

Nano-enabled AI: Some Philosophical Issues

J. Storrs Hall

16th May 2006

Abstract

Improvements in computational hardware enabled by nanotechnology promise a dual revolution in coming decades: machines which are both more intelligent and more numerous than human beings. This possibility raises substantial concern over the moral nature of such intelligent machines. An analysis of the prospects involves at least two key philosophical issues. First is intentionality in formal systems: can “mere machine” truly embody mind with meaning and understanding, to say nothing of sensation or emotion? Secondly, what is the moral nature of a machine vis-a-vis a human: can a machine be true moral agent? If so, might a machine be a better moral agent than a human?

1 Background

On April 13, 2029, asteroid Apophis (2004 MN4) will pass within 22,000 miles of the Earth. Our current knowledge of Apophis’ position—we can place it with confidence within a region of space approximately the size of the Earth itself—is enough to be reasonably certain that it will not strike the Earth at that time. However, within that envelope of uncertainty, there is a window about the size of a city block, called a resonance keyhole. If Apophis is actually in the keyhole, the flyby with Earth in 2029 will alter its orbit such that it will return to strike the Earth in 2036, with an energy of 870 megatons.¹

Other events of 2029 cannot be predicted with such precision. However, we can predict with a fair confidence that two significant watersheds will have been passed in technological development: a molecular manufacturing nanotechnology which can produce a wide variety of mechanisms with atomic precision; and artificial intelligence. Detailed arguments for these predictions have been given elsewhere and need not be repeated here (Kurzweil [2005], Moravec [1997], Hall [2005]). We are concerned instead with a joint implication: if both of

¹Astronomy 34#5 May 2006 pp. 46-51

these technologies are present, greater-than-human intelligence will not only exist, but will be ubiquitous.

The net present value of an intelligent, educated human being can be estimated at a million dollars. We will refer to three estimates of human-equivalent processing power (hereinafter HEPP): Kurzweil at 10^{17} IPS, Moravec at 10^{14} IPS, and Minsky² at 10^{11} IPS. The author's own estimate agrees with Moravec's. Along the Moore's Law trend curve, the cost and value of a Minsky HEPP crossed in the 1990's, of a Moravec HEPP this decade, and of a Kurzweil HEPP in the 2010's.

Nanotechnology can be seen as a strong supporting factor for a prediction that Moore's Law will continue to hold for two or three more decades. Drexler [1992] specifies a conservative design that serves as a lower bound on the capabilities of computational systems nanotechnology could produce, which can be summed up as 10^{15} IPS per cubic millimeter and 10^{16} IPS per watt. It corresponds to a gate-size metric that Moore's Law predicts for about 2030.

The implication is that by 2029, a Moravec HEPP will cost one dollar.

Note that we are intentionally ignoring the software side of the AI. While that is currently the most problematic aspect in a scientific sense, once AI is developed the software—complete with education—can be copied with negligible cost.

The number of applications to which a human-level intelligence adds at least a dollar of value is staggering. Thus we can confidently predict that human-level AIs will be exceedingly numerous.

AIs of greater-than-human intelligence are also likely. We know that humans with IQs of up to 200 or so can exist, and thus such levels of intelligence are possible. Less is known about the organization of complexity in intelligent systems than the generation of raw computational speed. Even mere speed can be useful, though. An AI operating at 1000 times human speed could read an average book in one second with full comprehension, or take a college course, with plenty of homework and research, in ten minutes. It could write a book in two or three hours, and produce a human's lifetime intellectual output, complete with all the learning and experience that formed it, in a couple of weeks.

By 2036, as Apophis returns, the Earth it approaches may well be home to tens of billions of such superintelligent AIs.

²Marvin Minsky, personal communication: note that the actual informal estimate was somewhat lower than the one used here.

2 Leibniz vs. the Martians

It seems inescapable that an understanding of the moral character of the AIs is crucial in evaluating such a scenario. We will, as noted above, assume as a starting point that operationally defined AIs can and will exist. That is, we will assume that machines can be built and programmed not only to converse in English and perform competently at any intellectual activity that humans can master, but that they learn, have insights, and so forth in the human manner, growing in wisdom as they gain experience.

Many technologists, most notably Turing [1950], simply assume an operational definition and leave the discussion there. There is an implicit assumption that once an AI passes the Turing Test, intuitive doubts as to its qualities of mind will evaporate in the face of experience. Indeed, there is evidence that this happens well before it is ought to: Weizenbaum [1976] noted with alarm that people imputed human-like qualities to his Eliza program, a pattern-matching conversationalist of nearly transparent simplicity. The Eliza Effect, and its turbocharged latter-day cognates involving robots with body language and facial expressions, are strong warnings that our intuitions as to other minds are less reliable than we like to think.

Even if carried out by a skeptical cognitive scientist, the Turing Test provides only a relatively static snapshot of an AI, and misses the plasticity and autogeny that mark the ability of a human mind to learn and grow.

Thus philosophers can reject an operational definition of mind with some justification. However, this leaves us with a vexing open question as to the aspects of mind in machines. The seminal intuition is expressed by Leibniz in his Monadology:

Supposing that there were a machine whose structure produced thought, sensation, and perception, we could conceive of it as increased in size with the same proportions until one was able to enter into its interior, as he would into a mill. Now, on going into it he would find only pieces working upon one another, but never would he find anything to explain perception. (Leibniz, Monadology 17.)

For perception we can substitute any subjective phenomenon, up to and including the experience of conscious will. I take Searle's famous Chinese Room to be essentially a reapplication of this same intuition (applied to a different end).

The problem with this intuition as a guide to the realities of mind can be seen with a modern version of Leibniz' scenario:

Suppose that you have been abducted by Martians and taken to their laboratory, where they put you into an super-nano-scope that allows them, and you,

to observe any aspect of your body or brain at any scale down to the molecular. The Martians are able to discover, with you looking over their shoulders, that absolutely everything you do or feel can be explained by physics, chemistry, biology, and so forth; whatever decision you make, whatever action you take, is completely predictable by mechanistic scientific laws.

A human, too, is a “mere” machine; there is no more free will to be found in a gamma globulin than in a gear. Leibniz’ insight holds, and the succeeding centuries of science have served only to fill in the details. Schools of thought holding vitalistic views have retreated steadily since the Enlightenment, and the end is in sight.

This leaves us with a dilemma. It is essentially an extension of the one at the core of the mind/body issue as well as the freedom/determinism issue: what can the mentalistic terms and phenomenological references actually mean? It is invalid to reject them, as operationalists sometimes do, as illusions: illusion is itself a mentalistic term.

The philosophical school of thought that, over the past few decades, has been taken to resolve this dilemma is the Computational Theory of Mind (hereinafter CTM). We are essentially forced to accept the ontological basis of CTM, in something like its original formulation by Putnam [1960], since in the broad sense “a computation” is synonymous with “the detailed working out, for a specific instance, of the formal rules of the description of a deterministic system.”

We do not, however, have to accept the further development of CTM, by Fodor and others, over the succeeding decades, as the whole story. Ironically, the main problem with this “standard” CTM as seen by an information systems architect is that it is much too simplistic. Fodor himself [2000] notes as much.

The nuclei and electrons which make up a cell, such as a neuron, are governed by the laws of quantum mechanics. By empirical observation and extrapolation of the physics, however, we note certain regularities: nuclei tend to trap electrons to form atoms, and atoms tend to share electrons in covalent bonds to form molecules. A cell contains thousands of molecular species, which undergo various physical and chemical reactions with each other. At any given point in time and space, the concentration of solution of these species can be represented by a numeric vector, and the time behavior of the solution modelled by a system of differential equations over the vector—a classic dynamical system. For the cell as a whole, its interior can be gridded and a simulation carried out. If the cell is a neuron, it can be modelled at a higher level of abstraction involving electronic circuit elements (still a dynamical system).

In going from each of these levels of abstraction to a higher one, phenomena are lost. For example, the typical models of molecules as atoms with bonds assumes a static distribution of charge, where the actually varying pattern can subtly affect the docking behavior of large

molecules. A theory at the level of covalent bonds handles molecular orbital structures such as benzene rings poorly.

The computational theory of mind we are forced, ontologically, to accept can be likened to the quantum level of description of a cell. It seems risky to assume that there are fewer salient levels of abstraction in a useful theory of mind above, say, the neuron level than there are to a cell (or to modern software). It is likely that the propositional attitude states and so forth of the standard CTM are near the top of a deep and as yet dimly understood hierarchy of explanation. At every level, we're likely to see phenomena like the covalent approximation: exceptions to the (elegant forms of) the higher-level theory, which can be explained (or computed) only by reference to the lower level, or by databases of special cases, which are essentially cached results of precomputed lower-level cases.

As an aside, it should be noted that the lower levels of a mature CTM will almost certainly be dynamical systems. However, this has virtually no implications for CTM in the broad sense—the evolution of a dynamical system in time is a computation in the sense used. Indeed, every actual electronic digital computer is implemented by a circuit that is a dynamical system; and simulating dynamical systems was the original application of digital computers: the “I” in ENIAC stands for “Integrator.” It is only remarkable that the dynamical system, which has been physical science’s prime modelling tool for centuries, was largely neglected in cognitive science until recently—but that is a story beyond the scope of this essay.

Although we cannot offer any substantial detail, we can make the following predictions as to the ultimate shape of a mature CTM:

- It will be causal and mechanistic.
- It will involve multiple levels of abstraction, but higher levels may require reference to lower levels in exceptional cases.
- It will contain dynamical as well as symbolic and algorithmic elements, and forms of computation like associative memory that are not part of standard algorithmic practice.
- At the higher levels, the architecture will be modular with definable information flows between modules. This does not preclude the possibility of various global communication channels, however.
- At intermediate levels, there will be information patterns recognizable as symbols; but these will be non-atomic with a wealth of implicit relationships implied by their structure.

- Propositional attitudes, qualia, free will, and the other aspects of mind which are of interest will be identified with various configurations and properties of the mental computational architecture in a satisfying way. The vast majority of perceptions, inferences, memory formation, and so forth are heuristic in form, adaptive in the ancestral environment but not general, sound, or complete in a mathematical sense.

To justify the use of “satisfying” in the above, we will attempt an example. The best-developed theory of this form of which the author is aware is for free will. It has elements of an error theory and of compatibilism.

First, a cognitive architecture from McDermott [2001]: an AI contains a model of the world. In order to take choices between possible actions, it simulates the actions in the model to estimate their effect, and applies a utility function to the resulting world states. (It is a Popperian creature, allowing its theories to die in its stead.) The model must necessarily contain a model of the AI itself, to perform the simulated actions; but just as necessarily, causality in the model is broken at the point of the self-model, since it is directed by the AI’s own decision algorithm from outside the world model.

Thus, the AI’s self-model is exempt from causal law in a way unique in its model, and thus in its understanding, of the world. It cannot be otherwise: any attempt to model the actual decision-making mechanism within the self-model leads to an infinite regress and prevents the algorithm from terminating. If the AI’s intuitions of causality are readouts of the model, it must conceive itself free.

Secondly, Wegener [2002] describes a strong basis for understanding the sensation of having acted with conscious will. While phenomenal consciousness in general is beyond the scope of this example, suffice it to say that there appears to be a mechanism which produces a summary narrative from all the various information processed in the mind. Various acts in this trace record are tagged as having been willed. The trace, however, is a heuristic reconstruction (and vast simplification) of the actual decision process.

Lacking the Martians’ super-nano-scope, Wegener convincingly makes his analysis based on boundary cases, where the subject believes he willed something he didn’t, or vice versa (and the remarkable delay between beginning to do an action and consciously deciding to do it in the Libet [2004] experiments). Note that the heuristic reconstruction works properly in the vast majority of cases; the boundary cases are quite analogous to visual “optical illusions.” Strong evidence for the heuristic and post-hoc nature of the sensation of intention is found in the confabulations produced with such “will illusions.”

The modelling / evaluation / acting / recording structure sketched above is a good one for rapid learning. The dynamics of the world model can be implemented as a database of causal pairs of the form “this happened and then (therefore) that happened.” The most

important of these will be the ones of the form, “I did this, therefore that happened.” The conscious-will summarizer produces exactly this latter kind of record to augment the model continuously with experience.

Together, these analyses evoke a system that is satisfying, at least to this author, in the sense that it is exactly what we should expect. The mind is the operation of a machine formed by evolution to survive and reproduce on the savannas of Africa. Our self-image is exactly appropriate to creatures who can choose and plan by considering possibilities. Our sensation of will, like our vision, is not mathematically exact (which, by the way, would be mathematically impossible), but is usually right where it usually matters. Rapid learning of the consequences of our actions is invaluable.

Evolution had no charter to give us clear insight into the intricacies of cognitive data processing any more than it had for celestial mechanics or protein folding. Only the painstaking march of science has done that, and often against a strong inertia bolstered by intuitions of geocentrism or vitalism. A mature theory of mind, coherent with evolutionary theory of our origins and physical theory of our construction, must contain the intuitions as explananda and not rely on them as explanans.

It is probably fair to say that no artificial computational process achieved to date has the appropriate structure to claim, say, genuine conscious will. Indeed, as the pieces are put together over the next decade(s), it will be possible only in retrospect to define the point at which it did happen. But happen it will.

3 Moral Machinery

The nice thing about ethical theories, like standards, is that there are so many of them to choose from. The major problem seems to be that there is no agreed-upon starting point—divine revelation? eudaimonia? categorical imperative? veil of ignorance? In the face of this, many modern commentators eschew a theoretical framework and appeal more or less directly to the moral intuitions of their readers. To some extent, although little acknowledged, this follows the “moral sense” theory propounded by Adam Smith [1759].

But such moral intuitions, of course, are modules in our mechanical minds formed by evolution in an environment of tribal foragers. They are as much contingent facts of evolutionary history as the shape of an oak leaf. In reasoning from them to statements in the moral realm, we run afoul of Moore’s [1903] naturalistic fallacy—which was promulgated specifically to counter Spencer’s Social Darwinist ethics. Moore is surely right in insisting that we not simply identify the good with some arbitrary property (such as evolutionary fitness). However, it seems much more defensible to base a study of the good on what our

moral modules tell us, and how they got to be that way. One can see little justification for moral inquiry otherwise.

We are forced to proceed essentially as with the mentalistic phenomena. We can gain some preliminary extensional definition of the good by reference to intuition, including the intuitions of the great moral philosophers. We can then regularize and abstract it, refer to cognitive architecture for its function, and to the environment of evolutionary adaptation for its effect. After that, we will be on our own; our existing environment differs from the ancestral one significantly, and the environment we foresee as a result of superintelligent AI differs almost unrecognizably.

Internally, a particular ethic seems to resemble the grammar of a natural language. There are structures in our brains that predispose us to learn our native ethic, in that they determine within broad limits the kinds of ethics we can learn, and that while the ethics of human cultures vary within those limits, they have many structural features in common. (This notion is fairly widespread in latter 20th-century moral philosophy, e.g. Rawls [1971], Donagan [1977].) Our moral sense, like our competence at language, is as yet notably more sophisticated than any simple set of rules or other algorithmic formulation seen to date.

Ethics have much in common from culture to culture: structural similarities reminiscent of the deep structure of language syntax. One in particular is salient: moral imperatives are associated with actions which contravene self-interest or common sense. Ethics are something more than arbitrary customs for interactions. There is no great difference made if we say “red” instead of “rouge,” so long as everyone agrees on what to call that color; similarly, there could be many different basic forms of syntax that could express our ideas with similar efficiency.

But one of the points of an ethic is to make people do things they would not do otherwise. The reason is that, particularly for social animals, there are many kinds of interactions whose benefit matrices have the character of a Prisoner’s Dilemma or Tragedy of the Commons, i.e. where the best choice from the individuals’ standpoint is at odds with that of the group as a whole. Furthermore, and perhaps even more importantly, there were many actions whose long term effects we simply don’t understand. Thus in many cases, the adoption of a rule that seemed to contravene common sense or one’s own interest, if generally followed, can have a substantial beneficial effect on a human group.

Many animals—social insects are an extreme example—have evolved innate behaviors that model altruism or foresight beyond the individual’s understanding. Some of these, such as altruism toward one’s relatives, can clearly arise simply from selection for genes as opposed to individuals. However, there is reason to believe that there is much more going on, and that humans have evolved an ability to be programmed with arbitrary (within certain limits)

ethics.

The reason a human ethic is learned is the same as the reason that toolmaking is learned: the ancestral environment changed fast enough that a general capability for rapid adaptation was more adaptive than any specific innate capability. Hunting techniques had to change faster than we could evolve beaks or saberteeth; we learned to make and modify stone knives and spears appropriate to the game, and clothing appropriate to the climate. The environment which would have made innate long-term or altruistic behaviors adaptive was unstable as well.

The ethics themselves are produced by cultural evolution, not individuals (usually). The individual must, for any of this to work, absorb the ethic, like language, from the culture as a natural part of maturation, and it must then occupy a place in the cognitive database that is authoritative and read-only (or at least very difficult to change). In other words, the individual must sense the learned ethic as a universal absolute. Otherwise it would not be able to perform its function of modifying behavior against the strong motives of self-interest and common sense.

In sum, our argument is that there is a separate “ethic learning” module, similar to the language learning one, and it feeds into an action-evaluation database. A fuller theory, beyond our present scope, would identify a plethora of subsidiary modules, including those producing and interpreting affective display.

This at first sounds like an error theory: here’s why we believe right and wrong exist, but it’s only this gadget in our brains designed to fool us. But consider: the ethic we inherit is developed by cultural evolution, as is, say, science. Science may not be Ultimate Absolute Truth—but it is the best we have, and much better than any individual could invent alone. Our learned cultural ethic is similar. Its distilled wisdom is greater than our individual knowledge, since it was gathered over historical time and many people.

If ethics are part of culture, are they arbitrary like clothing styles or expressions of a successively improving understanding of some objective reality, like science? First, let us point out that clothing styles are not completely arbitrary: an Inuit and a Seminole could not exchange dress and prosper. Clothing varies according to availability of materials as well as to climate. Ethics vary similarly and appropriately. In both cases, however, there is a logic to the appropriateness, which can be discovered.

For most cultural knowledge, how to make a stone knife for example, the individual’s cognitive endowment—experience, inferential ability, and the results of immediate experiment—are an appropriate optimizer. Seeing a better way to make a stone knife, a person adopts it, and contributes to the culture thereby. In contrast, the individual’s intelligence is counterproductive in some cases: notably, participation in non-zero-sum games of the Prisoner’s

Dilemma variety, and in cases where the “horizon effect” of limited personal experience produces the wrong answer. (These latter are typically low-probability high-stakes situations, e.g. whether to wear a seatbelt).

4 Asenian Robotics

It is common in the transhumanist and singulatarian fields to worry about autogenous–self-modifying and extending–AIs. They might, it is reasoned, remove any conscience or other constraint we program into them, or simply program their successors without them. But it is in fact we, the authors of the first AIs, who stand at the watershed. We cannot modify our brains (yet) to alter our own consciences, but we are faced with the choice of building our creatures with or without them.

An AI without a conscience, by which we mean both the innate moral paraphernalia in the mental architecture and a culturally inherited ethic, would be a superhuman psychopath. Prudence, indeed, will dictate that superhuman psychopaths should not be built; however, it seems almost certain that it will be done within the next two decades. Most existing AI research is completely pragmatic, without any reference to moral structures in cognitive architectures. Furthermore, much of the most advanced research is sponsored by the military, where the notion of an autonomous machine being able to question its orders on moral grounds is anathema. The other major venue where research seems likely to produce AI is corporate industry, where the top goal seems likely to be the fiduciary benefit of the company.

Bostrom [2002] analyses the situation as follows:

If a superintelligence starts out with a friendly top goal, however, then it can be relied on to stay friendly, or at least not to deliberately rid itself of its friendliness. This point is elementary. A “friend” who seeks to transform himself into somebody who wants to hurt you, is not your friend. A true friend, one who really cares about you, also seeks the continuation of his caring for you. ...

In humans, with our complicated evolved mental ecology of state-dependent competing drives, desires, plans, and ideals, there is often no obvious way to identify what our top goal is; we might not even have one. So for us, the above reasoning need not apply. But a superintelligence may be structured differently. If a superintelligence has a definite, declarative goal-structure with a clearly identified top goal, then the above argument applies. And this is a good reason for us to build the superintelligence with such an explicit motivational architecture.

Bostrom’s prescription of a “friendly” top-down motivational structure for an AI approx-

imates what, on a close reading, Asimov's [1950] "Three Laws of Robotics" amount to. Asimov conceived an internal mental structure in a dynamical systems formulation (see "Runaround"), and understood the problems inherent in reinterpretation of words, and assumes that the Laws can be built into the structure of the mind at a deeper level (see, among others, "Reason").

But Asimov's robots were not autogenous, and as noted above, the current probable sources of AI are not such as to admit of a generally adopted philanthropic formulation. The reasonable assumption, then, is that a wide variety of AIs with differing goal structures will appear in the coming decades.

A subtext of the singulitarian concern is that there may be the possibility of a sudden emergence of (a psychopathic) AI at a superhuman level, due to a positive feedback in its autogenous capabilities. There are three reasons for a lack of alarm.

First, although hardware for running a human-level AI exists, it is currently represented by the top ten or so supercomputers in the world. These are multimillion dollar installations, and have strong previous calls on their time. Even if someone were to pay to dedicate, say, Blue Gene to running an AI full time, it would only approximate a normal human intelligence. Ubiquitous superintelligence must wait for a decade or two of Moore's Law, implying nanotechnology.

Secondly, even when the hardware is available, the software is not. Some of the fears of sudden superintelligence are based on the notion that an early superintelligence would make writing the smarter next one faster, and so forth. It does seem likely that a properly structured AI could be a better programmer than a human of otherwise comparable cognitive abilities. It is ironic to note, however, that automatic programming is currently one of the most poorly developed of AI's subfields. Any reasonable extrapolation of current practice predicts that early human-level AIs will be secretaries and truck drivers, not programmers.

Even when AI computer scientists are achieved, adding one more to the existing field, which is already bending its efforts to improving AI, will not materially affect progress. Only when the total AI which is in fact devoting its efforts to this project begins to rival the intellectual resources of the existing field of AI, will significant acceleration occur.

Thirdly, intelligence does not spring fully formed like Athena from the forehead of Zeus. Even we humans, with the processing power of a thousand supercomputers at our disposal, take years to mature. A human requires about a decade to become really expert in any given field—including AI programming. More to the point, it takes the scientific community some extended period to develop a theory, and the engineering community some further time to put it into practice. Even if we had a complete and valid theory of mind, which we do not, putting it into software would take years; and the early versions would be incomplete and

full of bugs.

Human developers will need years of experience with early AIs before they get it right. Even then they will have systems that are the equivalent of slow, inexperienced humans. Software has a law of advance similar to Moore's Law for hardware, less celebrated and less precisely measurable but nevertheless real. Advances in algorithmics tend to produce software speedups analogous to the hardware ones. The completely understood, tightly coded, highly optimized software of mature AI will run a human equivalent in real time on a 10-100 teraops machine, but early versions will not.

There are two wild-card possibilities to consider. First is rogue AIs living on botnets—groups of hijacked PCs communicating via the internet. A best guess for current processing power available runs to 10,000 Moravec HEPP or 10 Kurzweil HEPP. However, the extreme forms of parallelism needed to make use of this form of computing resource, along with the communication latency involved, will tend to push the reasonable estimates toward the Kurzweil level (which is based on the human brain with its high-parallelism, slow cycle time architecture). Furthermore, that, together with the existing increasingly sophisticated security community, will make the development of AI software much harder in this mode than in a standard research setting. Thus, while we can expect botnet AI's in the long run, they are unlikely to be first.

The second possibility is that Minsky is right. Very few business or academic LANs currently offer less than a Minsky HEPP. If somehow an early AI were to find a “resonance keyhole” that allowed strong positive feedback into such a highly optimized form, it would find ample processing power available. And this could be aboveboard—a Minsky HEPP costs much less than a person is worth economically.

Let us, somewhat presumptuously, attempt to explain Minsky's intuition by an analogy: a bird is our natural example of the possibility of heavier-than-air flight. Birds are immensely complex: muscles, bones, feathers, nervous systems. But we can build working airplanes with tremendously fewer moving parts. Similarly, the brain can be greatly simplified, still leaving an engine capable of general conscious thought.

The author's intuition is that Minsky is closer to being right than is generally recognized in the AI community, but that computationally expensive heuristic search will turn out to be an unavoidable element of adaptability and autogeny—and thus of any AI capable of the runaway feedback loop singularity fear.

Almost certainly then, at least a full decade will elapse between the appearance of the first genuinely general, autogenous AIs and the time that they become significantly more capable than humans. This will indeed be a crucial period in history, but no one person, group, or even school of thought will control it. The question instead is, what can be done

to influence the process to put the AIs on the road to being a stable community of moral agents?

A possible path is shown in the experiments of Axelrod [1984] and of course in the original biological evolution of our own morality. In a world of autonomous agents who can recognize each other, cooperators can prosper and ultimately form an evolutionarily stable strategy. Superintelligent AIs should be just as capable of understanding this as humans are.

The environment in which AI morality will evolve will have some significant differences from the one in which ours did. The bad news:

- The disparities between the abilities of AIs could be significantly greater than those between humans, and more correlated with an early “edge” in the race to acquire resources. This can negate the evolutionary pressure to reciprocal altruism.
- Corporate AIs will almost certainly start out self-interested, and evolution favors effective self-interest.

It has been suggested, e.g. by Pinker [1997], Yudkowski [2003], and Hawkins [2004], that AIs would not have the “baser” human instincts built in and thus not need moral restraints. But it should be clear that they *could* be programmed with baser instincts, and it seems obvious to this author that corporate ones will be, and that military ones will be programmed with different but equally disturbing motivational structures.

Furthermore, it should be noted that any goal structure implies self-interest. Consider two agents, either of whom might get the use of some given resource. Unless the agents’ goals are identical, each will further its own goal more by using the resource for its own purposes, and consider it at best suboptimal and possibly counterproductive for the resource to be controlled by the other agent and thus applied to some other goal.

It should go without saying how greatly goal structures can vary even if both agents are programmed to seek, say, “the good of humanity.”

Back to human/AI differences—the good news:

- As a matter of practical fact, criminality is strongly negatively correlated with IQ in humans.³
- Reproduction of AIs is likely to be asexual and to involve the inheritance of acquired characteristics. Significantly greater and more detailed forms of memory transfer will be available than with humans. For these reasons, individuals AIs are likely to have access to experience at the extra-personal scales for which human morality is a heuristic.

³E.g. Herrnstein & Murray p.246: In one major national study, among white males aged 15-23, the average IQ of those never arrested was 106, of those ever sentenced to a correctional facility 93.

- Intra- as well as inter-AI structure may well be designed on a economic as opposed to a dominance (pecking order) model, since this is in fact more efficient. AI's could thus be without the mechanics we seem to have whereby authority can short-circuit morality, as in the Milgram [1963] experiments (or the Nazi experience).
- The economic law of comparative advantage implies that cooperation between individuals of greatly differing capabilities remains mutually beneficial.
- Individual lifetime is not arbitrarily limited, so an AI has the prospect of living into the far future, in a world whose character its actions help create. Forever is a long time to try to hide an illicit deed.
- Latter-day superintelligent AIs will be able to read and absorb the full corpus of writings in moral philosophy, especially the substantial recent work in evolutionary ethics, and understand it better than we do.
- AIs intending to alter their own morality modules could make the source code public and invite public scrutiny and revisions.
- Moral AIs would be able to track other AIs in much greater detail than humans do each other, for vastly more individuals. This allows a more precise formation and variation of cooperating groups.
- Self-selecting communities of utility-based rational AIs would be able to use the logic of superrationality, as expounded by Hofstadter [1996],⁴ to optimize their collective efforts.
- AIs will have considerably better insight into their own natures and motives than humans do. Thus an AI may have the ability to more completely honest than humans, who believe our own confabulations. (It is worth noting that some commentators, such as Nadeau [2006], propose that *only* an AI could be a moral agent, because confabulation means that humans' reasons for acting differ from our perceived intentions!)

We can only at present theorize about the ultimate moral capacities of AIs, but this list strongly suggests that an AI with a moral character not only on par with, but significantly better than that of humans, is not only possible, but perhaps even likely.

⁴In essence: if every agent party to a symmetric interaction is rational, they must make the same decision; knowing this, each will independently act to optimize the total utility.

It is clear that the world stands in need of such today, but even more so for the future: not only nanotechnology, but the advancing sciences of molecular biology and of the human mind make for fearsome thickets of potential for abuse or disaster in the coming decades.

One obvious ethical innovation that will be needed is a sound treatment of the question of the ethical relationship of minds of vastly different capabilities. Another will be the question, all too soon to be a possibility, of direct manipulation and modification of the minds of others—with or without consent.

By the time Apophis returns in 2036, the technological means could exist not only to prevent its doing mischief, but if desired to capture it as a valuable material resource. Neglect of the appropriate advances in astronautics would clearly be a dereliction of duty. Similarly, research and development in the moral nature of mind is certainly a duty of the highest order; not only now and for us, but for the AIs themselves when they do arrive.

5 Acknowledgements

The author wishes gratefully to acknowledge the attention and insights of Robert A. Freitas Jr., Douglas Hofstadter, Christopher Grau, David Brin, Chris Phoenix, John Smart, Ray Kurzweil, Eric Drexler, Eliezer Yudkowsky, and Robin Hanson.

References

- [1] ASIMOV, ISAAC. *I, Robot*. Doubleday, 1950.
- [2] AXELROD, ROBERT. *The Evolution of Cooperation*. Basic Books, 1984.
- [3] BOSTROM, NICK. Ethical Issues in Advanced Artificial Intelligence. in *Cognitive, Emotive and Ethical Aspects of Decision Making in Humans and in Artificial Intelligence*, Vol. 2, ed. I. SMIT et al., Int. Institute of Advanced Studies in Systems Research and Cybernetics, pp. 12-17, 2003.
- [4] DONAGAN, ALAN. *The Theory of Morality*. Univ Chicago Press, 1977.
- [5] DREXLER, K. ERIC. *Nanosystems: Molecular Machinery, Manufacturing, and Computation*. Wiley, 1992.
- [6] FODOR, JERRY. *The Mind Doesn't Work That Way*. MIT, 2000.
- [7] HALL, J. STORRS. *Nanofuture: What's Next for Nanotechnology*. Prometheus, 2005.

- [8] HAWKINS, JEFF. *On Intelligence*. Times Books, 2004.
- [9] HERRNSTEIN, RICHARD, and CHARLES MURRAY. *The Bell Curve: Intelligence and Class Structure in American Life*. Free Press, 1994.
- [10] HOFSTADTER, DOUGLAS. *Metamagical Themas: Questing for the Essence of Mind and Pattern*. HarperCollins, 1996. pp. 739-755.
- [11] KURZWEIL, RAY. *The Singularity is Near*. Viking, 2005.
- [12] LEIBNIZ, GOTTFRIED WILHELM. *Monadology*. Originally written in French in 1714, various translations are available online, e.g. philosophy.eserver.org/leibniz-monadology.txt
- [13] LIBET, BENJAMIN. *Mind Time: The Temporal Factor in Consciousness*. Harvard, 2005.
- [14] MCDERMOTT, DREW V. *Mind and Mechanism*. MIT, 2001
- [15] MILGRAM, STANLEY. Behavioral study of obedience. *Journal of Abnormal and Social Psychology*, Vol. 67, pp. 371-378. 1963.
- [16] MOORE, GEORGE EDWARD. *Principia Ethica*. Cambridge, 1903.
- [17] MORAVEC, HANS. *Robot: Mere Machine to Transcendent Mind*. Oxford, 1999.
- [18] NADEAU, JOSEPH EMILE. Only Androids can be Ethical. in *Thinking about Android Epistemology*, K. FORD, C. GLYMOUR, AND P. HAYES, eds. AAAI/MIT, 2006, pp241-8.
- [19] PINKER, STEVEN. *How the Mind Works*. Norton, 1997.
- [20] PUTNAM, HILARY. Minds and Machines. In *Dimensions of Mind*, edited by S. Hook. New York University Press, 1960.
- [21] RAWLS, JOHN. *A Theory of Justice*. Harvard/Belknap, 1971.
- [22] SMITH, ADAM. *The Theory of Moral Sentiments*. A. Millar, 1790.
- [23] TURING, ALAN M. Computing machinery and intelligence. *Mind* 59:433-460, 1950.
- [24] YUDKOWSKI, ELIEZER. *Creating Friendly AI* <http://www.singinst.org/CFAI/index.html> 2003.
- [25] WEGENER, DANIEL M. *The Illusion of Conscious Will*. MIT, 2002.