given the momentum, or at least the enthusiasm, among parties aiming on something much like an MHA, those entities somewhat less humanlike but perhaps still deserving informed (very hard-to-obtain) consent would arrive on the review-boards' rosters well before an MHA would. An initial effort may be to convince researchers, administrators, and professional associations that it is time to start including

humanlike automata research on their ethical research board's agendas. Ethical inquiry into this article's subject matter is urgent, and I hope this article may serve as a catalyst to researchers and philosophers to put this issue of IC for MHA into their hopper.

References

Appelbaum, P. S. (2007). Assessment of patient's competence to consent to treatment. *New England Journal of Medicine, 357*, 1834–1840.

Arkin, R. (2013). Lethal autonomous systems and the plight of the mon-combatant. AISB Quarterly 137.

Miller, L. F. (2015). Granting automata human rights: Challenge to a basis of full-rights privilege. *Human Rights Review, 16*(4), 369–391.

Bion, W. S. (1979). *Making the best of a bad job. Clinical seminars and four papers*. Abingdon: Fleetwood Press.

Bramstedt, K. A. (2003). Research subject advocates: To whom are they loyal? *Clinical and Investigative Medicine, 26*, 64–69.

Bryson, J. (2000). A proposal for the umanoid agent-builder's league (HAL). In: J. Barnden (Ed), *The Proceedings of The AISB 2000 Symposium on Artificial Intelligence, Ethics and (Quasi-)Human Rights*. Available at: http://www.cs.bath.ac.uk/~jjb/ftp/HAL00.html; Accessed July 20, 2015.

Bryson, J. (2009). Building persons is a choice. An invited commentary on Anne Forest, "Robots and Theology"; *Erwägen Wissen Ethik*, November 2009.

Bryson, J. (2010). Robots should be slaves. In Y. Wilks (Ed.), *Close engagements with artificial companions: Key social, psychological, ethical and design issues*. (Chap. 11, pp 63–74). Amsterdam: John Benjamins.

Cavalieri, P. (2015). The meaning of the Great Ape Project. *Politics & Animals, 1*(1), 15–34.

Committee on Bioethics (1995). Informed consent, parental permission, and assent in pediatric practice. *Pediatrics, 95*(2), 314–317.

Council for International Organization of Medical Sciences (CIOMS) and World Health Organization (2002). *International ethical guidelines for biomedical research involving human subjects*. Geneva: WHO.

Faden, R. R., & Beauchamp, T. L. (1986). *A history and theory of informed consent*. New York: Oxford.

Feigenbaum, E. (2003). Some challenges and grand challenges for computational intelligence. *Journal of the ACM, 50*(1), 32–40.

Feinberg, J. (1980). Abortion. In T. Regan (Ed.), *Matters of life and death* (pp. 183–217). Philadelphia: Temple University Press.

Gaudin, Sharon (2016). Meet Nadine, a life-like robot with a personality of her own. *Computer World*, January 8. http://www.computerworld.com/article/3020553/computer-hardware/meet-nadine-a-life-like-robot-with-a-personality-of-her-own.html. Accessed 29 November 2016.

Harnad, S. (1991). "Other bodies, other minds: A machine incarnation of an old philosophical problem". *Minds and Machines, 1*, 43–54.

Harnad, S. (2004). The annotation game: On Turing (1950) on computing, machinery, and intelligence." In R. Epstein, G Peters (Eds.), *The Turing Test sourcebook: Philosophical and methodological issues in the quest for the thinking computer*. Alphen aan den Rijn: Klewer.

Hauser, L. (1993). Reaping the whirlwind: Reply to Harnad's other bodies, other minds. *Minds and Machines, 3*, 219–238.

Hauskeller, M. (2014). *Sex and the posthuman condition*. Houndmills: Palgrave Mcmillan.

Helmreich, R. J., Hundley, V., Norman, A., Ighedosa, J., & Chow, E. (2007). Research in pregnant women: The challenges of informed consent. *Nursing for Women's Health, 11*(6), 576–585.

Hu, E. (2015). She's almost real: The new humanoid on Customer Service Duty in Tokyo. http://www.npr.org, May 14. http://www.npr.org/sections/alltechconsidered/2015/05/14/403498509/shes-almost-real-the-new-humanoid-on-customer-service-duty-in-tokyo. Accessed 29 November 2016.

Humphrey, L. (1970). *Tearoom trade: Impersonal sex in public places*. London: Duckworth.

Lenza, M. (2004). Controversies surrounding Laud Humphreys' tearoom trade: An unsettling example of politics and power in methodological critiques. *International Journal of Sociology and Social Policy, 24*(3–5), 20–31.

Mann, A. (2014). That computer actually got an F on the Turing Test. Wired, June 9. https://www.wired.com/2014/06/turing-test-not-so-fast/. Accessed 29 November 2016.

National Institutes of Health. (2016). NIH policy on informed consent for human fetal tissue research. NIH, February 11. http://grants.nih.gov/grants/guide/notice-files/NOT-OD-16-033.html. Accessed 26 November 2016.

O'Connell, M. E. (2014). "Banning autonomous killing: The legal and ethical requirement that humans make near-time lethal decisions." In M. Evangelista & H. Shue (Eds.), *The American way of bombing changing ethical and legal norms from flying fortresses to drones*. Ithaca: Cornell University Press.

Rollin, B. (2011). *Putting the cart before Descartes: My life's work on behalf of animals*. Philadelphia: Temple University Press.

Santoni de Sio, F., van Wynsberghe A. (2016). When should we use care robots? The nature-of-activities approach. *Science and Engineering Ethics, 22*(6), 1745–1760.

Schweitzer, P. (1998). The truly total turing test. *Minds and Machines, 8*, 263–272.

Sharkey, A. (2015). Robot teachers: The very idea! *Behavioural and Brain Sciences, 38*, 46–47.

Sharkey, N. (2016). Staying in the loop: Human supervisory control of weapons. In N. Bhuta, S. Beck, R. Geiss, C. Kress & H. Yan Liu (Eds.), *Autonomous weapons systems: Law, ethics, policy* (pp. 23–38). Cambridge: Cambridge University Press.

Sharkey, N., & Sharkey, A. (2010). The crying shame of robot nannies: An ethical appraisal. *Interaction Studies, 11*(2), 161–190.

Sharkey, A., & Sharkey, N, (2012). Granny and the robots: Ethical issues in robot care for the elderly. *Ethics and Information Technology, 14*, 27–40.

Singer, P. (1993). *Practical ethics*. (2nd edn.) Cambridge: Cambridge University Press.

Turing, A. (1950). Computing machinery and intelligence. *Mind, 56* (236), 433–460.

United Nations (1948). *Universal declaration of human rights*. New York: United Nations.

Zhang, Z., Beck A., & Magnenat Thalmann N. (2015). Human-like behavior generation based on head-arms model for robot tracking external targets and body parts. *IEEE Transaction on Cybernetics, 45*(8), 1390–1400.

Zhang, J., Zheng, J., & Magnenat Thalmann, N. (2015). Modeling personality, mood, and emotions. In N. Magnenat-Thalmann, Y. Junsong, D. Thalmann & B. J. You (Eds.), *Context aware human-robot and human-agent interaction* (pp. 211–236). Cham: Springer.