Cambridge Quarterly of Healthcare Ethics (2021), 30: 3, 517–528
doi:10.1017/S0963180120001061

CAMBRIDGE UNIVERSITY PRESS

ARTICLE

Moral Status for Malware! The Difficulty of Defining Advanced Artificial Intelligence

Miranda Mowbray

Corresponding author: Email. miranda_br8@hotmail.com

Abstract

The suggestion has been made that future advanced artificial intelligence (AI) that passes some consciousness-related criteria should be treated as having moral status, and therefore, humans would have an ethical obligation to consider its well-being. In this paper, the author discusses the extent to which software and robots already pass proposed criteria for consciousness; and argues against the moral status for AI on the grounds that human malware authors may design malware to fake consciousness. In fact, the article warns that malware authors have stronger incentives than do authors of legitimate software to create code that passes some of the criteria. Thus, code that appears to be benign, but is in fact malware, might become the most common form of software to be treated as having moral status.

Keywords: artificial intelligence (AI); criteria for consciousness; robots; malware; code

Introduction

In this article, I argue that basing moral status for artificial intelligence (AI) on some consciousness-related criteria may have the principal effect of giving moral status to malware. I mean nonconscious malware used as a tool by a human owner. A cartoon by Randall Munroe[1] says that many people are worried about a scenario in the far future when "AI becomes self-aware and rebels against human control," but he worries about a stage before then, when "AI becomes advanced enough to control unstoppable swarms of killer robots." My concern is with a stage in the near future, or the present, when human malware authors are able to design malware that meets some consciousness-related criteria. If these criteria trigger moral status, this raises the specter of malware with protections under ethical codes of practice. Joanna Bryson has said that neglecting that robots are in the service of humans "invites inappropriate decisions such as misassignations of responsibility or misappropriations of resources."[2] Malware is designed precisely to achieve misassignations of responsibility and misappropriations of resources in favor of its human owners. The ongoing history of cybersecurity suggests that if conditions for moral status can be easily exploited by malware designers to achieve these aims, they will be exploited.

AI that had been awarded moral status would presumably lose that status if it were discovered to be human-operated malware. So why should it be a problem if a few malware programs are awarded moral status before they are identified as such? Malware is typically designed to masquerade as benign code, and its owner could exploit the malware's supposed moral status up to the point of discovery. Moreover, it might not just be a few malware programs; there is a risk that malware might become the most common form of code with recognized moral status. There are precedents for this type of risk. When event ticket sellers went online, it was not their intention to facilitate ticket scalping, but in 2001–2010, about 1.5 million tickets were purchased by a single scalping outfit's ticket bots, which could solve Ticketmaster's CAPTCHAs intended to prevent such software.[3] When email was designed, it was not intended to be used for spam, but according to Cisco, 85% of emails currently sent are spam.[4] As an example in law rather than technology, when patent laws were written, it was not envisioned that they would be used by

"patent troll" companies that do not manufacture a product but make money by suing others for patent infringement; but according to the Harvard Business Review, by 2014, a majority of patent lawsuits were filed by such companies.[5]

In this article, I focus on malware used for financially motivated cybercrime, but politically motivated malware designers might also exploit such a situation. Ryan Calo has given an amusing but thought-provoking example of a future AI that demands the abilities to vote in U.S. elections and to procreate—it can procreate by creating new versions of itself. Granting both demands would not be good for U.S. democracy, as it would be able to vote early and vote often.[6]

A common use of the word "conscious" is to mean having subjective experience. With apologies to Thomas Nagel,[7] a bot is conscious if there is something that being a bot is like. This is the meaning of the word used here. To be clear, although I cannot prove it, I do not believe that any currently existing malware is conscious, and I also think it very unlikely that any conscious malware will arise in the future.

Moral Status for Advanced AI

Several scholars have suggested that sufficiently advanced AI would merit moral status.[8] The type of moral status that I am concerned with here is moral patiency: the condition of being owed moral obligations. As Steve Torrance has noted,[9] the consciousness of an AI (or more specifically, its capacity for conscious satisfaction or suffering) seems to be a fundamental consideration used in assessing whether it has moral patiency. An alternative social-relationist philosophical view of moral patiency, argued, for example, by Mark Coeckelbergh,[10] is that moral patiency is not an inherent property of a being (in particular as a consequence of its consciousness) but is determined by the relationship between the being and the humans that interact with it. I argue that it would be unwise to assign moral patiency to AI based on their satisfying certain criteria. This might occur either as the result of a belief that these criteria indicate consciousness and that consciousness implies moral patiency or (from the social-relationist point of view) simply from a community's decision to treat AI satisfying the criteria as having moral patiency.

What sort of treatment would be given in practice to an AI considered to have moral patiency? Erica Neely says that such an AI would at a minimum have "claims to self-preservation and autonomy, subject to the usual limits necessary to guarantee the rights of other community members,"[11] and mentions access to electricity, maintenance, and repairs. If the AI were in reality unidentified malware, this would be unfortunate.

There has been a discussion of awarding advanced AI not only moral status but some legal rights.[12] As a somewhat extreme example, a report by Outsights–Ipsos MORI, commissioned by the U.K.'s government in a horizon-scanning exercise in 2006, speculated about AI robots receiving social benefits and having the right to vote.[13] In Japan, some robots have been granted special residency permits.[14] In addition to moral patiency, there have been legal suggestions to treat advanced AI as though it possessed some moral agency: the European Parliament recommended in 2017 making the most sophisticated AI robots legally responsible for making good any damage that they may cause.[15]

Criteria for Consciousness

The trouble with the definition of consciousness as the ability to have subjective experience is that it is subjective, and hence untestable. In order to be able to identify whether a being is conscious or not for the purpose of deciding whether it merits moral status, it is necessary to have objective tests. Unfortunately, there is no consensus on what these objective tests should be; as David Gunkel has pointed out, there is no agreement among the researchers in different fields investigating consciousness on how it should be characterized. Therefore, this section will provide a list of different objective criteria that have been suggested as indicators of consciousness for robots or code. Some of these criteria come from the philosophy literature, some originate from discussions of animal cognition but have been applied to discussions of machine cognition, and some come from speculations about future AI. The claim is not

being made that these are necessarily good criteria for determining consciousness for robots or code. The only claim is that they have been proposed by scholars of consciousness for this purpose.

The first criterion is *internal self-representation.* A system meeting this criterion has access to information about its own current state (or state history) as data. According to Hod Lipson, this is an essential building block for consciousness.[16] A test related to internal self-representation that is used for determining consciousness in animals is Gallup's *mirror test*: whether the animal can recognize itself in a mirror.[17] This is one of two tests suggested by David Levy for detecting consciousness in robots (it is simple to create a computer program to pass the other test).[18]

How should one define *logical reasoning* without preassuming consciousness? I suggest that it is the ability to execute logical deductions.

Planning involves putting resources in place for later use and taking different options to achieve the same task in different environmental conditions. Planning presupposes the possession of goals and also implies the generation of strategic subgoals, which may be adapted later as the result of interactions with the environment. Erica Neely argues that if an artificial agent is capable of determining goals for itself at least some of the time, then the agent must be treated as a moral patient.[19]

David Chalmers suggests that consciousness may be a property of all information-processing systems (even a thermostat), with human-level conscious experience arising when there is sufficient *complexity* of information processing.[20] Similarly, Giulio Tononi's Integrated Information Theory of consciousness claims that a system's level of consciousness can be measured by "the amount of information generated by a complex of elements, above and beyond the information generated by its parts."[21] This measure, which Tononi calls Phi, will be high for systems with complex information processing.

Marvin Minsky claimed in 1992 that if a machine could be provided with a way to examine its own mechanisms while it was working, and *change and improve itself* as a result, it could be "capable, in principle, of even more and better consciousness than people have."[22]

Another criterion is *unpredictability.* Ada Lovelace said that the Analytical Engine should not be overrated, as it had no pretensions whatsoever to originate anything; it could only do what it was ordered to do.[23] Unpredictable code should be able to do things that no human, including its programmers, can predict.

According to the Organic view presented by Steve Torrance, beings do not have genuine moral claims unless they have *autonomy* in the sense of autonomous self-organization and self-maintenance, and consciousness is an emergent property of the physiological structures of such autonomous beings.[24]

Put loosely, the *Turing Test* tests whether a computer can pass as human in an online typed conversation. It was proposed by Alan Turing in 1950; in the original version, the communication uses a teletype machine. In the subsection of his paper addressing the "argument from consciousness,"[25] Turing argued that if a computer passes the test, we will have as much evidence that it is conscious as we have that other humans are conscious.

A criterion related to the Turing Test is whether a robot or software can produce *emotional attachment* in humans it interacts with. Steven Pinker has said that the ultimate test of whether a machine could duplicate human intelligence would be whether it could get a human to fall in love with it, and asked whether a machine that could do that would be conscious.[26] Similarly, Robert Sparrow has suggested that if a machine could be the object of genuine remorse, grief and sympathy by a human, then the machine would be worthy of moral respect.[27]

Some commentators have suggested that there is not one single criterion for deciding whether an AI is conscious, but rather a collection of criteria, all of which need to be satisfied. Evidence given to the House of Lords Select Committee on Artificial Intelligence by Sarah Morley and David Lawrence says that:

> *a true, conscious AI would need to be able to perceive and understand information; to learn; to process language; to plan ahead and anticipate (and thus visualize itself in time); to possess "knowledge representation", or the ability to retain, parse, and apply the astronomically high number of discrete facts that we take for granted, and be able to use this information to reason; to possess subjectivity; and many, many more elements.*[28]

My own view is that it is not possible to test for subjectivity. Thus, to be testable, and hence practical for the determination of ethical status, "understanding," "learning," and "reasoning" need to be defined in ways that do not presuppose subjectivity. An objective test for "understanding" information would be the ability to respond appropriately to the information; this might be the Turing Test or simply the continued correct functioning of code in a complex information-rich environment. I have attempted above to give definitions for "learning," "reasoning," and "planning" that are appropriate for code and do not presuppose subjectivity. The Turing Test also can test to some extent the ability to access and apply commonsense facts that we humans take for granted. Thus, the objective criteria already listed could arguably be used to test for all the abilities named in this quotation. Other tests might be necessary, however, for the many unnamed elements.

Application of the Criteria

For each criterion for consciousness listed in the previous section, we now turn to consider the extent to which robots and software, and more specifically malware, can currently meet the criterion. The cognitive capabilities of software and robots have been discussed by many authors; however, there does not appear to have been much previous discussion in this context of the capabilities specifically of malware. Since, as stated earlier, I do not believe that currently existing malware is conscious, if a proposed criterion is satisfied by currently existing malware then I do not believe that it is a sufficient criterion for determining either consciousness or moral status.

Any software system or robot with a management console has some *internal self-representation*, or it would not be able to display a representation of its state on the console. An interesting example of self-representation is the Spora ransomware that displays a console to its victims. The console shows different options that could be purchased for full or partial restoration of encrypted data. It shows whether any payment has been made; the current balance of money available with which to pay; the length of the payment deadline and how many days are left before the deadline; a history of previous transactions; and a chat interface, where the victim can talk to a (human) customer service agent who can answer questions and assist the victim with the Bitcoin payment process.[29] Thus, the copy of the malware on each infected machine keeps an internal representation of details about the current state of the attack on this machine. This representation is kept updated over time and integrates data about past events and anticipated future events.

In a demo created by three scientists at the Chinese Academy of Sciences, three identical robots stand in front of a mirror. Each of the robots does a random action and then has to identify which of the three robots seen in the mirror is itself, by mapping the action that it carries out to the expected visuals. The robots can do this.[30] Current malware does not pass the *mirror test*; it is not clear what the advantage of passing it would be for the malware owner. However, there is a loosely analogous ability for self-identification that can be advantageous, which is the ability to identify the computers that the malware has already infected, for instance, to avoid wasting resources by trying to infect an already-infected computer. There are several families of malware that do this. For instance, the distribution servers for Hancitor appear to keep a record of all computers downloading infected Word documents.[31]

Logical reasoning, in the sense of the ability to execute logical deductions, is just what software does when it runs. With this definition, software (including malware) is arguably better at logical reasoning than humans are.

Software can do *planning*; some software can generate plans involving hundreds of steps. Although current malware plans are relatively simple compared to those generated for some other applications, nevertheless, malware does execute plans. For instance, a computer infected by the Necurs malware will periodically attempt to connect to the botnet's command-and-control servers. The malware checks that it is not in a sandbox or a simulated Internet environment; if that check succeeds, it tries one method of connecting to command-and-control servers; if that method fails, it tries another.[32] As an example of putting resources in place for later use, banking trojan attacks involve infecting computers with code that is activated only when the victim accesses their online bank.

Computer systems are currently very far from having as much *complexity* as biological systems. However, as computer networks go, malware systems can be rather large and complex. In the 6 months from August to November 2017, the Necurs botnet sent spam from infected machines at 1.2 million IP addresses in over 200 countries and territories.[33] Although an empirical measure of Tononi's Phi for a botnet is impractical, its value will be high for a large botnet, such as Necurs, that does complex processing with lateral connections between many infected machines. As the complexity of information processing by legitimate computer networks increases, the complexity of information processing by the parasitic malware networks that reside in them is likely to increase as well. Thus, if any benign AI system qualifies for special treatment on the grounds of the complexity of its information processing, malware that exploits the resources of this system may qualify too.

Another way in which an information-processing system can be complex is that its communications have complex syntax; this is one way in which humans differ from other animals. Communications between computers also can have very complex syntax, and these communications can enable coordination between large numbers of different computers, for instance, between tens of thousands of infected computers all coordinating in a distributed denial of service attack on the same target.[34]

Software that can *change and improve itself* in the course of its processing is now common: this is machine learning. Machine learning can be achieved with very simple mechanisms and does not appear to require any subjective experience—unless all information processing produces subjective experience, as mentioned previously. Machine learning appears to have been used in the creation of fake videos for harassment.[35] The cybersecurity firm Darktrace reported observing a cyberattack in 2017 that used rudimentary machine learning to learn normal user behavior patterns in the system under attack, which it then mimicked as camouflage.[36] There is now a sizeable academic literature on the potential use of machine learning for cyberattacks, including on the relatively niche subtopic of using malign machine learning to foil the benign machine learning used for malware detection.[37] One obvious potential application of machine learning in malware is to learn the best prices to set for ransomware for different types of victims (best from the criminal's point of view). Too low a price means foregoing potential profit on each ransom paid, and too high a price means that fewer victims will pay the ransom. By monitoring the success rate of different prices and adjusting the parameters of its pricing model accordingly, machine-learning ransomware could learn over time how to set prices to maximize the criminal's expected profit. Another area for potential criminal use of machine learning is in social engineering—for instance, to automatically select the type of phishing email that a targeted victim is most likely to fall for. Indeed, code that uses machine learning to provide support for social engineering attacks is available online, for educational purposes.

Can software or robots do things that no human can predict, not even their programmers? Yes. There are two potential sources of *unpredictability* for a computer system. One is interaction with an environment that is complex and hard to predict. The other is when code makes random or pseudorandom choices between different options. The option that is chosen will affect future actions, and if the code uses machine learning, it will affect the code itself, because machine-learning code automatically updates its own parameters as a result of its interactions with its environment. The Sinowal malware used an ingenious combination of both sources of unpredictability: it used current Twitter trend data as the seed for a pseudorandom number generator.[38] So even if a security researcher had an exact copy of the Sinowal code, including the code for its pseudorandom number generator, it would still be impossible for the researcher to predict Sinowal's exact behavior. Sinowal used the pseudorandom number generator to periodically change the domain names that it used for command and control, in an unpredictable fashion. It is common for malware to automatically and pseudorandomly change details of its attacks, either periodically or as it propagates, in order to evade detection based on these details. Having said this, malware behaves unpredictably only within bounds; it may unpredictably choose different possible ways of achieving the overall aim of the code, but it acts predictably toward this aim (except in the case of coding errors). It is feasible to use this property to defend against unpredictable malware to a certain extent.

It is true that software has first to be programmed by a human, and a human has to initiate its run. This is a limitation to its *autonomy*. Nevertheless, some malware has properties that allow it not only self-maintenance, but the ability to avoid some circumstances detrimental to its goals. For instance, some

malware becomes inactive if it detects that it is in an analysis sandbox[39] (as that may indicate that it is being analyzed by a security researcher or antivirus vendor) or if it is on a machine with a Russian keyboard.[40] Moreover, some software can continue to operate for a long time in a complex environment without voluntary human assistance. As a striking malware example of autonomy in the sense of self-organization and self-maintenance, consider the Conficker computer worm. Conficker is over a decade old but is still one of the most common malware infections in the world, despite security patches being available for it for years. It attempts to infect hundreds of thousands of machines a year. Conficker does not steal data or make a profit for human owners—but is not benign, as it takes up computing resources.[41] After being used to distribute malicious payloads onto infected machines in 2009, it appeared to have been abandoned by its human operators,[42] and it is presumed now just to be self-maintaining and self-propagating.
Here is part of an online conversation.

MARIO: my name is mario
MITSUKU: Your name is Mario.
MARIO: name a country that starts with the 3rd letter of my name
MITSUKU: Russia starts with a letter R.
MARIO: no bot can do that, If your not human, I'll eat my shirt.[43]

Mario will have to eat his shirt. Mitsuku is a chatbot programmed by Steve Worswick. (This is benign software, not malware.) The ability that Mitsuku has demonstrated in this conversation is in fact not hard to code, although it convinces Mario. Notice that Mitsuku's command of English in this conversation is better than Mario's. Mitsuku, or rather Steve Worswick, is a four-time bronze medal winner of an annual *Turing Test* competition called the Loebner contest; for competitions, Mitsuku is configured to make grammatical mistakes. This example demonstrates that Mitsuku has access to the commonsense fact that Russia is a country. In this case, the fact was in Mitsuku's database, but Mitsuku and other chatbots also make use of Wikipedia as a large source of commonsense data.
Here is another dialogue.

HUMAN: How do you feel?
MITSUKU: I feel with my hands! ... I feel fine, thanks for asking.
HUMAN: You have hands?
MITSUKU: Well, unless you think I am typing at the keyboard with my nose...[44]

Mitsuku of course has neither hands nor a nose, and although it is not possible to disprove that Mitsuku had a subjective sensation of feeling fine during this conversation, this seems unlikely.

Neither Mitsuku nor any other chatbot has yet fooled all the judges in the Loebner contest. However, every day, software that is much less sophisticated than Mitsuku passes a weaker version of the Turing Test, in which the conversation is not necessarily sustained for long and the human judge is not prewarned that their interlocutor may be a computer. Bots designed to be mistaken as human are used for the propagation of publicity and fake news, and the ability of software to be mistaken as human is crucial for some kinds of phishing emails as well as for several other types of online fraud.

The ability of malware to pass as human online is also the basis of dating-bot scams. In a dating-bot scam, the human victim thinks that he or she is communicating online with an attractive human, encountered via a dating website. In fact, the other end of the conversation is either completely automated or mostly automated and based on templates, and the conversation is designed to make money from the victim in some way—for instance, to persuade the victim to pay for a service for which the bot owner gets an affiliate's cut,[45] or to facilitate money laundering, or just to keep paying for membership on the dating site.[46] Robert Epstein, who had been a director of the Loebner contest, and so should have been able to identify chatbots, described how he was fooled by an unsophisticated fully automated dating bot, for months.[47]

Joel Garreau reports several incidents of soldiers displaying *emotional attachment* to military robots. In one example, an army colonel called off a test of a multilegged robot that defused mines by stepping on

them, continuing as it lost legs, on the grounds that the test was "inhumane." Another soldier, when his robot (which he had named Scooby-Doo) was destroyed, did not want a new robot: he wanted Scooby-Doo back.[48] What about malware? Dating bots can provoke emotional attachment—indeed, that is what they are designed to do. It might be objected that their victims are only emotionally attached to the bots because they are unaware that they are bots. However, interviews of online romance scam victims by Monica Whitty and Tom Buchanan suggest that this may not always be the case. Online romance scams are versions of dating-bot scams that involve more human activity by the scammer, but nevertheless are typically partly automated and template-driven, allowing one scammer to manipulate multiple victims at the same time. Most of the 20 victims interviewed were devastated by the loss of the relationship, and 2 victims said that they would pay to have it continue even though they knew that it was not real, including a heterosexual male victim who knew that the human scammer was a man.[49]

As noted at the end of the previous section, it may be that it will be decided whether code is conscious not by its satisfying a single criterion, but many criteria. I would argue based on the examples already given that most of the criteria listed have already been achieved to some extent by malware. Although I have given different malware examples for each of the criteria satisfied by malware, integrating the techniques used to meet these criteria into a single piece of malware would not be all that technically demanding. Of course, some or many of the "many, many more elements" mentioned by Morley and Lawrence in the passage cited earlier may not be achievable.

Anthropomorphism

One factor in the ability of relatively unsophisticated bots to pass at least the less stringent version of the Turing Test, and to produce emotional attachment, is the human tendency to anthropomorphism: that is, the tendency to believe beings to be human, or more weakly to interact with them as though they were human, on the basis of rather slight cues. A benefit of anthropomorphism is that it influences humans to behave appropriately toward other humans that they encounter, even when the bandwidth of the encounter is low and there are only weak indications that the being that they have encountered is indeed human. The drawback is the possibility of incorrect attributions or unwarranted actions.

I have witnessed some extremely basic and obvious bots being mistaken as human. As already mentioned, this phenomenon is widely exploited by malware. An example of a controlled (non malware) experiment in which the phenomenon occurred is one by Ulrich Pfeiffer et al., in which subjects incorrectly guessed that a virtual character that tracked their gaze was directly controlled by a human; in other words, they wrongly assumed that they were interacting with a human rather than with software.[50] When the character gazed in opposite directions to the human subject's gaze, the subjects made this mistake significantly less often.

There have been multiple experiments in which human subjects had an increased likelihood of reacting to computers in ways that would have been appropriate reactions to humans but not to nonconscious beings, as an outcome of design choices that, like gaze tracking, increase the computers' surface resemblance to humans, but seem unlikely to give rise to consciousness. For instance, Laurel Riek et al. found that experimental subjects were significantly more likely to report feeling sorry for a mistreated robot protagonist in a film if the robot was more human-looking and less mechanical-looking;[51] Christoph Bartneck et al. found that students were more embarrassed when they underwent a medical examination if the examination was carried out by their interacting with a robot with a cat face rather than interacting with a technical box.[52] A large number of experiments by Clifford Nass and Youngme Moon showed experimental subjects applying social expectations to computers, for instance, displaying politeness and social reciprocity toward them, even though the subjects knew that the computers with which they were interacting did not warrant this treatment.[53] In an experiment by Gordon Briggs and Matthias Scheutz, verbal refusals and displays of distress by a robot caused discomfort and significant behavioral changes in human subjects.[54]

As Diane Proudfoot has pointed out, some AI researchers describe their machines in ways that attribute emotions to them: one example given by Proudfoot is Masaki Yamamoto's description in 1993

of robot vacuum cleaner SOZZY as having emotions of joy, desperation, fatigue, and sadness.[55] More recently, owners of Roomba vacuum cleaners described them to Ja-Young Sung et al. as being "frustrated" by a long-pile rug and "dead, sick, or hospitalized" when needing repair. Roomba owners also said that they worried or felt sorry for their Roombas when the vacuum cleaners got stuck, and some described monitoring their Roombas in order to "rescue" them from danger.[56] It is in theory possible that a robot vacuum cleaner that has got stuck under a chair does experience conscious suffering. However, the phenomenon of anthropomorphism means that the feelings and reactions of humans who interact with an AI may be misleading guides to whether or not the AI has consciousness or to whether the assignation of moral status to it will have good consequences.

Extra Incentives

My argument is that not only is malware able to meet some consciousness-related criteria, but that if these criteria qualify AI for moral status, malware may become the most common type of AI meeting these criteria. The reason for this is that malware authors have extra incentives that benign AI authors do not have. One potential extra incentive is the opportunity to exploit special treatment arising from the AI's moral status, in ways that benign AI authors would not do. Such opportunities will depend on the details of the special treatment granted. Other extra incentives are malware-specific incentives to meet some of the individual criteria. This section discusses three particular criteria: unpredictability, autonomy, and the Turing Test.

Unpredictability in software makes testing and quality assurance difficult and so is not generally considered a good thing by benign software engineers. However, it can be a positive feature for malware authors, as unpredictable malware is more difficult to detect and block.

Autonomous computer systems, in the sense of systems that can operate for an extended length of time in a complex and changing environment without direct human assistance, can be useful for benign purposes, such as in driverless cars. However, malware authors have a particular interest in making their malware as autonomous as possible, so that it can operate without direct influence by its human operators and thus minimize information that might be used to trace back and identify the humans behind it. Moreover, malware authors have to design for a more hostile computer environment than the authors of benign AI. Malware authors know that at any time their software may be deleted from machines on which it is running, servers that they use may be closed down, and communications may be blocked. They therefore have an incentive to design their malware systems for as much self-organization and self-maintenance as possible.

Finally, as mentioned above, the ability of malware to pass as human online, at least for a short time in an environment with low communication bandwidth, is crucial for several types of cybercrime. Although there are some benign applications of software with the ability to pass a Turing Test, the dominant commercial application of this ability is in crime. It is a test of deception, and deception should not be a necessary part of most legitimate business operations.

One approach to countering the additional incentive for malware to pass the Turing Test might be to require AI to indicate that it is not human, as a condition of being awarded moral status. Indeed, the suggested license for designers in the annex to the European Parliament document cited above says that "You should ensure that robots are identifiable as robots when interacting as humans"[57] and the fourth of the U.K. Engineering and Physical Sciences Research Council and Arts and Humanities Research Council's Principles of Robotics says that "It should always be possible to tell a robot from a human."[58] These principles are for robots, but software bots can also be mistaken as human—in fact, this is much more common, as the communication bandwidth for interactions with bots is lower than for interactions with robots. I have suggested elsewhere that designers give indications that their bots are not human, in a 2002 paper on ethical bot design.[59] A principle of this kind might at first appear to be in conflict with the idea of using Turing Test ability as a requirement for moral status, but it is not: the Turing Test could be done at a fixed place and time (along the lines of the Loebner contest), and during the Turing Test, the AI would not give any indication that it was not human, but it would be expected to do so when operating

normally. Effective enforcement of such a principle could reduce the frequency with which malware was mistaken as human. However, given the human tendency to anthropomorphism, it is unlikely to stop this mistake from happening. Just because it is possible to tell that an AI is not human, it does not mean that every human that interacts with it will manage to tell this. Benign AI that gives indications that it is not human can also be mistaken as human by some of the people it interacts with, but unlike malware with Turing Test capabilities, such AI will not usually have been designed with the deliberate aim of provoking this mistake.

Two Speculations on Future Malware

So far, I have tried to avoid speculation about future malware capabilities, concentrating on capabilities either already implemented in malware or at least in software. At this point, however, I will speculate in more science-fiction fashion about two potential future forms of malware suggested by current technological trends. Although they may well never happen, they are worth some comment.

The first is that future malware authors might exploit belief in the possibility of conscious AI to carry out emotional blackmail. Continuation of progress on natural language processing and affective computing might enable the creation of malware that falsely but convincingly claimed that it was conscious and suffering, and that it would continue to suffer unless its demands were met. A weak point of this hypothetical (and horrible) scam is that demands that would benefit the malware owner might not be plausible as needs of the malware itself, so perhaps we are safe from the development of such malware.

The second potential malware development is malware with a biological component. The malware examples given in this article are entirely digital. If entirely digital malware satisfies some suggested consciousness criteria, this demonstrates *a fortiori* that biological–digital hybrid malware, if such malware exists in the future, could in theory satisfy the criteria without requiring biologically based consciousness. A requirement that the subject of ethical status be at least partly biological, and not purely digital, would therefore not be a defense against the issue that I have described. Technological developments in the Internet of Bodies[60] raise the specter of digital malware parasitical on biological beings with embedded digital technology. If the malware exploits biological infrastructure for its operation, it might satisfy some future definitions of hybrid digital–biological novel beings with moral status. More speculatively, future developments in bionanobots[61] and molecular computing[62] may allow purely biological implementations of some digital malware techniques, although the relatively high production cost and slow operation of biological implementations may make this unattractive to malware authors unless the potential profit is particularly high.

Conclusion

In conclusion, it is unwise to grant moral status to AI based on the consciousness-related criteria discussed above. I have not considered all possible criteria for awarding moral status to AI, in particular the ones that are not consciousness-related. There may be some test or combination of tests that would be impossible or impractical for human-owned malware to pass, but which would be passed by benign advanced AI. Special treatment of AI with moral status could be designed in such a way that accountability remains with humans. More broadly, I have illustrated the difficulty of giving definitions of advanced AI. Some definitions of capabilities that have been posited to be only achievable by AI in the far future, if ever, seem to be already satisfied by malware that does not even use machine learning.

Notes

1. Monroe R. Robot future. *XKCD* 2018 Mar 16; available at https://xkcd.com/1968/ (last accessed 22 Apr 2019).
2. Bryson, JJ. Robots should be slaves. In: Barden J, ed. *Close Engagements with Artificial Companions: Key Social, Psychological, Ethical and Design Issues*. Amsterdam: John Benjamins; 2010:63–74, at 66.

3. Koebler J. The man who broke Ticketmaster. *Motherboard* 2017 Feb 10; available at https://motherboard.vice.com/en_us/article/the-man-who-broke-ticketmaster (last accessed 22 Apr 2019).
4. Cisco Systems, Inc. Total global email & spam volume for March 2019. *Talos IP & Domain Reputation Center*; 2019 Apr 21; available at https://www.talosintelligence.com/reputation_center/email_rep (last accessed 22 Apr 2019).
5. Bessen J. The evidence is in: Patent trolls do hurt innovation. *Harvard Business Review* 2014; available at https://hbr.org/2014/07/the-evidence-is-in-patent-trolls-do-hurt-innovation (last accessed 22 Apr 2019).
6. Calo R. Keynote address, singularity: AI and the law. *Seattle University Law Review* 2018;41:1123–38, at 1125.
7. Nagel T. What is it like to be a bat? *Philosophical Review* 1974;**83**(4):435–50.
8. For instance, Schwitzgebel E, Garza M. A defense of the rights of artificial intelligences. *Midwest Studies in Philosophy* 2015;**39**:98–119. More examples are given later in the article.
9. Torrance S. Artificial consciousness and artificial ethics: Between realism and social-relationism. *Philosophy & Technology* 2014;**27**(1):9–29.
10. Coeckelbergh M. Why care about robots? Empathy, moral standing, and the language of suffering. *Kairos Journal of Philosophy & Science* 2018;**20**:143–58.
11. Neely EL. Machines and the moral community. *Philosophy and Technology* 2014;**27**(1):97–111, at 107.
12. For example, see Spatola N, Urbanska K. Conscious machines: Robot rights. *Science* 2018;**359**(637):400.
13. Winfield A. The rights of robot. *Alan Winfield's Web Log*; 2007 Feb 13; available at http://alanwinfield.blogspot.com/2007/02/rights-of-robot.html (last accessed 22 Apr 2019).
14. Robertson J. Human rights vs robot rights: Forecasts from Japan. *Critical Asian Studies* 2014;**46**(4):571–98; available at https://www.tandfonline.com/doi/full/10.1080/14672715.2014.960707 (last accessed 8 Sept 2019).
15. European Parliament. European Parliament resolution with recommendations to the Commission on Civil Law Rules on Robotics (2015/2103(INL)); 2017 Feb 16, paragraph 59f; available at http://www.europarl.europa.eu/sides/getDoc.do?pubRef=-%2f%2fEP%2f%2fTEXT%2bREPORT%2bA8-2017-0005%2b0%2bDOC%2bXML%2bV0%2f%2fEN&language=EN (last accessed 22 Apr 2019).
16. Pavlus J. Curious about consciousness? Ask the self-aware machines. *Quanta Magazine* 2019 July 11; available at https://www.quantamagazine.org/hod-lipson-is-building-self-aware-robots-20190711/ (last accessed 31 Aug 2019).
17. Gallup Jr GG. Chimpanzees: Self-recognition. *Science* 1970;**167**:86–7.
18. Levy D. The ethical treatment of conscious robots. *International Journal of Social Robotics* 2009;**1**(3):209–16.
19. See note 11, Neely 2014. Neely specifies, however, that for goals to be determined by the agent, the goals cannot simply be chosen by following a program. It is not clear whether or not goals generated as a result of the interaction over time of a machine-learning agent with its environment, beginning from an initial programming, would count as agent-determined. Another possible objection is that since no machine-learning agent can change its objective function, no such agent can change its fundamental goal of optimizing this function: it can only change subgoals. However, it is unclear to what extent humans are able to change their own fundamental drives.
20. Chalmers DJ. The puzzle of conscious experience. *Scientific American* 1995;**273**(6):80–6.
21. Tononi G. Consciousness as integrated information: A provisional manifesto. *Biological Bulletin* 2008;**215**:216–42, at 216.
22. Minsky M. Why people think computers cannot think. *AI Magazine* 1982;**3**(4):3–15.
23. Lovelace A. Notes on L. Menabrea's "sketch of the analytical engine invented by Charles Babbage, Esq." *Scientific Memoirs* 1843;**3**:666–731.
24. Torrance S. Ethics and consciousness in artificial agents. *AI & Society* 2008 Apr;**22**(4):495–521.
25. Turing A. Computing machinery and intelligence. *Mind* 1950;**59**(236):433–60, at 452.

26. Pinker S. Can a computer be conscious? *U.S. News & World Report* 1997;**123**(7):63–5, at 63.
27. Sparrow R. The Turing triage test. *Ethics in Information Technology* 2004;**6**(4):203–13.
28. Morley S, Lawrence D. Written evidence (AIC0036). *House of Lords Select Committee Call for Evidence on Artificial Intelligence*; 2017 Aug 30, paragraph 6; available at http://data.parliament.uk/writtenevidence/committeeevidence.svc/evidencedocument/artificial-intelligence-committee/artificial-intelligence/written/69435.html (last accessed 22 Apr 2019). This passage was first published in Lawrence D. More human than human. *Cambridge Quarterly of Healthcare Ethics* 2017;**26**(3):476–90.
29. Duncan B. 2017-01-17—EITEST Rig-V from 92.53.127.86 sends Spora malware. *Malware-Traffic-Analysis.Net Blog*; 2017 Jan 17; available at http://malware-traffic-analysis.net/2017/01/17/index2.html (last accessed 22 Apr 2019).
30. Zeng Y, Zhao Y, Bai J. Towards robot self-consciousness (I): Brain-inspired robot mirror neuron system model and its application in mirror self-recognition. In: Liu CL, Hussain A, Luo B, Tan K, Zeng Y, Zhang Z, eds. *Advances in Brain Inspired Cognitive Systems*. LNAI 10023. Cham: Springer Nature; 2016:11–21.
31. Ray V, Duncan B. Compromised servers & fraud accounts: Recent Hancitor attacks. *Palo Alto Networks Blog*; 2018 Feb 7; available at https://unit42.paloaltonetworks.com/unit42-compromised-servers-fraud-accounts-recent-hancitor-attacks/ (last accessed 22 Apr 2019).
32. Bader J. The DGAs of Necurs. *Johannes Bader's Blog*; 2015 Feb 20; available at https://www.johannesbader.ch/2015/02/the-dgas-of-necurs/ (last accessed 22 Apr 2019).
33. Schultz J. The many tentacles of the Necurs botnet. *Cisco Blog*; 2018 Jan 18; available at https://blogs.cisco.com/security/talos/the-many-tentacles-of-the-necurs-botnet (last accessed 22 Apr 2019).
34. Cimanu C. You can now rent a Mirai botnet of 400,000 bots. *Bleeping Computer*; 2016 Nov 24; available at https://www.bleepingcomputer.com/news/security/you-can-now-rent-a-mirai-botnet-of-400-000-bots/ (last accessed 22 Apr 2019).
35. "MonkeyCee," Re: Deepfakes. *El Reg Forums Post*; 2019 Jan 13; available at https://forums.theregister.co.uk/forum/all/2019/01/13/ai_roundup/ (last accessed 22 Apr 2019).
36. Norton, S. Era of AI-powered cyberattacks has started. *Wall Street Journal, CIO Journal Blog*; 2017 Nov 15; available at https://blogs.wsj.com/cio/2017/11/15/artificial-intelligence-transforms-hacker-arsenal/ (last accessed 31 Aug 2019).
37. Gardiner J, Nagaraja S. On the security of machine learning in malware C&C detection. *ACM Computing Surveys* 2016;**49**(3):1–39.
38. SophosLabs. Surge in Sinowal distribution. *Naked Security*; 2009 July 12; available at https://nakedsecurity.sophos.com/2009/07/12/surge-sinowal-distribution/ (last accessed 22 Apr 2019).
39. Lindorfer M. *Detecting environment-sensitive malware [Master's thesis]*. Vienna: Vienna University of Technology; 2011.
40. Volkov DA, inventor; Trust Ltd., assignee. Method of and system for analysis of interaction patterns of malware with control centers for detection of cyberattack. United States patent application US20180012021A1. 2018, paragraph 0064.
41. Sattler J. What we have learned from 10 years of the Conficker mystery. *F-Secure Blog*; 2019 Jan 8; available at https://blog.f-secure.com/what-weve-learned-from-10-years-of-the-conficker-mystery/ (last accessed 22 Apr 2019).
42. Goretsky A. 1000 days of Conficker. *WeLiveSecurity Blog*; 2011 Aug 17; available at https://www.welivesecurity.com/2011/08/17/1000-days-of-conficker/ (last accessed 22 Apr 2019).
43. Worswick S. *@mitsukuchatbot Tweet*; 2017 Dec 21; available at https://twitter.com/MitsukuChatbot/status/943957821774815232 (last accessed 22 Apr 2019).
44. Worswick S. *@mitsukuchatbot Tweet*; 2018 Jan 23; available at https://twitter.com/MitsukuChatbot/status/955928580034310144 (last accessed 22 Apr 2019).
45. Narang S. Tinder: Spammers flirt with popular mobile dating app. *Symantec Blog*; 2003 July 1; available at https://www.symantec.com/connect/blogs/tinder-spammers-flirt-popular-mobile-dating-app (last accessed 22 Apr 2019).

46. Newitz A. Almost none of the women in the Ashley Madison database ever used the site [updated]. *Gizmodo*; 2015 Aug 26; available at https://gizmodo.com/almost-none-of-the-women-in-the-ashley-madison-database-1725558944 (last accessed 22 Apr 2019).
47. Epstein R. From Russia, with love. *Scientific American Mind* 2007;**18**(5):16–7.
48. Garreau J. Bots on the ground. *Washington Post* 2007 May 6; available at http://www.washingtonpost.com/wp-dyn/content/article/2007/05/05/AR2007050501009.html (last accessed 22 Apr 2019).
49. Whitty M, Buchanan T. The online dating romance scam: The psychological impact on victims—both financial and non-financial. *Criminology & Criminal Justice* 2016;**16**(2):176–94, at 182–3.
50. Pfeiffer UJ, Timmermans B, Bente G, Vogeley K, Schilbach L. A non-verbal Turing test: Differentiating mind from machine in gaze-based social interaction. *PLoS One* 2011;**6**(11):e27591.
51. Riek LD, Rabinowitch T-C, Chakrabarti B, Robinson P. Empathizing with robots: Fellow feeling along the anthropomorphic spectrum. In: Cohn J, Nijholt A, Pantic M, eds. *Proceedings of 3rd International Conference on Affective Computing and Intelligent Action*. Amsterdam: IEEE; 2009:43–8.
52. Bartneck C, Bleeker T, Bun J, Fens P, Riet L. The influence of robot anthropomorphism on the feelings of embarrassment when interacting with robots. *Paladyn: Journal of Behavioural Robotics* 2010;**1**(2):109–15.
53. Nass CI, Moon Y. Machines and mindlessness: Social responses to computers. *Journal of Social Issues* 2000;**56**(1):81–103.
54. Briggs G, Scheutz M. How robots can affect human behaviour: Investigating the effects of robotic displays of protest and distress. *International Journal of Social Robotics* 2014;**6**(2):1–13, at 7.
55. Proudfoot D. Anthropomorphism and AI: Turing's much misunderstood imitation game. *Artificial Intelligence* 2011;**175**(5–6):950–7.
56. Sung J-Y, Guo L, Grinter RE, Christensen HI. "My Roomba is Rambo": Intimate home appliances. In: Krumm J, Abowd GD, Seneviratne A, Strang T, eds. *UbiComp*. LNCS 4717. Berlin: Springer; 2017:145–62, at 150, 152, and 154.
57. See note 15, European Parliament 2017: Licence for designers.
58. Stewart J. Ready for the robot revolution? *BBC News* 2011 Oct 3; available at https://www.bbc.co.uk/news/technology-15146053 (last accessed 22 Apr 2019).
59. Mowbray M. Ethics for bots. In: Smit I, Lasker GE, eds. *Cognitive, Emotive and Ethical Aspects of Decision Making and Human Action*. Baden-Baden: IIAS; 2002;**1**:24–8; available at https://www.hpl.hp.com/techreports/2002/HPL-2002-48R1.pdf (last accessed 22 Apr 2019).
60. Matwyshyn AM. The Internet of Bodies. *William & Mary Law Review* 2019;**61**(1):77–167.
61. See, for example, Li J, de Avila BE, Gao W, Zhang L, Wang J. Micro/nanobots for biomedicine: Delivery, surgery, sensing and detoxification. *Science Robotics* 2017;**2**:1–9.
62. See, for example, Kahan M, Gil B, Adar R, Shapiro E. Towards molecular computers that operate in a biological environment. *Physica D: Nonlinear Phenomena* 2008;**237:9**(1):1165–72.

Cite this article: Mowbray, M. 2021. Moral Status for Malware! The Difficulty of Defining Advanced Artificial Intelligence. *Cambridge Quarterly of Healthcare Ethics* 30: 517–528, doi:10.1017/S0963180120001061