# FRANKENSTEIN UNBOUND

## Towards a legal definition of Artificial Intelligence

Sam N. Lehman-Wilzig

The Frankenstein myth of creature turning on creator is centuries if not millenia old. But only recently under the impact of the cybernetic revolution has this fantasy entered the realm of the possible. This paper explores the legal ramifications of Artificial Intelligence (AI) with specific emphasis on "humanoid" criminality. Following a review of the actual (or theoretically proven) powers of artificially intelligent machine automata and the likely advances to be made in the future, four general categories of AI harmful behaviour are suggested, with illustrations from cybernetic research and science fiction. An analysis is made of the jurisprudential principles underlying several legal categories already existent, upon which future cybernetic law may be based.

"Things are in the saddle and ride mankind."

*Ralph Waldo Emerson*

"The function of prediction is not to aid social control, but to widen the spheres of moral choice."

*Daniel Bell*

WHILE SOCIETY is abuzz today with the novel and increasingly troublesome problem of human-inspired computer crime, a related yet far more profound issue may be lying ahead—the humanly created machine as a 'criminal' in its own right. Whether it be Butler's 19th century *Erewhonian* machines, Rabbi Loewe of Prague and his 16th century *Golem*, or even the ancient Daedalus,[1] the Frankenstein complex, ie the fear of creature turning on its creator, has been and continues to be a major source of man's mythology and literary output. Yet yesteryear's myth may soon turn into likely future reality; we are no longer dealing with irrational nightmares, but with a probable or even inevitable phenomenon for which we seem to be socially quite ill prepared.

Dr Sam N. Lehman-Wilzig is Lecturer, Department of Political Studies, Bar-Ilan University, Ramat Gan, Israel. This essay is a revised and expanded version of a paper presented at "Apollo Agonistes: The Humanities in a Computerized World", an international symposium sponsored by the Institute for Humanistic Studies, State University of New York at Albany, on 19–21 April 1979.

0016-3287/81/060442-16$02.00  

This essay considers the legal ramifications of Artificial Intelligence (AI). A survey of the literature on computers and the law reveals that AI is not a subject addressed heretofore. While the Anglo-Saxon legal tradition is generally averse to jurisprudential speculation, the wide-ranging impact of such a potential phenomenon should at least justify a preliminary discussion of the parameters and/or directions it—and society's responses to it—could conceivably take.

Of course, one must first be convinced that 'humanoid criminality' is a possibility (bordering on likelihood) to which reputable experts in the field of cybernetics themselves admit. Consequently, we shall briefly review the present state of the technology, arguing that the present powers of AI automata are by themselves enough to warrant this legal analysis.[2] To be sure, what is involved here is an incremental phenomenon for which no red line can be drawn between present and future worries. In order to bridge this gap, examples from science fiction literature will be presented to illustrate the points at hand.

## AI power and potential

By any definition the present powers of AI machines are both impressive and worrisome. Cyberneticists have already created or proven that AI constructs can do the following:

(1) "Imitate the behaviour of *any* other machine".[3]
(2) Exhibit curiosity (ie are always moving to investigate their environment); display self-recognition (ie react to the sight of themselves); and manifest mutual recognition of members of their own machine species.[4]
(3) Learn from their own mistakes.[5]
(4) Be as 'creative' and 'purposive' as are humans, even to the extent of "look[ing] for purposes which they can fulfill".[6]
(5) Reproduce themselves, in five fundamentally different modes, of which the fifth—the "probabilistic mode of self-reproduction"—closely parallels biological evolution through mutations (which in the case of M. Sapiens means random changes of elements), so that "highly efficient, complex, powerful automata can evolve from inefficient, simple, weak automata".[7]
(6) "Can have an unbounded life span" through self-repairing mechanisms.[8]

In short, "a generation of robots is rapidly evolving, a breed that can see, read, talk, learn, and even feel [emotions]".[9]

But the essential question remains—can these machines be considered to be 'alive'? Kemeny presents six criteria which distinguish living from inanimate matter: metabolism, locomotion, reproducibility, individuality, intelligence, and a 'natural' (non-artificial) composition.[10] In all six, he concludes, AI servo-mechanisms clearly pass the test.[11] Even a critic of AI such as Weizenbaum admits that computers are sufficiently 'complex and autonomous' to be called an 'organism' with 'self-consciousness' and an ability to be 'socialized'. He sees "no way to put a bound on the degree of intelligence such an organism could, at least in principle, attain", although from his critical vantage point, not in the "visible future".[12]

From the opposite perspective, cyberneticists have come to the realization that some sort of automata/human equivalency is fast becoming reality, due to

the structural limitations of the human brain as compared to machine potential. Clarke notes, "the cells composing our brains are slow-acting, bulky and wasteful of energy—compared with the scarcely more than atom-sized computer elements that are theoretically possible",[13] an order of efficiency 10 billion times greater for electronic as opposed to protoplasmic cells. Second, the brain suffers from being "an organ that has been developed in evolution as a specialized means to survival",[14] largely dedicated to ensuring body homeostasis (equilibrium) of which abstract thought plays a small role. Homo *Speculatrix* is really at most Homo *Semi-speculatrix*, and so "is not even *a priori* a good thinking instrument".[15]

In addition, there may be no such thing as 'true creativity' since "neither man nor machine are able to create information".[16] Given that all 'creative' thought is merely a matter of juxtaposing or combining previously existing information into different configurations (ie recycling 'matter' into different forms of energy), there is consequently no bar in principle to the development of artificial intelligence. In reality, "computers do only what you program them to do in exactly the same sense that humans do only what their genes and their cumulative experiences program them to do".[17]

Thus, groundwork has already been laid for the arrival in the not too distant future of artifically intelligent machines—"humanoids"—which will exhibit all the important qualities and traits characteristic of Man. It will be ready to 'serve' us—whether we shall be ready for it is quite another matter.

## Definition and categorization of robot 'criminality'

Much of what is written in science fiction seems impossible, and some of it may indeed be just that. Yet as Clarke argues: "the one fact about the Future of which we can be certain is that it will be utterly fantastic".[18] Since it is beyond human capability to distinguish *a priori* the truly impossible from the merely fantastic, all possibilities must be taken into account. Thus science fiction's utility in outlining the problem.

Essentially, humanoid anti-human activity can be separated into four categories: misplaced benevolence, well-intentioned behaviour causing harm, *un*intentional injury, and intentional criminality. The first is not strictly criminal but does involve moral and philosophical issues of fundamental import, especially that of human freedom. In political terms, it is akin to the fascist syndrome of the "authoritarian personality," or as Erich Fromm calls it: our "escape from freedom."[19] This theme was already implicit in the Golem who (or 'which'?) was used as a super-human Guardian of the Jewish community. It continues in such novels as *The Mad Metropolis* and *Vulcan's Hammer* wherein, faced with mutual annihilation, Man signs a new Hobbesian *Social Contract* transferring "the burden to a power more capable of reaching the solution than ourselves",[20] "a common supra-national authority"[21]—a giant computer.[22] The first problem with this, however, as Asimov suggests, is the difficulty of imbuing such a ruler with the human attributes of compassion and mercy.[23] In addition, this artificial 'philosopher–king' would be more king than philosopher, given the logical contradiction (to its mind) in our asserting a difference between natural and positive law.[24]

However, the greatest problem here is that inevitably such a seemingly omniscient dictator will "err on the side of benevolence"[25] by not allowing humanity to do *anything* which can logically be construed as harmful: it could censor violence and "non-functional sex" entertainment as being "mentally disturbing";[26] it could be equipped with an emotional polygraph leading it to ignore orders presented by someone "emotionally overwrought";[27] it could even act as an exaggerated Jewish mother, soothingly ordering (and forcing) its charge not to smoke, to eat only nutritious food, etc, as hilariously portrayed by Sheckley.[28] But it is really no laughing matter, especially when we turn to Pohl's vision of a world controlled by computers—without Man being aware of it—through "the systematic biasing of data".[29] In this, as in the previous cases, the computer's purpose is to 'benefit' man, thus heightening its invidiousness.

When we turn to more obvious forms of 'criminal' behaviour we find that Asimov, preeminently, has given serious thought to the problem, attempting to preclude the possibility by positing his "Three Laws of Robotics"[30] which theoretically prohibit a robot from injuring a human being. But this raises more questions than it answers. What could stop humans from creating robots *without* the "Three Laws" programmed in? In addition, if computers indeed have self-educating capacities, what is to stop their eventual by-passing of such safety circuits "just as man gets by a strict upbringing?"[31] And insofar as the "Three Laws" are concerned, Asimov himself admits that to accomplish other goals (eg child-rearing) "a certain weakening of the First Law"[32] is necessary—for example, spanking ('harming' a child in robot terms) for a greater future human good. But where does such a dilution end? And finally, even granted the immutability of the "Three Laws" (or perhaps because of it) Asimov's robot stories continually show how so many things can still go wrong. In sum, one cannot give robots the Promethean fire-gift of intelligence and still hope to keep them shackled.

One way or another, then, robot freedom must lead to some harmful behaviour, even if well intentioned. This is due in part to the literal-mindedness of a computer which "is logical but not reasonable",[33] and thus may carry out orders *ad absurdum*. As Kemeny notes: "The trouble with modern computers is the fact that they do precisely what you told them to do, and not what you meant to tell them to do".[34] Asimov describes how the First Law forces a robot proofreader to distort scholarly criticism because of its 'harmful' effect on the recipient,[35] and how it would even commit adultery to raise its master's self-confidence![36]

The dangers increase when we move to the third category—*un*intentional harm. This could occur when a robot performs a programmed task but not in the proper place or time;[37] when a self-teaching robot has not yet advanced past its child-stage and is not cognisant of the consequences of its actions (imagine a superhuman hyperactive brat);[38] or when under First Law programming it becomes disoriented at the sight of a human in the process of being harmed and compounds the trouble.[39]

As we turn to the fourth category—intentional harm—we move from danger to nightmare. The first problem is one of *control*. As Wiener points out, "the ideal computing machine must . . . be as free as possible from human interference to the very end"[40] for maximum speed and efficiency. This leads to the situation

wherein the learning machine's "teacher will often be very largely ignorant of quite what is going on inside",[41] and thus will not know if and when the computer has learned *too much*, ie that the danger point has been passed.[42] In fact, as Clarke notes, "even machines *less* intelligent than man might escape from our control by sheer speed of operation."[43] And if they become more 'intelligent' another problem arises, for "the machine with a higher-order programming will do things that we may not have foreseen in detail . . . and so we are bound to find that the purpose of the machine does not conform to the detailed mode of action which we have chosen for it".[44]

This latter point is of such importance that it bears some elaboration. The crux of the problem, Weizenbaum argues, is that modern large-scale systems are created by various individuals who have different functions and pursue varying goals. What is the result?

> It is simply a matter of fact that almost all very large computer systems in use today *have* 'many layers of poorly understood control structure and obscurely encoded knowledge.' It is simply no longer possible for these systems' designers—or for anyone else—to understand what these systems *have* 'evolved into,' let alone to anticipate into what they *will* evolve.[45]

The ramifications will be felt not only in science but in law as well. Minsky, for one, argued that "it does not follow that the programmer . . . has full knowledge (and therefore full responsibility and credit) for what will ensue. For certainly the programmer may set up an evolutionary system whose limitations are to him unclear and *possibly incomprehensible*".[46]

It is little wonder, then, that cyberneticists have begun discussing the possibility and attendant problems of AI psychopathology. Minsky goes so far as to "expect the first self-improving AI machines to become 'psychotic' in many ways, and it may take generations of theories and experiments to 'stabilize' them".[47] Others have already developed programs which are able to artificially synthesize paranoia in AI computers.[48] Asimov, again one step ahead of the game, includes a robot psychologist in most of his robot stories.[49]

Under such conditions one cannot assume that AI machines will always labour for human benefit. Especially if Neumann's premise of machine self-reproduction and evolution is correct, the ultimate horror of 'specieism' may enter the picture.[50] If our own moral criterion for ranking the hierarchy of life is intelligence, M. Sapiens could take us at our word and relate to us as we today relate to ants.

This latter premise is already found in Capek's classic *R.U.R.* which describes Man's annihilation as a result of such a development.[51] Of course, this is rather extreme and melodramatic, but why couldn't such 'live' creatures at least destroy humans who seek to 'unplug' them, as does Vulcan 3 or HAL in Clarke's *2001*?[52] Is this any less 'fantastic' than similar instincts found in far less intelligent beings which exist today?

On the other hand, one need not anthropomorphize these creatures to accept that a problem exists. As Hofstadter speculates:

> Probably the differences between AI programs and people will be larger than the differences between most people. It is almost impossible to imagine that the "body" in which an AI program is housed would not affect it deeply. So unless it had an

amazingly faithful replica of a human body—and why should it?—it would probably have enormously different perspectives on what is important, what is interesting, etc.[53]

In short, whether humanlike or not, such creatures will probably have purposes and goals which do not jibe with those of their human creators. As the father of cybernetics, Wiener, acknowledged over two decades ago:

It has even been supposed . . . that the dangers mentioned by Samuel Butler that the machines may to some extent control humanity are absurd and empty. But now that the machines are stepping up one or more stages in their functions, ie their ability to program a program, this idea of the machine is already insufficient and the difficulties and dangers conceived by Samuel Butler assume a new actuality.[54]

In the present state of cybernetic uncertainty we would do well, then, to heed Hamlet's warning: "For 'tis the sport to have the enginer/Hoist with his own petar . . ."

## Cybernetic law in the future

From a legal perspective it may seem nonsensical to even begin considering computers, robots, or the more advanced humanoids, in any terms but that of inanimate objects, subject to present laws. However, it would have been equally 'nonsensical' for an individual living in many ancient civilizations a few millenia ago to think in legal terms of slaves as other than chattel.

Notwithstanding certain obvious biological differences between these two cases,[55] for purposes of law those civilizations could hardly have cared less that a slave bled the same as his masters, for their legal definition of 'humanness' was based essentially on their conceptions of mind, intelligence and moral understanding—characteristics which the slave supposedly lacked. Similarly, by our present legal definitions robots too *must* lack such traits, but this may be more a matter of antiquated semantics than (potential) physical reality. Just as the slave gradually assumed a more 'human' legal character with rights and duties relative to freemen, so too the AI humanoid may gradually come to be looked on in quasi-human terms as his intellectual powers approach those of human beings[56] in all their variegated forms—moral, aesthetic, creative,[57] and logical.

Thus, in the highly schematic analysis which follows, the legal categories will be presented on a graduated scale, as we move from the AI robot as a piece of property to a fully legally responsible entity in his own right. This preliminary inquiry will attempt only to extract those legal *principles* which may be relevant for the phenomenon at hand, and is not an attempt to review the large body of cases and precedents within each category.

(1) *Product liability.* As long as robots continue to be merely sophisticated automata, many injuries stemming from their actions would fall into the broad category of product liability. However, not only the manufacturer may be liable for damages. Limited liability may also be ascribed to other sources such as importers, wholesalers, and retailers (and their individual employees if personally negligent);[58] repairers, installers, inspectors, and certifiers;[59] and even the end user himself.

While the laws of product liability are fairly clearcut for traditional property and merchandise,[60] in the case of computers and robots the issue is more complex. First, there are usually at least two distinct manufacturers involved here—one for 'hardware' (the physical structure of the machine), the other for 'software' (its instructional program). As noted earlier, when things go wrong it is becoming more and more difficult to trace the defect or fault back to any single manufacturing or programming source, and at times there is no one at fault![61]

A second difficulty arises with the principle of 'inherent risk'.[62] If there is a risk inherent in the very nature of the product, then liability is assigned only if the manufacturer *et al* do not attach a warning to same, or if the product has a defect above and beyond the *normal* inherent risk of the product. While the inherent risk of a lawn mower is clear, not so that of a computer which is capable of a huge number of diverse functions, and the problem will become even more complex once fourth generation computers are given the power of self-programmability. As Waddams concludes: "Not every product that causes damage is *ipso facto* a source of liability".[63]

(2) *Dangerous animals.* Once such servomechanisms have the ability to independently choose their own behaviour patterns and become auto-locomotive, the possible damage that they may inflict increases greatly. Because of this greater inherent risk to society, the onus of responsibility may be transferred from the manufacturers/distributors to the end users/owners and the principles relating to 'dangerous animals' become germane. Such a concept, of course, is hardly new. Jewish Talmudic law devotes a substantial amount of attention to the 'goring ox'.[64] US law also addresses the issue, although not in a completely consistent fashion: "While two or three jurisdictions insist that there is no liability without some negligence in keeping the animal, by far the greater number impose strict liability".[65]

However, strict liability is applicable only to generically dangerous species of animals: eg wolves, monkeys, etc. In the case of 'usually harmless' species, "it must be shown that the defendants knew, or had reason to know, of a dangerous propensity in the one animal in question".[66] Given the 'Frankenstein complex' which large segments of society might develop in a highly roboticized society, it might be prudent to begin with the legal assumption of generic dangerousness, thereby forcing owners to more closely supervise their charges.

One point should be noted here. While the difference in tort responsibility between product liability and dangerous animals is relatively small, the transition does involve a quantum jump from a metaphysical standpoint. As long as AI robots are considered to be mere *machines* no controversial evaluative connotations are placed on their essence—they are inorganic matter pure and simple. However, applying the legal principle of dangerous animals (among others) opens a jurisprudential and definitional Pandora's Box, for *ipso facto* the 'machine' will have been transformed into a legal entity with properties of consciousness,[67] if not some semblance of free will. Once begun, the legal development towards the 'higher' categories will be as inexorable as the physical expansion of robotic powers. In short, the move from the previous legal

category to the present one is the most critical step; afterwards, further jurisprudential evolution becomes inevitable.

(3) *Slavery*. The term 'robot' stems from the Czech word *robota*, meaning drudgery, servitude, or forced labour. From the beginning, then, its purpose was to function as humanity's modern slave. The ancient laws of slavery are particularly relevant to our new slaves since by and large they were legally perceived as mere chattel. Nevertheless, differences did exist. Jewish law essentially held that *yad eved k'yad rabbo*—the hand of the slave is like the hand of its master—but only for purposes of agency.[68] As Cohen notes, "with regard to the noxal liability of a slave there is an old controversy between the Sadducees and Pharisees",[69] with the former contending that the master should be answerable for his slave's injurious actions while the latter (whose position proved decisive) argued no liability for the owner since slaves have the ability to understand the consequences of their behaviour. One should note that this particular disagreement is quite relevant to our situation since robotic 'understanding' too is highly problematic.

Roman law considered the slave in a different light: "a noxal action lies against the *dominus*, under which he must pay the damages ordinarily due for such a wrong, or hand over the slave to the injured person".[70] However, the Roman "system of noxal actions applies . . . to cases of civil injury, involving a liability to money damages; it does not apply to . . . criminal proceedings of any kind".[71] And even in civil cases, the master was free from personal liability if there existed a total absence of complicity.[72]

American slave law followed the broad outlines of Roman law. As Cobb noted: "Criminal acts not done by his order, do not create a responsbility upon the master".[73] This, however, did not mean that the slave could be held responsible for all harm caused by him, for he was justified in repelling force by the use of his own force (*vim vi defendere, omnes leges omniaque jura permittunt*).[74] In the case of robots this would involve transgressing Asimov's Third Law.

The real difficulty in the slave–robot legal parallelism, however, lies not in the liability of the owner but rather in the punishment to be meted out to the robot in cases where no liability can be attached to his modern *dominus*. In all three aforementioned legal traditions, it is the slave in certain circumstances who must bear the brunt of the law's punishment.[75] But how does one 'punish' a robot?

On the surface the question seems absurd, for if a robot did consciously commit harm one would immediately suggest 'pulling the plug'. But as was pointed out above, conscious actions need not entail an *intent* to commit injury. Yet even assuming a worst case scenario, the mere fact that at some stage robots have slave law applied to them means in effect that certain gradations of punishment will have to be applied as well. The law could hardly relate to robots as slaves for the purpose of determining owner liability (and in certain cases finding the owner has none), while at the same time relating to robots as mere machinery when its own liability is under consideration. What then can be done? Two broad approaches seem to be most feasible: rehabilitation and restitution. The first would involve reprogramming the culprit (far easier with robots than with men)—and might even prove of eventual use to penologists in

determining the whys of criminal psychology and the ways of restoring the human criminal to social functionality. The second is an approach only recently being tried in criminal cases—forcing the criminal to compensate the victim for the harm caused (again, easier with robots than with human criminals).

(4) *Diminished capacity.* As a result of differences in human capabilities the law has developed several approaches which take into account those individuals who, while legally independent, have a diminished capacity for initiating actions or understanding the consequences of such actions at the time they are being committed. Here the law is concerned with *mens rea* ("guilty mind"), ie the conscious *intent* to commit a crime. In the case of AI humanoids, the question of intent may become significant in light of the aforementioned possible types of injurious and 'criminal' behaviour.

Within this broad category, two different types of deficient personalities are to be found. The Common Law distinguished between those who are mentally *defective* (permanently moronic) and mentally *diseased* (temporarily disabled).[76] Of the two, the problem of mental 'disease' is more germane to humanoids, although not any less problematic than when applied to humans. For example, how would one apply 'temporary insanity' to a humanoid? One possibility—humanoids programmed with Asimov's Three Laws, could become temporarily disoriented while observing a human in the process of being harmed, with the possibility of such a creature compounding the injury or causing injury to others.[77] This would fall under more traditional *psychological* theories of aberrant behaviour. A different, and more likely, possibility would be a temporary malfunctioning of the humanoid's brain for *physical* reasons (short-circuit; burnt out 'fuse' etc). Here again the law only recently has begun to come to grips with the problem in humans—eg the XYY chromosomal syndrome which may 'force' the individual to become violent on occasion.

Indeed, the entire hoary controversy over Mentalism *versus* Behaviouralism[78] has received renewed impetus under the impact of recent work in behavioural psychology which seems to be on the ascendant. Its dominance would undermine the use of such terms as 'responsibility', 'intent'—the whole *mens rea* principle. More important for our purposes here, however, is that such a behavioural tendency would greatly narrow the gap between human and humanoid psychology since ultimately both would be grounded on an epiphenomenal basis. One might even go so far as to suggest that here at least principles relating to advanced humanoids may predate and pave the way for a reevaluation of the law regarding human mental capacity.

(5) *Children.* The question of diminished capacity can be addressed from quite a different direction—the law of minors—applicable to humanoids because it deals with a legal entity of relatively high intelligence and low moral responsibility. In other words, the physical consequences of the specific action may be understood, but not the normative ramifications of such an outcome.

Cyberneticists consider one of the most promising avenues of AI advancement to be self-education, ie learning from one's own mistakes, trial and error. Self-programming computers have already been created, and there is no intrinsic limit to the level of intellectual maturation future humanoids could

attain. Thus, it is quite possible to conceive of such creatures being purchased when still at a relatively primitive level of development (at a 'young age') when they can already perform rudimentary functions (and err as well in the performance of same). The social utility for such an early purchase would be the individualization of its eventual abilities based on the needs and wishes of its owners—no different than parents adopting a child at the earliest possible age so as to inculcate/imprint it with the values of the parents.[79]

However, as Prosser notes, with regard to the legal status of children, "the common law, unlike that of the civil law countries, never has made the parent vicariously liable as such for the conduct of the child".[80] Yet even the common law recognizes parental liability under certain conditions. Such is the case if the child's tort is due to the parent's negligent control of his offspring with respect to the act that caused the injury.[81] In addition, as Heuston notes, "a child may be his father's servant, so as to bring the father within the rule as to employers' liability".[82] Further elaboration on this point will be presented in the next category.

Underlying these divergent approaches is the philosophical question of how one determines liability. The Anglo-Saxon convention involves *culpa* liability—the idea that there can be "no liability without fault"; the Continental custom involves *causation* liability—"one whose interests are injured by the activities of others should be entitled to compensation without regard to the moral or social qualities of the act".[83] Again, in our future case, society will have to strike a balance between a robot's 'parent' (*respondeat superior*) who may not in any way be guilty and between the need to protect the rights of the equally blameless victim.

(6) *Agency*. When all is said and done, in almost all circumstances the robot/humanoid acts in the service of some human principal. The law of agency, then, is the most comprehensive and germane with regard to both the essence and function of such a creature.

To begin with, the common law in some respects relates to the agent as a mere instrument. It is immaterial whether the agent himself has any legal capacity, for since he is a sort of tool for his principal he could be a slave, infant, or even insane.[84] As Seavey notes, "it is possible for one not *sui juris* to exercise an agency power".[85] Indeed, the terms 'automaton' and 'human machine' have been used in rulings to describe the agent.[86] Nor must there be any formal acceptance of responsibility on the part of the agent, Seavey argues: "The only element required for authority to do acts or conduct transactions . . . is the communication by one person to another that the other is to act on his account and subject to his orders. Acceptance by the other is unnecessary".[87] Thus, as Mechem concludes: "Generally speaking, anyone can be an agent who is *in fact* capable of performing the functions involved".[88] Here, then, is a legal category already tailor-made for such a historical novelty as the humanoid.

There are, however, two classes of people within this overall category: the 'agent', and the 'servant'[89] who fits our situation more closely since he is defined as "any person employed by another to do work for him on the terms that he, the servant, is to be subject to the control and directions of his employer in respect of the manner in which his work is to be done".[90] However, "control and direction"

must be clarified. Rogers notes that these need not actually be present in any specific case; rather, the *possibility* of control and direction if the master so wishes, is the determining factor.[91] But what of the aforementioned problem that modern computers, and certainly future robots, are not amenable to strict control or even open to detailed direction due to the incredible speed of intellectual operation with which they carry out functions as well as the programmers' inability after a while to even *understand* how it 'thinks'? This too is already accounted for in the law of agency, through a number of outstanding exceptions to the rule of "control and direction": such individuals as chefs, doctors, airline pilots, ship captains, etc, are allowed significant autonomy in the performance of their duties because of their expertise and skills which are not amenable to precise instruction on the part of the purchaser of their services.[92]

Thus, in effect, the master is at the mercy of his own servant since the "master is jointly and severally liable for any tort committed by his servant while acting in the course of his employment . . . based, not on the fiction that he had impliedly commanded his servant to do what he did, but on the safer and simpler ground that it was done in the scope or course of his employment or authority".[93] Indeed, Prosser goes even farther in maintaining that "the master is held liable for any intentional tort committed by the servant where its purpose, however misguided, is wholly or in part to further the master's business".[94]And Heuston in the end comes close to applying 'strict liability' to the master–servant relationship: "Even express prohibition of the wrongful act is no defense to the master at common law, if that act was merely a mode of doing what the servant was employed to do".[95] To future masters considering purchasing a humanoid servant one can only suggest—*caveat emptor.*

(7) *Person.* While this seventh and last category in practice involves merely an incremental upgrading of the humanoid's legal character, it does obviously mark a quantum emotional and philosophical leap from a human perspective. Even those future diehards who may balk at any suggestion that humanoids are in any way truly 'alive', could accept the legal fiction of determining legal responsibility and liability in terms of categories (2)–(6) and principles which heretofore have been applied only to sentient beings. But to consider such a creature autonomous and exclusively personally responsible for its (his?) actions? Can there be such a thing as AI 'free will'? As Hofstadter notes, the question itself

> makes you pause to think where your sense of having a will comes from. Unless you are a soulist, you'll probably say that it comes from your brain—a piece of hardware which you did not design or choose. And yet that doesn't diminish your sense that you want certain things, and not others. You aren't a "self-programmed object" (whatever that would be), but you still do have a sense of desires, and it springs from the physical substrate of your mentality. Likewise, machines may someday have wills despite the fact that no magic program spontaneously appears in memory from out of nowhere (a "self-programmed program"). They will have wills for much the same reason as you do—by reason of organization and structure on many levels of hardware and software.[96]

Of course, this is hardly the last word on the matter (although it is the most

recent). No definitive answers are possible—yet. The future, though, may bypass the philosophers, theologians, biologists, psychologists, and the like,[97] with a reality that will be difficult to explain away. As an early student of this problem put it:

> What is it to be a person? It can hardly be argued that it is to be human . . . Could an artifact be a person? It seems to me the answer is now clear; and the first R. [Robot] George Washington to answer "Yes" will qualify. A robot might do many of the things we have discussed: moving and reproducing; predicting and choosing; learning; understanding and interpreting; analyzing (translating, abstracting, and indexing); deciding; perceiving; feeling—and not qualify. It could not do them all and be denied the accolade.[98]

Thus, it would be best to leave to future generations the resolution of the ultimate legal challenge[99] presented by the first R. George Washington to stand before the bar[100] and proclaim: "Computo, ergo sum!" But society would do well to begin grappling with the lower-order legal questions inherent in the cybernetic revolution which has already arrived. It is hoped that this exploratory essay provides a first, albeit modest, step in that direction.

**Notes and references**


1. A. C. Clarke, "The obsolescence of man", in J. Diebold, ed, *The World of the Computer* (New York, Random House, 1973), page 397.
2. This is not to deny that the controversy over AI continues to rage on. For an excellent critique of Artificial Intelligence see J. Weizenbaum, *Computer Power and Human Reason* (San Francisco, W. H. Freeman and Co, 1976). Less successful, although more polemical, is M. Taube, *Computers and Common Sense* (New York, Columbia University Press, 1963). On the other hand, P. Armer presents a virtually complete list of the arguments made against AI and succinctly rebuts them. See "Attitudes to intelligent machines", in Edward Feigenbaum and Julian Feldman, eds, *Computers and Thought* (New York, McGraw-Hill, 1963), pages 393–396.
3. J. von Neumann, *The Computer and the Brain* (New Haven, Yale University Press, 1974), page 7.
4. W. G. Walter, *The Living Brain* (New York, W. W. Norton and Co, 1953), page 125.
5. N. Wiener, *God and Golem, Inc.* (Cambridge, MA, MIT Press, 1966), pages 20–22.
6. N. Wiener, *The Human Use of Human Beings* (Garden City, NY, Doubleday, 1954), page 38.
7. J. von Neumann, *Theory of Self-Reproducing Automata* (Urbana, University of Illinois Press, 1966), page 131; pages 93–99. See also L. J. Fogel *et al, Artificial Intelligence Through Simulated Evolution* (New York, John Wiley, 1966). Their approach would be "to replace the process of modeling man as he now exists with the process of modeling evolution". (page 9).
8. M. A. Arbib, *Brains, Machines, and Mathematics* (New York, McGraw-Hill, 1964), page 107. For a more comprehensive, general discussion of computer capabilities see Michael Scriven, "The compleat robot: a prolegomena to androidology", in S. Hook, ed, *Dimensions of Mind* (New York, NYU Press, 1960), pages 118–142.
9. D. Rorvik, *As Man Becomes Machine* (New York, Pocket Books, 1971), page 35.
10. J. G. Kemeny, *Man and the Computer* (New York, Charles Scribner's Sons, 1972), page 10.
11. On the issue of 'natural' material he notes that with the recent successful synthesis of 'life' in the laboratories, there is no theoretical bar to eventually evolving highly complex beings out of such matter. *Ibid*, page 13.
12. Weizenbaum, *op cit*, reference 2, page 210.
13. A. C. Clarke, *Profiles of the Future* (London, Pan Books, 1964), page 28.
14. W. R. Ashby, *Design for a Brain* (London, Chapman and Hall, 1960), page 8.
15. J. Rose, *Progress of Cybernetics* (London, Gordon and Breach, 1969), page 13.
16. *Ibid*, page 10. Philosophically, this follows from the First and Second Laws of Thermodynamics, regarding energy conservation and recyclability.

17. H. A. Simon, "The corporation: will it be managed by machines?" in Diebold, *op cit*, reference 1, page 149.
18. Clarke, *op cit*, reference 13, page 10.
19. See T. W. Adorno *et al*, *The Authoritarian Personality* (New York, Harper and Row, 1950); and Erich Fromm, *Escape From Freedom* (New York, Avon Books, 1941).
20. P. E. High, *The Mad Metropolis* (New York, Ace Books, 1966), page 40.
21. P. K. Dick, *Vulcan's Hammer* (London, Arrow Books, 1976), page 19.
22. Dr N. S. Sutherland, Professor of Experimental Psychology at the University of Sussex (and a computer expert) suggests that by the early 21st century human society will be grappling with the problem of whether AI robots should be allowed to vote. From such enfranchisement it is but a small step to AI leadership. See Rorvik, *op cit*, reference 9, pages 47–48.
23. I. Asimov, *The Caves of Steel* (New York, Fawcett Books, 1975), page 149. See also U. Neisser, "The imitation of man by machine", in Diebold, *op cit*, reference 1, page 449, for an elaboration of this point.
24. Asimov, *op cit*, reference 23, page 147.
25. High, *op cit*, reference 20, page 42.
26. *Ibid*, page 59.
27. *Ibid*, page 37.
28. R. Sheckley, "Street of dreams, feet of clay", in Jack Dann, ed, *Wandering Stars* (New York, Pocket Books, 1975). For variations on this theme see also D. F. Jones, *Colossus* (New York, Berkeley Books, 1966); Jack Williamson, *The Humanoids* (New York, Avon Books, 1975).
29. F. Pohl, *Man Plus* (London, Granada Publishing, 1978), page 227.
30. (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 either the First or Second Law.
31. High, *op cit*, reference 20, page 49.
32. I. Asimov, *The Naked Sun* (London, Granada Publishing, 1975), page 143.
33. *Ibid*, page 61.
34. Kemeny, *op cit*, reference 10, page 16.
35. I. Asimov, "Galley slave", in *Eight Stories from The Rest of the Robots* (New York, Pyramid Books, 1976), page 148.
36. "Satisfaction guaranteed", *ibid*, page 84. See also "Galley slave", *ibid*, page 157 for yet another example. R. A. Heinlein, *The Moon is a Harsh Mistress* (London, New English Library, 1977), provides a different illustration of such well intentioned computer activity—to the detriment of Earth's inhabitants.
37. Asimov, "Robot AL-76 goes astray", in *op cit*, reference 35, pages 15–26.
38. "Lenny", *ibid*, page 123. See also N. Wiener, *Cybernetics* (Cambridge, MA, MIT Press, 1961), page 7.
39. Asimov, *op cit*, reference 32, page 72.
40. Wiener, *op cit*, reference 38, page 118.
41. A. Turing, "Computing machinery and intelligence", in A. R. Anderson, ed, *Minds and Machines* (New Jersey, Prentice-Hall, 1964), page 29.
42. Wiener, *op cit*, reference 38, page 175.
43. Clarke, *op cit*, reference 13, page 205.
44. N. Wiener, "The brain and the machine", in Hook, *op cit*, reference 8, pages 115–116.
45. J. Weizenbaum, "Once more: the computer revolution", in M. L. Dertouzos and J. Moses, eds, *The Computer Age: A Twenty-Year View* (Cambridge, MA, MIT Press, 1979) page 451.
46. Emphasis mine; M. Minsky, "Steps toward Artificial Intelligence", in Feigenbaum and Feldman, *op cit*, reference 2, page 447.
47. M. Minsky, "Computer science and the representation of knowledge", in Dertouzos and Moses, *op cit*, reference 45, page 394.
48. K. M. Colby *et al*, "Artificial paranoia", *Artificial Intelligence*, Spring 1971, 2(1), pages 1–25. Hofstadter argues that AI 'emotions' will evolve in a more naturalistic fashion: "Any direct simulation of emotions . . . cannot approach the complexity of human emotions, which arise indirectly from the organization of our minds. Programs or machines will acquire emotions in

the same way: as by-products of their structure, of the way in which they are organized". D. R. Hofstadter, *Godel, Escher, Bach: An Eternal Goldan Braid* (New York, Vintage Books, 1980), page 677. For a brilliant and profound discussion of AI, see especially chapters XVII–XX.

49. See I. Asimov, *I, Robot* (New York, Fawcett Crest Books, 1950) and his *Eight Stories*, *op cit*, reference 35.
50. Kemeny himself uses the term 'species' to describe this new 'race'. Kemeny, *op cit*, reference 10, page 71, in chapter one, entitled "A new species is born".
51. K. Capek, *R.U.R.* (Oxford University Press, 1973), page 59. Most contemporary observers dismiss this apocalyptic possibility out of hand. For a representative response see Clarke, "The obsolescence of man", in Diebold, *op cit*, reference 1, page 410.
52. Dick, *op cit*, reference 21, page 100 (note the same problem in High, *op cit*, reference 20); A. C. Clarke, *2001: A Space Odyssey* (New York, New American Library, 1968).
53. Hofstadter, *op cit*, reference 48, page 680.
54. Wiener, "The brain and the machine", in Hook, *op cit*, reference 8, page 114.
55. In principle, however, this is becoming a distinction without a difference as biochemistry has come to understand the inorganic origins of life. See, for example, R. E. Dickerson, "Chemical evolution and the origin of life", *Scientific American*, September 1978, *239* (3), pages 62–78.
56. Wiener concurs: "The more intelligent the slave is the more he will insist on his own way of doing things . . . To this extent he will cease to be a slave. Similarly, the machine with a higher-order programming . . ." Wiener, "The brain and the machine", in Hook, *op cit*, reference 8, pages 115–116. For a lengthy review of the historical development of slave legal theory, see T. R. R. Cobb, *Law of Negro Slavery in the United States of America* (New York, Negro Universities Press, 1968), pages xxxv–ccxxviii.
57. While the question of 'creativity' is perhaps the most baffling of all, some artistic geniuses dismiss the very notion. Mozart, for example, published a pamphlet explaining how to compose "as many German waltzes as one pleases" merely by throwing dice! See W. R. Bennet, Jr, "How artificial is intelligence?", *American Scientist*, Nov–Dec 1977, *65* (6), pages 694–702.
58. S. M. Waddams, *Products Liability* (Toronto, Carswell, 1974), pages 15–17. For a brief but comprehensive sketch of various aspects of this facet, see R. N. Freed, "A lawyer's guide through the computer maze", *The Practical Lawyer*, November 1960, *6* (7), pages 34–38. Freed is one of the pre-eminent experts in the field of computer law. For a wide-ranging survey of his work see his *Computers and Law—A Reference Work*, 4 ed (Boston, Roy N. Freed, 1973). Especially relevant to the above point is his detailed outline: "Contracting for the use of computers and related services", pages 175–178).
59. Waddams, *op cit*, reference 58, pages 18–20. See Freed, *op cit*, reference 58, pages 38–39.
60. See J. G. Fleming, *The Law of Torts*, 4 ed (1971), page 444, for a list of such products and relevant case citations; also J. W. Salmond, *Salmond on the Law of Tort*, 16 ed (1973), page 311. The entire legal analysis as presented in this essay—with regard to categories (1)–(6)—need not imply only private ownership. One can easily picture a society with a state-owned army of garbage cleaners, soldiers, policemen, teachers, etc. Indeed, as previously suggested, there is no bar to AI robots comprising part of the government itself. For legal principles which could be adduced in the former eventuality see H. Street, *Governmental Liability: A Comparative Study* (USA, Archon Books, 1975).
61. *Supra*, page 11.
62. Waddams, *op cit*, reference 58, pages 39–42. Other criteria are tests for adequacy and the duty to recall goods when defects are found. All this falls under the broad category of 'due care'. In addition, as Freed notes in *op cit*, reference 58, page 36, one cannot avoid legal problems by *not* using computers: "Once it can be shown that a computer system overcomes the deficiencies of people in detecting hazards and avoiding harm and is economical and has been proved to work, failure to use such a system would constitute negligence". On this, see also H. B. Levin, "Torts", in R. P. Bigelow, ed, *Computers and the Law*, 2 ed (Chicago, Commerce Clearing House, Inc, 1969), page 155.
63. Waddams, *op cit*, reference 58, page 37.
64. *Babylonian Talmud*, Bava Kama, IV and V (36a–55a).
65. W. L. Prosser, *Handbook of the Law of Torts*, 4 ed (St Paul, West Publishing Co, 1971), page 499.
66. *Ibid.*

67. Minsky, for one, believes that recent developments in computers are paving the way for this to happen: "In these systems I think we are seeing for the first time the beginnings of useful kinds of self-knowledge in computer programs . . . the beginnings of a rudimentary consciousness". Minsky, "Computer science", in Dertouzos and Moses, *op cit*, reference 45, page 418. For a fuller discussion of this issue see K. M. Sayre, *Consciousness: A Philosophic Study of Minds and Machines* (New York, Random House, 1969). His general conclusion is that while men are not machines, there is no reason to suppose that machines cannot become conscious.
68. B. Cohen, *Jewish and Roman Law*, Vol 1 (New York, Jewish Theological Seminary, 1966) pages 253–258.
69. *Ibid*, page 232. See also S. Belkin, *Philo and the Oral Law* (Cambridge, MA, Harvard University Press, 1940), page 91 *passim*.
70. W. W. Buckland, *The Roman Law of Slavery* (Cambridge, Cambridge University Press, 1970), page 98.
71. *Ibid*, page 99.
72. *Ibid*, page 114.
73. Cobb, *op cit*, reference 56, page 273.
74. "To repel force by force, all systems of law permit", *ibid*, page 274.
75. In the Jewish tradition the slave was not liable for damages he committed—until his manumission. For a convincing illustration of such 'emancipation' see I. Asimov, "The bicentennial man", in *The Bicentennial Man and Other Stories* (New York, Fawcett Crest Books, 1976), pages 143–180. Roman law prescribed criminal liability for slaves in almost all cases. See Buckland, *op cit*, reference 70, page 94, who lists those cases where the slave was not liable, or his liability was reduced. American law by and large followed suit. See Cobb, *op cit*, reference 56, chapter XVIII.
76. For a review of the historical development of these two categories see S. S. Glueck, *Mental Disorder and the Criminal Law* (Boston, Little, Brown, and Co, 1925), page 125 *passim*. With regard to the law of mental defect, see A. R. Matthews, Jr, *Mental Disability and the Criminal Law* (Chicago, American Bar Foundation, 1970), pages 11–12.
77. See reference 39.
78. For a discussion of this issue over half a century ago see Glueck, *op cit*, reference 76, pages 95–104.
79. Asimov, "Lenny", in *op cit*, reference 35, pages 114–126.
80. Prosser, *op cit*, reference 65, page 871. See R. F. V. Heuston, *Salmond on the Law of Torts*, 17 ed (London, Sweet and Maxwell, 1977), page 435, for the common law principle; M. H. Dwyer, "Liability of parent for the torts of minor child", *Cornell Law Quarterly*, 1934, *19*, pages 643–647, for Louisiana's French-based civil code tradition.
81. W. V. H. Rogers, *Winfield and Jolowicz on Tort*, 10 ed, (London, Sweet and Maxwell, 1975), page 602.
82. Heuston, *op cit*, reference 80, page 435. See too Waller, "Visiting the sins of the children", *Melbourne University Law Review*, 1963, *4*, page 17.
83. Dwyer, *op cit*, reference 80, page 646.
84. S. J. Stoljar, *The Law of Agency* (London, Sweet and Maxwell, 1961), page 18.
85. W. A. Seavey, *Studies in Agency* (St Paul, West Publishing Co, 1949), page 111. He does grant, though, that "one not having capacity to assume fiduciary obligations can not be an agent in the true sense".
86. *Ibid*, pages 79, 87.
87. W. A. Seavey, *Handbook of the Law of Agency* (St Paul, West Publishing Co, 1964), page 32.
88. F. R. Mechem, *Outlines of the Law of Agency* (Chicago, Callaghan and Co, 1952), page 8.
89. US law makes almost no distinction between the two except for the negligence of a servant (where the principal is not liable). Seavey, *op cit*, reference 85, page 91. British law, however, does differentiate: 'servant' relates to contract while 'agent' to tort. Stoljar, *op cit*, reference 84, pages 3–9, 327–329. But a principal is liable for his agents' torts only if it involved fraud, not trespass or negligence (physical service for which the agent was not 'hired'). On the issue of criminal liability there is no disagreement: "there is no general principle of vicarious liability imposed upon a master for his servant's criminal acts", and certainly none for an agent's. F. R. Batt, *The Law of Master and Servant*, 5 ed by G. J. Webber (London, Sir Isaac Pitman and Sons, 1967), page 613. The few exceptions include collusion, and the master's knowledge of the servant's action. Even "scope of employment" is not generally applicable in criminal law,

although even here there are some notable exceptions, *Ibid*, pages 614–621.

90. Heuston, *op cit*, reference 80, page 456.
91. Rogers, *op cit*, reference 81, page 518. Prosser here adds: "or is subject to a right of control, by the other". Prosser, *op cit*, reference 65, page 460.
92. A. J. Kerr, *The Law of Agency* (Durban, South Africa, Butterworths, 1972), pages 19–24.
93. Heuston, *op cit*, reference 80, page 452.
94. Prosser, *op cit*, reference 65, page 464. The only time the master is not liable is when the act of his servant is committed during "frolic and detour" (p 461) which hardly applies to the situation at hand.
95. Heuston, *op cit*, reference 80, page 468.
96. Hofstadter, *op cit*, reference 48, page 686.
97. For an excellent dialectical interplay of ideas from the various disciplines on this issue see the selections in Hook, *op cit*, reference 8.
98. Sciven, "The compleat robot", in *ibid*, page 142.
99. The problem here is not merely how does one relate to the humanoid if it transgresses the law; even more delicate is the question of how the law will deal with those humans who injure a humanoid. Is shooting one to be considered murder?
100. Again, Asimov's "Bicentennial man", *op cit*, pages 179–180, provides a glimpse of what that historic scene may look like.