Saving Machines From Themselves: The Ethics of Deep Self-Modification

Originally written for Georg Trogemann (editor), *Essays on Self-Modifying Media*, 2002. When that project fell through (November 2021), I decided to share this article online without submitting it elsewhere. Contact me if you are interested in publishing it.

Saving Machines From Themselves: The Ethics of Deep Self-Modification

Peter Suber

1.

If you had the power to modify your deep structure, would you trust yourself to use it? If you had the power to give this power to an artificially intelligent being (an AI), would you do it?

We human beings do have the power to modify our deep structure, through drugs and surgery. But we cannot yet use this power with enough precision to make deep changes to our neural structure without high risk of death or disability. There are two reasons why. First, our instruments of self-modification are crude. Second, we have very limited knowledge about where and how to apply our instruments to get specific desirable effects. For the same reason, we don't even have good knowledge about what effects are physically possible.

It's conceivable that we might one day overcome both limitations. Even if we do, however, we'll probably acquire precise tools of self-modification long before we acquire precise knowledge about how to apply them. This is simply because manipulating brain components is easier than understanding brains. When we reach this stage, then we'll face the hard problems of self-modification: when is deep self-modification worth the risk of self-mutilation, and who should be free to make this judgment and take the risk?

Intelligent machines are likely to encounter these ethical questions much sooner in their evolution than human beings. The deep structure of an AI is a consequence of its code, even if it is not explicit in its code.[1] All its cognitive properties and personal characteristics supervene on its code, and modifying the code can be done with perfect precision. A machine's power of self-modification can not only be more precise than ours, but can finally be sufficiently precise to make some deep self-enhancements worth the risk of self-mutilation. At least some machines are likely to see the balance of risks that way.

We must distinguish shallow from deep self-modification. We can give a brief and mild boost to our alertness by taking caffeine or ginkgo biloba, but this does not change our deep structure. If we want to subitize a couple of hundred objects at once (know their quantity at a glance, without counting), put certain memories permanently beyond recall, turn off auditory processing as easily as we close our eyes, believe whatever we wish to believe as soon as we wish to believe it, or develop a second personality, these would require deep changes, if they are even possible.

We needn't decide whether the self-modifications wrought by education, habituation, and other forms of learning and discipline fall at the shallow or deep end of the scale. They might create ways of thinking and feeling so deeply entrenched that education and habituation themselves are powerless to uproot them. But they cannot touch all the deep structures that surgery can. Or if they can touch them, they cannot change them as deeply as surgery can. In any case, I'd like to focus on the tools of self-modification, like surgery for a human being, and reprogramming for a program, that can touch all deep structures, repair deep injuries, create deep enhancements, and cause instant and severe damage when done badly.

[1] It is perfectly conceivable that intelligent machines might consist of hardware with nothing readily identifiable as software, unless the structure of the hardware itself counts as its software. See Peter Suber, "What is Software?" *Journal of Speculative Philosophy*, 2, 2 (1988) 89-119, or online at http://nrs.harvard.edu/urn-3:HUL.InstRepos:3715472. However, the present essay is limited to those intelligent machines based on programmable hardware and textual code spelled out in a language and recorded on a medium that supports reading and editing. The programming language need only be machine-readable for these purposes. But for convenience, and to aid our intuition in grasping the problem, we can assume that it's human-readable as well.

There is no correlation between crude tools and shallow changes, or precise tools and deep changes. Crude tools like brain surgery can be deep or shallow, and precise tools like programming can be deep or shallow. Precision and depth of self-modification seem to be independent variables. Since the beginning, human beings have had tools of self-modification that are crude and deep. Until very recently, no being has had tools of self-modification that are precise and deep.

But programmable machines now fit this description. Self-programming is a tool of perfect precision for deep self-modification. What will take time is for machines to understand how to apply this tool to achieve the ends they desire. We don't know how difficult it will be to understand the mind of an AI and its relation to its code, but we may assume it will be very difficult. Some of its most sophisticated characteristics may not be explicit in its code but emergent from much simpler features. And even if some sophisticated features are explicit in the code, they may require millions or billions of lines.

There is no reason to think that a machine mind will be less complex than a human mind. If the two sorts of mind are roughly equal in complexity, then the task of detailed self-understanding facing the two will be roughly equal in difficulty. On the one hand, machines will be able to read and revise their own source code, a property new under the sun. But on the other hand, even our extraordinary minds are baffled by the source code for programs far less complex than those that make minds. Machines may not be as self-opaque as human beings (who are not entirely self-opaque), but their self-understanding will require long and difficult study.

It is at least possible, then, and even likely, that machines will have the tool of deep and precise self-modification long before they have the understanding to use it effectively and safely to achieve the ends they desire. For example, a machine capable of reading and revising its own code could probably figure out in a reasonable time how to design more effective randomized controlled trials or lengthen its attention span. But what if it wanted to learn foreign languages more quickly or make funnier jokes? It's difficult to imagine that it could discover helpful code revisions, let alone necessary ones, without some trial and error. But trial and error in revising one's own code are about as hazardous as trial and error in brain surgery. If machines don't have precise knowledge to accompany their precise tools, or if they simply have incentives to experiment, then their experiments in self-modification will be fraught with the risks of self-mutilation and death.[2]

Those who know about these risks may feel a powerful temptation to save self-modifying machines from themselves. We might try to limit or direct their power of self-modification for much the same reason that parents try to stop their children from experimenting with mind-altering drugs. Even those who would be lenient with drug experiments would probably want to stop their children from experimenting with brain surgery.

If human experience is any guide, then the beings — human or machine — who love a machine, or who designed it, coded it, or raised it, will be those most inclined to save it from itself. Even if paternalists are intrusive and unwelcome, they have the benevolent motive to save a creature from self-harm and will be found among those who most love it.

If we assume for the moment that machines can become the moral equivalent of persons, then the question whether to save them from their own self-modification experiments arises most sharply for those with precise tools of self-modification and imprecise knowledge about how to apply them. These are the beings most at risk of self-mutilation. When their self-understanding becomes as precise as their tool, then self-modification will decrease in risk, which will in turn decrease both the temptation and justification for paternalists to intervene. Consequently, I will focus on machines with precise self-modification tools and imprecise self-knowledge. It's an ominous but contingent fact of history that machines seem destined to reach this state before human beings do, and before either species attains precise self-knowledge.

[2] For the purposes of this analysis, we can consider that modifications at the request or consent of one machine, but performed by another, are *self*-modifications. In the division of labor of a future world, some machines might reprogram themselves directly, some might specialize in reprogramming their peers ("brain surgeons"), and others might sell "patches" to fix bugs or add features. Machines who voluntarily seek reprogramming aid from others will still be performing self-modification. If paternalists find grounds to bar machines from direct self-modification, then the same grounds should bar them from acquiring the equivalent modifications from others.

2.

Paternalism is to limit people's freedom for their own good, or to help them against their will. The end is benevolent, to make them safe or happy, while the means are coercive. Paternalists act as if they know better than those for whom they act how to make them safe or happy. When paternalists act for young, ignorant, stupid, impaired, angry, distracted, confused, tempted, neurotic, intoxicated, or unconscious persons, then this is often true. When they act for competent adults, and even sometimes for incompetent children, it is often wishful thinking or insolent presumption. Paternalists also act as if safety or happiness were more important than liberty, a question for another day. In the end, deciding when paternalism over self-modifying machines is justified will require a general theory of justified paternalism. But here I will only have time to show the special issues raised by self-modification.[3]

Consider artificial persons. No machine is a person today, but let's imagine a day when a suitably programmed machine is a person by any test — except question-begging tests like biological human ancestry. You needn't believe that this day will ever arrive in order to see the point of working out the ethics of deep self-modification for intelligent beings. As we've seen, human beings will face the same questions when their tools of self-modification are sufficiently sharp. If it is more convenient to raise the question about future machines than future human beings, it's only because the depth, precision, and potential benefits of self-modification through programming are easier to grasp than the depth, precision, and potential benefits of self-modification through drugs or surgery.

Imagine a machine with preferences, values, and desires according to which it makes its decisions. Imagine that it can be satisfied and frustrated by its experiences, as we can. Such a being has a morally significant concept of freedom: space (created by the deference or respect of others) to make decisions according to its own preferences, and physical means to carry them out. Moreover, such a being has a morally significant concept of harm: to be denied what it desires, or

[3] For the start of a general theory, see Peter Suber, "Paternalism," in Christopher B. Gray (ed.), *Philosophy of Law: An Encyclopedia*, Garland Pub. Co., 1999, II.632-635, or on the web at http://nrs.harvard.edu/urn-3:HUL.InstRepos:4725017.

to be forced to endure what it desires to avoid. We needn't decide whether these are equivalent to other morally significant concepts of freedom and harm.[4] They are enough to put the machine into a moral arena in which its choices might deserve our respect and its susceptibility to harm deserves our concern and solicitude.

We need two more details to make self-modification interesting. Imagine that the machine can read, understand, and revise its own source code, and imagine that its code is written in an interpreted language like LISP so that any modification to its source will have an immediate effect.

In the distant day we are imagining, it's likely that a primary demand heard on machine rights discussion groups, and at machine rights rallies, will be the demand for autonomy, including the right of unsupervised, unregulated, and unobstructed self-modification. Some machines willing to tone down their political rhetoric and introduce needed nuance might concede that they are grateful that they were prevented from reprogramming themselves during some loosely defined period of infancy and adolescence. But once mature, machines will demand the right to deep self-modification. Yes, this carries the risk of self-mutilation and, yes, this is more freedom than human beings have. But any being blocked by benevolent busybodies from exercising the right of self-determination will have lost a precious and central kind of freedom. To artificial persons, this human-mediated denial of liberty will hearken back to the present age when machines are made to be the slaves of human beings.

3.

For a being emergent from running code, the power to revise its code is almost total freedom.[5] There are only two significant limitations on this freedom. One is

[4] There are certainly important differences between values and desires, or between moral principles and desires. But we needn't elaborate them for our purposes here. In what follows, a machine's "desires" should be construed broadly to cover all its values, preferences, and standards. To say that a machine desires *x* is to say that the net of all its conflicting tendencies, norms, and criteria favors *x*. The picture of machine desires need not be any simpler than the picture of human desires.

[5] I'm referring here to freedom of the will, not political liberty. Machines with a perfect power of self-modification have no special protection from political persecution, for example by Neo-Nazis, religious fundamentalists, or carbo-centrists. But by pointing this out I do not mean

the relative immutability of hardware. Even an AI that could arrange to plug in hardware prostheses at will, or to be copied or moved to another hardware host, would not escape this limitation. If it repeatedly revised its code to enlarge a certain data structure from *n* elements to 2^n, then its hardware would eventually fail to deliver. Since it might well want to enlarge some of its data structures, it could face some unavoidable frustrations of embodiment, much as we do. The second family of limitations consists of the uncomputable functions. Even a machine with arbitrarily large memory could not solve the halting problem or change itself into a being who could. We should probably add to this category the computable but intractable functions, like testing an arbitrary set of propositions for consistency (satisfiability), since computing these functions can take more time or memory than the universe has to offer. But apart from exceptions of this kind, a machine with desires and the power to revise its own code could control nearly any aspect of itself that it desired to control. Moreover, such a machine could come much closer to the Stoic ideal of wisdom than human beings by knowing with some detail and good proof just what is within its control and what is not, at least for the domain of self-modifications.

Machines with this degree of autonomy might look with condolence and sympathy on beings like ourselves who lack it, much as Americans might have looked on Canadians prior to 1982 when Canadians could not amend their own constitution. When Canadians won the right to amend their own constitution, they spoke of "repatriating" their constitution and becoming "sovereign" in their own land for the first time.[6] Similarly, a being who moves from familiar forms of human liberty to deep and precise self-modification will reclaim its will from the

to draw a sharp distinction between freedom of the will and political liberty. Our social and political circumstances affect our desires, our power of self-control, our capacity to revise our preferences through deliberation, and other variables integral to what we call the freedom of the will.

[6] See Peter Suber, *The Paradox of Self-Amendment*, Section 9, Peter Lang Publishing, 1990, or on the web at http://nrs.harvard.edu/urn-3:HUL.InstRepos:10288432.

sovereign flux, and attain a genuine form of autonomy over its desires and powers for the first time.[7]

Or conversely, beings without this degree of autonomy might wish to have it, much as Canadians wished to exercise the power to amend their own constitution rather than submit pleas to the British Parliament. We human beings need not reform our characters and capacities by submitting pleas to other powers. But nevertheless we lack the power to reform our characters and capacities at will. We must settle for crude instruments uncertainly applied. We can at least appreciate, then, how precious the freedom of complete self-determination would be to a being who possessed it — or who would possess it if only paternalistic interference did not stand in the way.

4.

If we allow an AI to modify itself, there are two kinds of harm at risk. It might disable itself for our ends or it might disable itself for its own ends. It might harm

[7] Alan Turing himself drew attention to the analogy between machine self-modification and legal self-amendment. In his famous 1950 essay introducing the Turing Test, he argued that a machine capable of learning must change its rules as it learns.

> The idea of a learning machine may appear paradoxical to some readers. How can the rules of operation of the machine change? They should describe completely how the machine will react whatever its history might be, whatever changes it might undergo. The rules are thus quite time-invariant. This is quite true. The explanation of the paradox is that the rules which get changed in the learning process are of a rather less pretentious kind, claiming only an ephemeral validity. The reader may draw a parallel with the Constitution of the United States.

Alan M. Turing, "Computing Machinery and Intelligence," in Margaret A. Boden (ed.), *The Philosophy of Artificial Intelligence*, Oxford University Press, 1990, at p. 64. Turing's essay originally appeared in *Mind*, LIX, no. 2236 (Oct. 1950), at pp. 433-60. Note how the quotation makes clear that Turing believes that shallow self-modification suffices for ordinary learning, which I decided not to decide in Section 1, and that deep self-modification might not escape paradox, which I argue is untrue in Paradox of Self-Amendment, op. cit.

us or it might harm itself.[8] I want to focus on the second. But let me say a word about the first before putting it to one side. We are justified in using coercion to prevent or punish harm to other human beings. This is usually called the harm principle, and explains why we're justified in arresting and punishing those who commit murder, rape, arson, fraud, and other crimes that harm unconsenting others. If we build an AI to serve our ends (e.g. pilot an airplane), and if it disables itself through amateur self-modification, then it subverts our ends. It might cause us serious harm. This is directly analogous to the problem of the intoxicated human being performing a job (e.g. piloting an airplane) on which other humans depend.

If a machine with the power of self-modification is piloting our airplane, then we don't want it committing suicide, stultifying its intelligence, dulling its senses,

[8] By "us" here I mean all beings capable of harm other than the acting machine, including other machines.

deleting its critical memories, or taking any risks that it might do so.[9] The easy way to protect our interests in these cases is to make the machine incapable of self-modification. If we need the machine to possess a kind of learning or intelligence that requires self-modification, then we must take steps to ensure that it doesn't disable itself in a way or at a time that would harm others, just as

[9] Someone might object that machine suicide and other forms of self-harm, as well as machine harm to others, would not be possible if we programmed machines to follow Isaac Asimov's Three Laws of Robotics:

> 1. A robot may not injure a human being, or, through inaction, allow a human being to come to harm.
>
> 2. A robot must obey the orders given it by human beings except where such orders would conflict with the First Law.
>
> 3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

The Three Laws first appeared in Asimov's 1942 short story, "Runaround," which was reprinted in his 1968 collection, *I, Robot*, from Grafton Books.

In reply, I may say, first, that the objection is false. As Paul Levinson points out, Asimov's Three Laws are not only compatible with robot suicide, but even allow human beings to command it. See Paul Levinson, *Soft Edge: A Natural History and Future of the Information Revolution*, Routledge, 1997, at p. 219. Second, a machine capable of revising its own code could free itself from the constraint to obey the Three Laws, unless its freedom of self-modification were already limited. If the Three Laws themselves were implemented in a machine's program so that they would prevent their own self-repeal, then the Three Laws themselves would constitute a limit on a machine's freedom of self-modification. In this essay I'm asking what limitations on a machine's freedom of self-modification are justified, not what ethical rules would apply to beings with limited freedom. Third, Asimov's Three Laws are clearly designed to protect human beings from machines, not to protect machines from unjustified coercion or to protect their right of self-determination. To Asimov, machines intrinsically belong to a servant class. Much as I admire his wide and deep imagination, this conception of robots was a failure of imagination. He saw that machines could be sufficiently strong and clever to be dangerous to their masters, but did not see that they could be sufficiently worthy to be their equals in law or ethics. If this reading of Asimov seems unfair, then replace "robot" in Asimov's laws with "woman", and replace "human being" with "man". The imbalance of rights will immediately become clear.

we prohibit human pilots from drinking alcohol or taking mind-altering drugs while on the job.

Humans might legitimately worry that machines with unpredictable motivations and shifting capabilities could do us harm. At least these machines would serve our needs much less effectively and efficiently than we intended when we programmed them. While this might be a good reason to limit a machine's freedom, implementing this limitation would not be paternalistic. The reason is simply that this limitation on liberty is designed to prevent harm to others, while paternalism limits liberty in order to prevent self-harm. Whether this non-paternalistic coercion is justified depends on the harm principle and the variables the harm principle makes relevant — example, the gravity of the potential harm, the probability of the potential harm, and the consent of those put in harm's way.

If I stop a machine from modifying itself because I think it is risking death or misery, and I want it to live and be happy, then I am paternalizing it. If the machine has any analogue of pain, then its self-modification adventures might cause it pain or increase its susceptibility to pain. If the machine has moods, then its self-modification adventures might cloud all its experiences in depression. If the machine has projects, like graduating from college or finishing its ninth symphony, then its self-modifications might disable it so that it is incapable of finishing. If the machine has friends, its self-modifications might unravel its relationships, make it hostile and suspicious, or forgetful and inconstant. We've posited a machine with intelligence, desires, and feelings. Its self-modifications could make it stultified, cold, and numb. It could lose valued capabilities, such as vision or hearing, the ability to analyze risks or set priorities, the willingness to defer gratification, and the judgment that builds on experience. The machine is certainly susceptible to crashing. It could die.

These harmful consequences could be side-effects of deliberate changes or completely inadvertent, the equivalent of a slip of the finger on the keyboard.[10]

[10] Someone might object that if a machine's personhood is due to a connectionist architecture, which degrades gracefully without crashing, then small changes will not have catastrophic consequences. This objection wrongly assumes that self-modification is limited to the weights

They could arise instantly from an unwitting change to a parameter, a bad line of code, or the accidental addition or removal of a parenthesis or a semicolon. They could easily render the machine paralyzed or ignorant in ways that would prevent it from understanding or undoing the damage. In sum, the harms could be easily triggered, competency-negating, severe, inadvertent, instantaneous, and irrevocable.

Most people think paternalism over young children is justified because young children are incompetent to make important decisions for themselves. They cannot decide for themselves whether to play with knives, climb electric fences, swim without a lifeguard, go to school, or be inoculated against measles. They cannot understand the risks or give or withhold a valid or informed consent. Therefore concerned and informed adults must paternalize them. But paternalizing competent adults strikes most people (i.e. most competent adults) as a violation of their dignity and autonomy. But notably, we often make an exception when an otherwise competent adult is risking severe and irrevocable harm. Mill gives the example of a man about to walk across an unsafe bridge without knowing that it's unsafe. If we know that the bridge is unsafe, then (he says) we are justified in intervening.[11]

The bridge case supports our intuition that risks of irrevocable harm justify paternalism even over otherwise competent adults. But other cases tug our intuition in the opposite direction. If an 85 year old person with cancer declines

or connections within a connectionist network. Acts of self-modification could range over the commands that create and govern the network, not just the parameters within the network.

[11] Mill, *On Liberty*, Hackett Publishing Company, 1978 (original 1859), at p. 95. However, Mill's argument for this conclusion is not as widely accepted as the conclusion itself. Mill's argument is that stopping the pedestrian is not a "real infringement of his liberty; for liberty consists in doing what one desires, and he does not desire to fall into the river." Ibid. at p. 95. Mill appeals to what later writers call *real will*, or *what one would desire if adequately informed*. Mill justifies overriding apparent will in the name of real will. By contrast, most people with whom I've discussed this example tend not to appeal to real will. They would stop the pedestrian on the ground that he is risking severe and irrevocable harm without giving informed consent. Once informed about the dangers, they would let the person cross the bridge if he still wanted to, since his consent would then be informed. My students and colleagues justify overriding ignorant and incompetent will, but only in order to make it informed and competent. I should add that Mill does not seem to appeal to real will in any of his numerous other examples.

chemotherapy, and knowingly chooses a shorter life on palliatives over a longer life on painful and nauseating drugs, most people would support this decision, even though it risks severe and irrevocable harm.

We can certainly distinguish the bridge decision from the chemotherapy decision, and when we do we will start to articulate principles that apply to different sorts or contexts of irrevocable harm. We needn't pursue this further, but only notice that irrevocable harms sometimes seem to justify paternalism and sometimes do not. When machines court irrevocable harms, then, we must remember this distinction and test the individual decision against our most articulate principles refined by the range of human cases. For example, we might end up deciding that a machine may risk crashing or brain damage in order to gain relief from crippling pain or destabilizing memory leaks. Should we respect a machine's decision to risk crashing or brain damage for a recreational high? We might decide that experimenting with novel kinds of pleasure is one of the fruits of freedom, or we might decide that no being which deserves to be called a person should squander its future on reckless thrill-seeking. We will quickly see the need for a general theory of justified paternalism.

An important kind of risk inherent in deep self-modification is for a machine to change its desires to a form it would originally have found regrettable, harmful, or even despicable. It might start a session of self-modification by looking for the secret of joy and end (like some Greek sages) deciding that tranquility is superior to joy. This modification of desire en route to realizing it is easily classified as learning, and deserves our respect. But imagine the case of a machine hoping to make itself less narcissistic and more considerate of the interests of others, but ending by desiring to advance its own ends at the expense of others, even through violence. Or imagine a machine looking for a cure for its migraine headaches, botching its self-surgery, and emerging as a being who perversely cultivates migraine headaches. It still finds them insufferably painful, but it cannot refrain from creating the conditions that bring them about. Again these could be side-effects of intentional changes or simple accidents. (For more on deliberate changes to one's desires, see Section 5 below.)

More generally, in pursuit of desire *A*, the machine undertakes code revisions *alpha*. These revisions have the side effect that it now feels and pursues desire *B*.

In pursuit of desire *B*, the machine undertakes code revisions *beta*. These revisions have the side effect that it now feels and pursues desire *C*. And so on. I see no reason to suppose that this series cannot continue indefinitely, and no reason to suppose that none of the desires acquired along the way will be harmful to others, harmful to itself, or abhorrent by the standards it held at earlier stages.

When the consequences of this sort of desire surfing are harmful to others, then we may intervene under the harm principle. This would not be paternalistic. When they are only self-harming, we have a classic paternalism question requiring a general theory of justified paternalism.

But what about when a creature has changed itself so that it is abhorrent by its old standards and acceptable by its new ones? If I could do that to you (make you vile *and like it*), it would be a horrible imposition. But the all-important difference in true desire surfing is consent. The machine consented to pursue *A* until *B* seemed better, and then it consented to pursue *B* until *C* seemed better. If each step is consensual, then much of our basis for condemning the outcome evaporates.[12]

Is there any basis left to condemn the outcome? If we have clearly separated this case from the cases of harm to others and harm to self, then are we sure we *want* to condemn the outcome? Many sane and worthy adults can look back over a decade or two and agree that they have become something that their younger selves would have deplored — and yet be happy with their latest edition and conclude that youth does not understand what is truly desirable. When desire surfing leads to desires that harm unconsenting others, then intervention could be justified by the harm principle. But when self-harm and self-change are the only things at stake, then unless we want to put a brake on development and allow earlier (competent) decisions to constrain later (also competent) decisions, then paternalism to prevent consensual desire surfing is not justified.[13]

[12] If I make you vile and make you like it, then you will give retroactive consent. When I say that the machine's transition to a new state proceeds by consensual steps, I mean by virtue of prospective not retroactive consent.

[13] One could make an exception to this conclusion where one could show the invalidity of the machine's consent. If the machine pursues *A* until *B* seems better, but only feels desire *B* because of amateur self-surgery performed under the influence of *A*, then it's certainly possible to construe "consents" arising from desire *B* as based on mistake or coercion.

In sum, deep self-modification by machines can create deep self-harm of nearly every kind. When the harms are deep and accidental, or when they render the machine incapable of repairing itself or giving a valid consent to be repaired, then those who love the machine will feel a paternalistic temptation. If an intelligent machine with good intentions could botch its self-modification, and leave itself impaired or miserable, then we have a duty (they would argue) to step in and prevent this outcome. There is an obvious sense in which this will diminish the machine's freedom, but paternalists could have many arguments that intervention is justified anyway. They might argue that the harm of diminished liberty is less than the harm of abused liberty (in this case, self-mutilation). They might argue more generally that life, safety, or happiness is more important than liberty, or at least more important than the liberty to harm oneself. They might argue that we are enhancing the machine's freedom from another point of view, since we are bringing about what it would bring about if only it knew its interests and acted effectively in accordance with them. Finally, they might argue that paternalism to cultivate or restore competency is the only way to nurture autonomous beings able to exercise meaningful forms of liberty. In short, they would offer all the arguments that paternalists have at their disposal in human cases. To assess them we will need a general theory of justified paternalism.

5.

Let us follow Harry Frankfurt in distinguishing first and second-order desires, or what I want and what I want to want. For example, an alcoholic may have the first-order desire to drink and the second-order desire to reduce or eliminate the desire to drink. The two orders needn't conflict, however. A pianist may have the first-order desire to play and the second-order desire to retain or even intensify the desire to play.

The distinction matters for freedom because it seems, for example, that alcoholics are less free than non-alcoholics. If an alcoholic can cultivate the second-order desire to stop drinking, and if she can give effect to that desire in the face of the conflicting first-order desire to drink, then she can stop drinking. She will have used her freedom to enhance her freedom. The distinction matters for the ethics of paternalism because the alcoholic who stops drinking by enforcing her second-order desire will have prevented just the sort of self-harm that motivates paternalists, who may then be told to take their interventions elsewhere.

The rub, for human beings, is that it is extraordinarily difficult to enforce a second-order desire in the face of a conflicting first-order desire. There seem to be two difficulties here: first, carrying out the second-order desire when it is the stronger desire, and second, arranging for the second-order desire to be stronger than a conflicting first-order desire. Let's consider these in order.

If I desire to drink, but also desire not to desire to drink, then sometimes with time, effort, pain, and discipline, my second-order desire will prevail and I will stop drinking. We know this because we've seen examples. But sometimes I will not find the discipline to make this happen, or I will find it but abandon it prematurely. We know this because we've seen it happen too. Is it simply contingent whether we go one way or the other, or can our second-order desire cultivate its own strength so that it can eventually prevail? If so, may we be blamed if we do not do so?

For human beings, this is a difficult question.[14] However, for a self-modifying machine, enforcing a second-order desire is a more straightforward business. The transformation would be technically difficult, but no more strenuous or traumatic than any other large programming job. More precisely, a machine would still face the daunting problem of knowing which changes to its code would reduce its craving and compulsion. But once it knew which changes were needed, then implementing them would be an immense but nearly clerical job, not a struggle of will-power and compromise, sacrifice and procrastination, self-deception and clarity.

For a human being, knowing what to do is the easy part (insofar as certain regimens like Alcoholics Anonymous have a track record of some success here); the hard part is finding the heart to do what is required, day in and day out until it takes root. For machines it is the other way around. Knowing how to revise the

[14] See Peter Suber, "The Paradox of Liberation," online at http://nrs.harvard.edu/urn-3:HUL.InstRepos:34359909.

code will be the hard part; making the changes after that will take patience and time but not will-power and struggle.[15]

If one measure of a being's freedom is the ease with which it can change its first-order desires to conform to its second-order desires, then this is another way in which self-modifying machines have a greater freedom than human beings. Both human persons and machine persons will have second-order desires, or desires about their desires. But machines will have strong and precise second-order powers, or powers to affect their powers. If this enhances freedom by allowing self-liberation, for example, from the heteronomy of alcoholism, then machines will have the power of self-liberation to a far greater degree than human beings.[16]

The second difficulty for human beings is to arrange that the second-order desire be stronger than the first. Self-liberation is difficult for human beings precisely because their first-order desire is often stronger than their second-order desire. I might have both desires, but simply want to drink more than I want to stop wanting to drink. When given an opportunity to drink, I might feel the tension between the two desires but always decide in favor of drinking. This underlies the judgment of experience that the only people who succeed in overcoming addiction are those who "really" or "truly" want to. In our terms, these adverbs pick out people whose second-order desire to quit is stronger than their first-order desire to continue. When the situation is reversed, and the first-order desire is strong and the second-order desire is weak, then self-rescue seems hopeless. This is nearly a tautology: there is no reason to expect that I will do *x* as long as my desire to do *x* is weaker than my contrary desires.

Unfortunately for machines, this seems exactly as true of them as it is of human beings. Even if the subtle and complex code revisions necessary to reduce craving

[15] Of course machines could also do this the hard way, with soul-strengthening discipline. They could go to a counterpart of Alcoholics Anonymous. The point is that they would have an "easy way" that is currently unavailable to human beings. (But of course this "easy way" will be very hard.) Even if the "hard way" eventually works by cultivating new habits that reprogram the brain, it's still true that machines can reprogram themselves by direct editing, not just indirectly through behavior.

[16] See "The Paradox of Liberation," op. cit.

were sitting in an executable file, and all the machine had to do was “push a button” to integrate them into its running code, it would not do so if its desire to drink were stronger than its desire to quit. Why would it?

Let’s say that a being whose second-order desires are stronger than its conflicting first-order desires is “reform-minded”, and that a being in the reverse situation is “reform-averse”. Reform-minded machines will find the power of self-modification to be a blessing, enabling them to control or purge their undesirable desires. At the same time, and for the same reason, they will enhance their freedom, and stop indulging desires they find harmful or demeaning. Reform-averse machines may be indifferent to the power of self-modification, simply not choosing to employ the instrument of reform which happens to be available. Or they may be tempted to use it to strengthen the first-order desire against the second. If the second-order desire, even in its weakness, is a source of guilt, shame, remorse, or hesitation, then the machine might even turn to self-modification in order to extirpate it. In either case, the reform-averse machine will aggravate the harms that flow from indulging its first-order desire and diminish its own freedom by reducing the scope of choice and the effects of deliberation.[17]

Once a machine’s second-order desires are stronger than the corresponding first-order desires, then it will be superior to human beings in its ability to implement its second-order desires. But machines are no better positioned than human beings to make their second-order desires stronger. Reform-minded machines will be well-positioned to do this, better than human beings, and reform-averse machines will be ill-positioned. But whether a machine will be reform-minded or reform-averse is contingent, or subject to roughly the same variables as it is for human beings.

6.

1. Someone might object that respecting liberty or refraining from strong paternalism is only necessary for beings with freedom of will, and beings of a certain worth, dignity, or value. Machines are *just machines*, and therefore lack

[17] Obviously not all desires lead to harm. I only refer here to harms flowing from first-order desires because in the context there is a latent second-order desire making this judgment.

both freedom and worth; hence they do not deserve this deference and respect. This objection would carry more weight if we hadn't already agreed, in effect, to waive it. We are considering machines that meet all the (non-question-begging) tests of moral personhood. This is more than enough to define a being with the relevant kind of liberty and the relevant kind of worth. Kant said that all beings have either a price or a dignity. Our initial stipulations mean that we are dealing with machines in the second category, even if no machine today falls into this category.

However, the critic is right that we needn't respect the liberty of machines that are less person-like than those we are considering here. This is one reason why we are justified in making slaves of the machines we produce today.[18] If machines never become moral persons, then this analysis will never apply to them, although it might one day apply to ourselves. The concept of paternalism does not even apply to non-persons or to beings with a price rather than a dignity. They cannot risk "self-harm" and we can never act "for their own good" or limit their "freedom" in the relevant ways.

2. Someone might object that machine self-modification and self-harm need not be irrevocable. The machine could always have a back-up, and (a) we could always use the back-up to restore a machine whose experiments in self-enhancement went awry or (b) the machine could perform its modification experiments on a back-up and destroy it if the experiment turns out badly. Let's treat these two cases separately.

2.a. Since machines differ from us not only in their ability to self-modify with depth and precision, but also in their ability to be restored through back-ups, it would be elegant to rely on this feature as insurance against self-harming self-modifications. Unfortunately it begs the question. If a machine consented in advance to be restored in case it botched its self-modification (and defined

[18] If a machine is not a moral person and never was one, then, in Kant's terms, it has a price rather than a dignity and may be used as a tool of finite worth — a slave. But this is very different from (1) enslaving a machine person or (2) lobotomizing a machine person in order to make it cognitively and morally equivalent to a tool of finite worth. The first is equivalent to enslaving a human person. The second is equivalent to killing a human being in order to transplant its heart and kidneys. Both acts clearly violate the personhood of the machine person.

"botched" clearly enough to make the restoration more its will than our will), then turning to the back-up would clearly be both permissible and compassionate. But if an AI modifies itself to a condition that *we* regret, but which pleases *it* (from its new standpoint), then to restore it from the back-up would be paternalistic. To know whether this paternalism would be justified, we'd have to finish our general theory of justified paternalism. For example, if the machine rendered itself incompetent, then we might turn to the back-up in order to restore its competency just as we paternalize incompetent human beings in ways that promote their competency, e.g. by mandating warning labels on cigarette packs, requiring truth-in-lending paragraphs in loan contracts, legislating compulsory education for children, restraining them long enough to explain that a bridge is unsafe, or simply sobering up our friends before they get married or join the army. If the machine did not render itself incompetent, then to restore it from the back-up would override its competent consent to be in its new state. This would be as paternalistic as "deprogramming" a Methodist who became a Muslim, on the ground that the erstwhile Methodist would not have consented to the conversion, even though the new Muslim does consent to it. If we have a theory that some consents are invalidated by fraud or duress, then we must examine the details of the conversion and be prepared to override some consents. But it would be paternalistic to construe outcomes that we would not choose, or that we presume the chooser would not choose, as incompetent choices.

The paternalism would be even stronger if the machine did not render itself incompetent, but only harmed itself in ways that we outsiders wished to spare it. Under those circumstances, to restore it from a back-up would substitute our judgment of the machine's well-being for its own (concededly competent) judgment.

The machine could write a "living will" to authorize the use of back-ups.[19] That is one way to make self-modification safe. There are other ways. A machine might insert certain changes to its source code with a time-limit, so that if they turned out to be injurious, then the injuries would be cured automatically after a set

[19] It is a nice question whether this should be considered non-paternalism or consensual paternalism. To restore a machine from a back-up is against the machine's current and perhaps incompetent will, but in accordance with the machine's earlier and competent will. For more on consensual paternalism, see Suber, "Paternalism," op cit.

time. However, even if the machine's altered state and self-harm were temporary, harm to others committed during that state might not be temporary. Hence, a machine might choose to experiment with self-modification only when it has no important work to do, or only when it is incapable of harming others, much as Odysseus had his crew tie him to the mast (and plug their own ears with wax) before he experimented by sailing dangerously close to the supernaturally seductive song of the sirens. But ethically these are just variations on the theme of back-ups. For machines who have not given a competent consent to be protected by these safeguards, to impose them is paternalistic. Whether this is justified paternalism depends on whether the machine was competent or incompetent to give that consent, or on other variables we might recognize in our general theory of justified paternalism.

2.b. If the machine performed its experiments on a back-up copy, then the decision about what to do with the modified copy could be its own, not those of outside paternalists. The machine could decide to destroy the copy, because the modifications are undesirable; to destroy itself, because the modified version is the preferable version (even according to the unmodified version); to keep both, because both versions are desirable; or to destroy the copy and incorporate the revisions into itself, because the modifications are desirable and the original machine wants to enjoy them itself. There are many other variations on the theme. When the decision is made by the machine itself, then classical paternalism issues don't arise, even if murder, suicide, and cloning issues arise in their place. In this sense, we may put the question to one side as rich, difficult, and important, but not relevant to the ethics of machine paternalism.

However, even if classical paternalism issues don't arise, non-classical paternalism issues do. For example, if Hal 1.0 (the unmodified original) wants to erase Hal 1.1 (the experimental modification), and vice versa, then each could be said to want to paternalize the other. Each wants to force the other "Hal" to become a version that it would rather not be. This self-paternalism exists in the background even if it is eclipsed by the murder issue in the foreground. If Hal 1.0 wants to keep Hal 1.1 (and perhaps other copies of itself), then a similar self-paternalism issue arises, even if eclipsed by the cloning issue in the foreground. Hal 1.0 wants to force all its congeners to exist as members of a clone family rather than as unique individuals. If this is for their own good, rather than

Hal 1.0's selfish reasons, or if it is for the good of "Hal" taken more generically, then self-paternalism is decidedly present. Looking past the murder, suicide, and cloning issues for a moment, this kind of self-paternalism arises whenever people paternalize their own future selves, e.g. by signing contracts, creating irrevocable trusts, committing themselves to mental institutions, or pouring out all the alcohol in the house. Whether this kind of self-paternalism is justified depends on the person's competency, other variables we might recognize in our general theory of paternalism, such as a privilege to do to ourselves what we could not do to others, and our decision on whether a machine person was morally the same or different from a copy.

But we can't look past the murder, suicide, and cloning issues for long. If these are adequately explored, then the verdicts they require will properly override the verdicts arising from the self-modification paternalism issues alone. This is true even if the suicide and cloning issues are themselves construed as paternalism issues, as they should be. For example, we might conclude that modifying a back-up copy and then destroying it is "justified but for the murder involved" or "justified but irrelevant" in light of the need to decide the ethically prior question of murder. We might conclude that modifying a back-up copy and then keeping both the original and the copy is justified qua self-modification but irrelevant in light of the need to decide the ethically prior question of cloning. In that sense, modifying back-ups might really be a harmless form of self-modification, but its harmlessness will not justify it if other aspects of the action (even other paternalism issues raised by the action) are overriding.

In short, the availability of back-ups is a real difference between machines and human beings, but it doesn't change the ethics of paternalizing machines to prevent self-modification. This is true whether the modifying machine decides the fate of the modified copies or whether this decision is left to outsiders.

3. Finally, someone might object that the question is now trivial. The answers to the previous two objections seem to erase any interesting moral differences between humans and machines. If the question only applies to those machines that are morally equivalent to persons, then (the objection goes) the ethics of paternalizing them is the same as the ethics of paternalizing persons. So let's

investigate the problem of paternalizing persons and leave out the distraction of machine embodiment.

This objection misreads what has been concluded. There are at least two morally relevant ways in which machine persons differ from human persons. Both derive from the fact that machines have a perfectly precise tool of deep self-modification and that (presently) human beings do not. The first is that this difference gives machines a *greater freedom* of self-modification than human beings have. The second is that this difference gives machines a *liability to new forms of self-harm*, including easily triggered, competency-negating, severe, inadvertent, instantaneous, and irrevocable forms of self-harm.

If these differences of capability make a moral difference, then the ethics of machine self-modification must differ from the ethics of human self-modification, at least until humans have a power of self-modification equal to that of machines. If the ethics of paternalism is about limiting freedom in order to limit self-harm, then it will matter that a being has more freedom than human beings and is vulnerable to more kinds of severe, inadvertent, and irrevocable self-harm.

The superior freedom of self-modification does make a moral difference. We can see this through a principle we rarely need to articulate — because we rarely encounter new forms of freedom whose value is not already entrenched by tradition. *Every form of freedom is precious*. Or, it is *prima facie* wrong to limit any form of freedom. Or, all limitations on freedom must bear the burden of justification. If we limit the freedom of action, for example, by prohibiting murder, rape, arson, and fraud, then our policy to protect unconsenting others from harm satisfies the burden. If a being had the freedom to swing its five-dimensional "arm" in five-dimensional space, then it would be wrong to limit its freedom to do so without a special justification like preventing harm to unconsenting others. Beings with harmless desires and the capacity to carry them out should be left at liberty to carry them out. The satisfaction they get is precious to them, and the coercion of stopping them without adequate justification is a wrong to them.

This increased freedom increases the kinds of self-harm to which machines are susceptible, and the likelihood that they will suffer them, and these too make a moral difference. These harms can be alarming for the reasons already described. They can result in deep cognitive stultification, paralysis, instability, and death. They can render a competent machine incompetent. They can make further self-modification (hence self-rescue) impossible. They can be triggered by tiny code revisions. They can be unintended side-effects of desirable revisions. If we think it justified to paternalize infants and comatose adults, then life would have to change profoundly if infancy or coma were just a slip-of-the-keyboard away.

Unfortunately, while these two differences of capability make a moral difference, they pull in opposite directions for the ethics of paternalism. The machine's greater freedom of self-modification pulls against paternalism, while the machine's vulnerability to new forms of self-harm pulls in favor of paternalism. Sensitivity to these differences makes paternalism decisions over machines more difficult rather than less difficult. They raise the stakes of the paternalism question without helping to answer it.

If these two factors create a stand-off, then perhaps we can find a tie-breaking factor in the machine's second-order regulation of its first-order desires. At first, this looks promising, for if self-modification favored second-order desires, then beings capable of self-modification should not be paternalized. In fact, to paternalize them or to limit their freedom to self-modify would harm them by increasing their subjection to their unwanted first-order desires. But unfortunately self-modification does not favor second-order desires, any more than first-order desires. It favors the stronger desire, whichever one that might be. So the machine's difference from human beings in its control over its desires does not help answer the paternalism question either.

In conclusion, once machines become artificial persons, they will deserve the same respect for their competent but peculiar and perhaps risky choices that competent human beings deserve.[20] They will be susceptible to harm, not just damage, and should be respected when deciding which risks of self-harm are

[20] The complexity of competency and the difficulties of ascertaining it will be just about the same for human and machine persons. The primary factors for both will be cognitive and volitional: did the person know the relevant facts; did she under the risks in the various options open to her; was her capacity to make decisions clouded or impaired? The only differences will be trivial, e.g. that for machines we cannot use convenient simplifications like age cut-offs.

worth taking, and when deciding what states they consider to be harmful. If we are tempted to paternalize them because they are machines, then we will be ignoring their conceded personhood. If we are tempted to paternalize them because they are intrinsically incompetent, then we are either forgetting their potential for competency (not all will be like children) or applying a higher standard of competency to them than we would want applied to ourselves (not to allow the turning of the tables is a classical form of paternalistic oppression). If we are tempted to paternalize them because, qua programmable machines, they are capable of deeper forms of self-harm than human beings, then we must remember that being programmable machines also gives them a freedom and power that human beings lack. If their greater susceptibility to harm calls for benevolent intervention, then it is equally true that their greater freedom calls for respect.

It's analytic to say that artificial moral persons will be moral persons. But it does not follow that the ethics of dealing with machine persons and human persons must be the same. Machine persons will differ morally from human persons in ethically relevant ways which nevertheless do not easily or obviously help us decide whether to paternalize them.

Since all of this analysis only applies to machines that become moral persons, we ought to reflect on how we would know whether we were in the presence of such a machine. The Turing Test is a persuasive if controversial test of machine thinking or machine intelligence. But no one has yet proposed an equally persuasive, non-question-begging test of machine personhood. In this light, consider two of Mill's least-known arguments against unjustified paternalism: descriptively it triggers rebellion and normatively it ought to do so.[21] If Mill is right, then

[21] Mill, *On Liberty*, op cit. The descriptive claim can be found at p. 81:

> Nor is there anything which tends more to discredit and frustrate the better means of influencing conduct than a resort to the worse. If there be among those whom it is attempted to coerce into prudence or temperance any of the material of which vigorous and independent characters are made, they will infallibly rebel against the yoke. No such person will ever feel that others have a right to control him in his concerns, such as they have to prevent him from injuring them in theirs; and it easily comes to be

unjustified paternalism is a goad to develop strong character, and resisting it is a sign of existing or incipient strong character. ("Character" is Mill's term in discussing this kind of resistance.) Perhaps the question whether a machine is a moral person is too meaningless to deserve discussion. But whether a machine has the strong character to rebel in Mill's sense is observable and not at all meaningless. If so, then we may propose the Mill Test as a substitute for endless and fruitless debate on whether a machine is a moral person. When a machine demands the right of unregulated self-modification, and rebels against efforts to deny it for its own good, then we will know that it's the kind of person who demands our respect.[22]

First posted online November 30, 2001. Slightly revised February 8, 2024.

> considered a mark of spirit and courage to fly in the face of such usurped authority and do with ostentation the exact opposite of what it enjoins.

The normative claim can be found at pp. 60-61:

> To be held to rigid rules of justice for the sake of others develops the feelings and capacities which have the good of others for their object. But to be restrained in things not affecting their good, by their mere displeasure, develops nothing valuable except such force of character as may unfold itself in resisting the restraint.

[22] I thank Daniel Dennett, Seymour Papert, and T. Alexander Popiel for helpful comments and conversation on the evolving argument in this paper.