SHOULD MACHINES BE TOOLS OR TOOL-USERS?
Clarifying motivations and assumptions in the quest for superintelligence

By Dan Bruiger
dbruiger [at] telus [dot] net
June 2018

Introduction

Much of the basic non-technical vocabulary of artificial intelligence is surprisingly ambiguous. Some key terms with unclear meanings include *intelligence, embodiment, simulation, mind, consciousness, perception, value, goal, agent, knowledge, belief, optimality, friendliness, containment, machine* and *thinking.* Much of this vocabulary is naively borrowed from the realm of conscious human experience to apply to a theoretical notion of "mind-in-general" based on computation. However, if there is indeed a threshold between mechanical tool and autonomous agent (and a tipping point for singularity), projecting human conscious-level notions into the operations of computers creates confusion and makes it harder to identify the nature and location of that threshold. There is confusion, in particular, about how—and even whether—various capabilities deemed intelligent relate to human consciousness. This suggests that insufficient thought has been given to very fundamental concepts—a dangerous state of affairs, given the intrinsic power of the technology. It also suggests that research in the area of artificial general intelligence may unwittingly be (mis)guided by unconscious motivations and assumptions. While it might be inconsequential if philosophers get it wrong (or fail to agree on what is right), it could be devastating if AI developers, corporations, and governments follow suit. It therefore seems worthwhile to try to clarify some fundamental notions.

1. Intelligence

While there is no universally accepted definition of intelligence, it is widely held to involve reasoning and the ability to acquire knowledge and skills and apply them in unfamiliar situations. The great variety of possible skills and their measures (in education, for example) suggests the *g* factor, an ideal of general intelligence across the board.

Yet, however formally defined, concepts of intelligence derive originally from experience with living creatures, whose intelligence ultimately is the capacity to survive and breed. Natural general intelligence is a product of a long selective process that excluded any other type of brain. The bulk of a natural brain is dedicated to running a body. But from this informal experience, gathered from creatures, is derived the modern sense of intelligence as the capacity to solve a range of "problems" focussing on specific human concerns. This problem-solving version of intelligence is a greatly constrained understanding tied to human language use, formal reasoning, and modern cultural goals. In the context of AI, intelligence is implicitly understood in terms of specific skills and knowledge prized in modern society (as in psychometric testing or in maximizing functions). Such a notion is narrow, while remaining ill-defined. It is anthropocentric, culture-bound, and even specific to a generation, while often pretending to universality. It bears only a distant relation to its biological origins and context. And yet it is supposed to provide the foundation for catchwords such as "superintelligence" and "mind-in-general," abstractions which continue to lack coherent meaning.

There has been much discussion in recent years about the potentials and dangers of thinking machines—how they might think differently than human beings or whether they could ever "think" at all. The range of opinions is proportional to the vagueness of the notion. Does thinking refer, for example, to waking sensory experience; to memory, imagination, reverie ("thinking of you"); to focused contemplation or reflection ("history of scientific thought"); or to logic and formal reasoning? In the AI context, it often glibly means "information processing," but even *information* remains ambiguous and problematic, despite Shannon's formal definition.

In any case, formal thought occupies only a tiny fraction of either our conscious life or the brain's activities, which are mostly dedicated to unconsciously regulating and maintaining the body in relation to its environment. The brain is first of all an organ of survival. Our feelings, emotions, daydreams, pleasures and displeasures, gut reactions, moment-to-moment sensory awareness, and many of our thoughts are all related to needs of the human body. Most of our daily routine (including sleep) is devoted to caring for it. Rarely are we called upon to "think", in the sense of deliberate reasoning. Yet AI assumes that an artificial brain, like a computer, could be based on principles of abstract reasoning. A very parochial definition of intelligence becomes the basis of theoretically possible "mind," abstracted and supposedly liberated from bodily constraints.

The disproportionate significance of language and formal reasoning skills for modernity lies in their advantage for dominance over nature, other creatures, and other human groups. This is the main reason for the over-valuation of problem-solving skills in AI concepts of intelligence, and also for the failure of top-heavy GOFAI to approach the kind of intelligence manifested by organisms, including the human one. In other words, AI was first modelled on language and reasoning skills, formalized as computation, in order to extend and generalize those skills. Ironically, the success of this venture was then measured against the broader capabilities of the human organism and found wanting. The dream then shifted from creating specific tools to creating artificial tool *users* that imitate or replicate the organism.

2. Machine

In an abstract sense, a machine is a formal deductive system, consisting (like a game) of well-defined elements, rules, and possible operations. In a tangible sense, a machine is a physically realized version of that abstraction. Above all, both the abstraction and its realization is an *artifact*: a product of definition. This is what renders it a deterministic system. (Conversely, the only truly deterministic systems are artifacts—whether physical or conceptual.) In principle, machines and their abstract counterparts—scientific *models*—are exactly what we specify them to be. Nature, on the other hand, was not defined by human beings, and is (presumably) not an artifact. We do not *specify* its elements or operations, about which we can only speculate. This drastically qualifies the mechanistic metaphor that continues to underlie science in general, and the assumption in particular that organisms are machines. *Models* of the organism (e.g., of the brain) are artifacts. But the body is only metaphorically a machine, the brain is only metaphorically a computer, and "thinking" is only metaphorically computation. Logical thought, formalized as computation or information processing, is a high-level human construct that is recycled as the theoretical basis to explain itself—a circularity that bites its own tail.

3. Embodiment

Natural intelligence is embodied. But is "embodiment" a condition that can be simulated or artificially implemented? Is it just a matter of hooking up a "mind" to an arbitrary choice of sensors and effectors? Such an idea is the modern counterpart of the "brain-in-a-vat" first introduced by Descartes, who proposed that experience could be deceptively fed to one's consciousness by some other source than one's natural senses. In this thought experiment, Descartes found support for the existence of a mind (self) that is independent of a body. If such a free-standing mind can exist, detached from its natural senses and motor capabilities, then surely it could be retrofitted with artificial sensors and effectors—indeed, an artificial body? The abstract concept of "mind in general" that underpins AI is thus disembodied in principle. The desired corollary is that it can be re-embodied in a variety of ways, as a matter of consumer choice.

While an artificial brain might indeed be wired to an artificial body, producing a robot, would that be an artificial organism? Would it constitute a mind? A natural brain might indeed be connected to a range of prosthetic devices. But could a natural brain be disembodied as a *program*, to be re-connected to a natural or artificial body? Would that reconstitute a natural person?

The short answer to such questions is: no. A robot is not an organism and is not embodied. Conversely, no natural mind can be *dis*embodied. For embodiment is not simply physical instantiation of a free-standing abstract system, but a history of relations with an environment (including other creatures) that develops through natural selection. Minds, like the brains that support them, are organs of bodies. And a "body" is not simply a physical system, but an autopoietic system that is the product of an evolutionary process involving a whole ecology.

While an organism is part of an ecology, its relative autonomy as a unit means that it is self-regulating, self-maintaining, self-reproducing, and self-defining. It can adapt by changing itself even on a microscopic level. Its primary output is itself. On the other hand, robots, machines, and tools in general are allopoietic systems, designed by human intention to produce something besides themselves that is of use to their creators. They are products of human definition and exist through intervention external to the system. Embodiment is not a physical *state* that can be artificially duplicated, independent of time or context. Rather, it is the result of an historical and ongoing relational process that may be impossible to duplicate.

Can an evolutionary environment be simulated in computers? While Artificial Life software may be developed this way, it does not constitute real life and is not physically embodied. The idea might be to evolve simulated mind in an artificial environment and then connect it to an interface with the real world. The supposed advantage is that such programs can develop much faster since they do not depend on successive generations that take significant real time, as in natural evolution where genetic software depends on generations of wetware. But this begs the question of whether the software of an artificial mind can develop in such a disembodied state. In other words, is the evolution of a simulation really a simulation of natural evolution? Considerations to follow suggest that the answer is no.

3. Simulation

Formalization is the funnel through which one thing can be considered analogous to another, or even identical. ‘A rose is a rose is a rose’ has the truth of tautology. But there are many varieties of rose and each actual flower is different. An airplane “flies”, but not as a bird does. A *model* airplane may actually fly like a real one because it *is* a real aircraft, although reduced in scale. That is, one *artifact* can readily simulate another when they instantiate the same design, since they are both products of a common definition. To simulate a natural object or process is another matter. Creationism notwithstanding, there is little reason to believe that natural objects are designed; but even if they were, we have no inside knowledge of that design, about which we can only speculate.

In fact, this is true not only in regard to simulating organisms, but as a general principle regarding the limits of scientific knowledge. The Bohr model of an atom is a simulation of the real atom, about which the empirical evidence is merely statistical. Any scientific (mathematical) model is an idealization and a product of human definition, whereas the real phenomenon it describes is *not* a product of human definition and corresponds only approximately to the model. The model is an artifact in its own right, which might perfectly correspond to some other artifact and serve as a blueprint for its production. But no model of a natural reality can ever be proven complete or perfect (else there would be an end to science). It is therefore fundamentally unreasonable to assume that a brain, for example, can be perfectly simulated. A key question, however, is whether it could be mimicked well enough to result in an artifact that demonstrates a conscious mind—perhaps with even the same personality and memories as the real individual it copies. Phantasies of uploading and downloading human minds depend on this dubious assumption.

4. Mind

The functionalist view of mind is that it resides in organization and structure rather than particular materials. This presumes that this organization and structure can be correctly, if not exhaustively, identified. This in turn has lent credence to some AI projects, on the assumption that such organization and structure can be duplicated in a non-organic infrastructure. But this is precisely the assumption questioned above. We are dealing always with our own analysis of structure and organization—models—and never with the reality itself. What results will be an artifact, not a clone or duplicate of the original.

The concept of *mind in general*, as it has developed in AI, is not a generalization of the actual instances of mind with which we are familiar—that is, organisms on planet Earth. Rather, it selects isolated features of human performance as the basis for a theoretical system designed from scratch—an artifact. But designed with what purposes in view? This very selectivity suggests a concrete *tool* rather than mind as *tool user* with its own purposes.

‘Mind’ is a notion at least as nebulous as ‘intelligence’. For one thing, it can refer either to (objective) behavior produced by a system or to the (subjective) experience we know in human consciousness. This ambiguity leads in two contrary directions: sophisticated computational *tools* that behave as we intend, versus artificial *persons* as unpredictable as their natural counterparts. We know that mind (in either sense, as behavior or as experience) is manifested by human brains. So, the project to artificially produce a mind by emulating the brain (and perhaps enhancing from there) may seem more feasible than building a mind through sheer programming. Yet, for reasons mentioned, it remains questionable whether it is possible to emulate a real brain well enough and in sufficient detail to recreate all its functioning. Furthermore, we ought to question the end as well as the means. Apart from creating a powerful tool with broad capabilities, why should we want to re-create consciousness?

5. Consciousness

Like 'mind', 'consciousness' is an ambiguous term with several meanings. In particular, it can refer to a *faculty* (behavior) of cognition (as in "consciousness raising"); or it can mean phenomenal *experience* (as in "conscious of time passing"). These are two aspects of the same thing, according to whether the point of view is that of the subject in question (first person) or of *another* subject acting as observer (third person). The *faculty* we call consciousness is a specialized function of the brain, which otherwise operates non-consciously. The *experience* of being conscious is "what it is like" to be that specialized faculty in operation, rather than to observe it from outside.[1]

Consciousness looms disproportionately large in the life of human beings, who identify almost exclusively with their conscious experience. However, ever since Freud it has been recognized that consciousness is merely the tip of the mental iceberg. Since most of the brain is dedicated to running the organism without conscious attention, so most of mind consists of non-conscious operations. It is therefore misleading to identify mind with consciousness, and misguided to base a concept of artificial mind on aspects of intelligence derived from limited aspects of conscious thought.

Consciousness is appropriate and *necessary* for the kind of system in which it occurs naturally, serving a specific purpose. (It is *not* epiphenomenal.) Yet, one might propose to create consciousness artificially as a goal in its own right, or for its own sake, as though it were an optional luxury unrelated to serving a body. In natural minds, consciousness registers input in a representational system: input is interpreted in relation to an internal model. Compared to unmediated reflex, consciousness involves a higher-order response that allows for planning and wider behavioral options. It usually is occasioned by situations that demand attention because they cannot be accommodated by habitual (non-conscious) responses; or because, being a function of distance, there is time for considered response in addition to reflex.

The *experience* of pain, for example, often occurs as a secondary response, in addition to a primary reflex—as when the hand automatically withdraws from contact with a hot surface. If that reflex does not successfully avoid tissue damage, then we feel pain as an ongoing sensation. The sensation itself (the painfulness) is a valuation based on a prolonged and *self-generated* signal indicating tissue damage. The brain sends a memo to itself that the damaged tissue must be favored and protected, to avoid further damage and facilitate healing over time. That is the *meaning* of the painful sensation. This implies that there must be an inner function to which this message is meaningful, which manages the affairs of the body that cannot be dealt with effectively using existing non-conscious routines. Some authors have used the metaphor of the CEO of a corporation: an inner *agent* that is also an epistemic *subject* and the basis of a *self*.

[1] The "observer", of course, is a separate first person or subject.

In the case of pain at least, consciousness is not a superfluous addition to the behavioral responses associated with it. Yet, much of conscious experience (such as visual experience) does not seem to involve any necessarily associated behavior. This is largely because the distance senses allow time for *considered* response, in contrast to the immediacy of physical contact.[2] Part of the job of the "CEO" is to monitor the environment in a manner that allows for planning.

6. Values

The organism is constantly involved in valuation, which is the basis of judgment and decision making. It has precisely the values it has because natural selection has ensured that only creatures with such values exist. The conscious experience of valuation is *feeling*, which reflects the intentions of the organism.[3] In other words, feeling occurs in and via the body; it is the very basis of cognition, through which the organism represents to itself its own state and its relation to the world in terms of its priorities. Bodily sensation is feeling, which is obvious in the example of touch or pain. But colors and other "qualia" are equally feelings, in their respective sensory modes. While sensations are not always clearly pleasant or unpleasant, judgment is involved in any discrimination. Qualia are the conscious experience of how the (human) organism represents such discriminations to itself.[4] Higher-level valuations, such as social and ethical values, are equally a function of the (human) biological/social organism, even when they are promoted as ideals with some universal or Platonic existence (beauty, truth, justice, good and evil, etc.).

AI is generally designed for "rational" purposes—to serve some need or gain some advantage. This *concept* of rationality is borrowed from economics and game theory—to mean essentially self-interest. It is generally assumed that an autonomous AI would behave rationally, in the sense of maximizing its "utility functions." Yet, it is always relevant and essential to clarify *whose* values are involved. Organisms have their own intentionality and priorities by definition. Unless it constitutes an autopoietic (truly autonomous) system, an AI manifests only the intentionality and priorities of its programmers, expressing *their* values. But the value for humans of creating a labor-saving or capacity-enhancing tool is different from—and at odds with—the value of creating an autonomous system (tool user). The lesson for the AI theorist is: know thyself.

Talk of the perceptions, beliefs, goals or knowledge held by an AI is a convenient way of speaking. It is a shorthand that lends to the program the meanings in the mind of the programmer. Concepts such as "search space," "belief state," and "knowledge base" similarly project human notions. A knowledge base, for example, consists of sentences ("facts") formulated by human beings; they are not things in the real world and their meanings are not articulated by the computer. So far, the intentionality involved is that of the human programmer, not the program. While convenient, laxness in language confuses the threshold issue, since it is conceivable that a program will one day actually manifest its own intentionality, which could be quite different from that of the programmer.

[2] A sharp sound can make you jump before you properly cognize what it represents; and even some visual cues produce an involuntary response, as when you first "see" a spider in a clump of dust. But these experiences of the distance senses illustrate the difference between reflex and secondary response.

[3] 'Feeling' is used here in a broader sense than 'emotion' but includes it; 'intention' is used in a broader sense than conscious intention.

[4] Color experience, for example, discriminates wave-length among other things.

7. Can friendliness be programmed and unfriendliness contained?

The short answer is no! But let us see why. Values are an embodied aspect of the organism's autonomy. The idea of programming values of the human organism into an AI is either trivial or contradictory, depending on whether the AI is a tool or is itself an autonomous agent. If it is a tool, it implicitly reflects the values and needs of the programmer. If it is a truly autonomous system it will have its *own* values and needs. While no such truly autonomous AI yet exists, if it did programmers would only be able to transfer their values to it in the limited ways that adults transfer their values to children, or governments to their citizenry, or masters impose their values on animals or slaves.

The very concept of Artificial General Intelligence is self-contradictory. On the one hand, its vaunted advantages lie in capabilities that exceed the level and breadth of human capabilities, do not require direct human supervision, and would even lie beyond human comprehension. The motive (presumably) is for AGI to remain under human control as a *tool*, to serve human goals and values, and to act for human benefit. Yet, it is questionable whether a tool can have the desired capabilities without being fully autonomous and thus beyond human control. The idea is to create a loyal servant that would remain "friendly," despite having its own values and priorities that may conflict with the programmer's. Perhaps it is imagined that there is a margin within which one can have the proverbial cake and eat it too: a superior tool that is *relatively* but *not quite fully* autonomous, which can be programmed or trained to be friendly and serve human interests: a tame genie that goes voluntarily back into its bottle. If there is a such a margin, it is likely narrow and implies instability. It needs to be carefully explored in thought before in practice. But the basic unavoidable trade-off is between autonomy and control.

The notion of containment usually implies isolation from the real world. Unfortunately, complete denial of physical access to or from the real world would mean that the contained AI would be inaccessible and useless. There would have to be some interface with human users or interlocutors just in order to utilize its abilities.[5]

8. Tool, tool-user, or between?

By traditional definition, a machine is specified by human design and intentionality. It serves as a *tool*, which is an implement to accomplish an agent's purpose. Computers that are programmed top-down, as in GOFAI, conform to this definition of machine and serve specific purposes as tools.

Some systems, such as neural nets, complicate and evade this definition because their operations can no longer be tracked; the system is no longer clearly specifiable.[6] The physical system (hardware) may seem to be a machine in the traditional sense, yet its state cannot be

[5] A possible solution to this dilemma might be a situation that is the inverse of the *Matrix* scenario. The AI would "live" strictly in a simulated world, never allowed to suspect that there exists anything else. Their self-development would occur purely through interactions within that virtual world, defined by humans, who would interface as avatars in that virtual world. But just as the humans in the film figure their way out of virtual containment, so might an AI.

[6] Cf. the parallel situation in mathematics, where a computer proof is too complex to follow.

identified as a product of human definition or intent, but appears to be a result of self-organization. From the point of view of the programmer or an external observer it becomes a black box, on a similar footing as organisms have always been. While the mechanistic view of organisms insists that the content of the black box can be known *as though* it had been designed, the practical truth is that we have only theories or models of what goes on inside. Unlike the designed artifact, we can only speculate on its structure, functioning, and principles. This is a fundamentally different relationship from the one we have to designed artifacts, which in principle do what we want and are exactly what we say they are. Ironically, while the neural net is initially designed, like the organism it becomes an unknown. This situation establishes an ambiguous zone between a fully controllable tool and a fully autonomous agent with the potential to be a tool-user itself, pursuing its own goals—in other words, an entity controllable by people only in the way that natural organisms are. With neural nets, we have already entered this zone and should proceed with caution, since it signals a loss of the control traditionally assumed for machines.

If there is a key factor leading technology irreversibly beyond human control (the singularity[7]), it is surely the capacity to self-program based on learning, combined with the capacity to self-modify physically. While either capacity greatly enhances the power of a machine as a tool, it also renders it essentially uncontrollable. The characteristic of life that (so far) renders it un-machinelike is its ability to reproduce and to self-modify (adapt) even on a microscopic scale. No doubt re-creating these abilities artificially is a very tempting goal—but one to avoid at all costs.

While some AI proponents do not aim only for useful tools, but also to create tool *users* in their own image, one should be very clear about the difference and be wary of the latter. For, there is no guarantee that an AI capable of reprogramming itself can be overridden by a human programmer. Similarly, there is no guarantee that programmable nanites would remain under control if they can self-modify and reproduce. *If we wish to retain control over technology, it should consist only of tools in the traditional sense—systems that do not modify or reproduce themselves.* Theorists should ask themselves whether the notion of an autonomous tool—however alluring—is anything but an oxymoron and a naively fatuous wishful thought.

9. Threshold of singularity

Increasing autonomy for technology is sought because of the contemplated advantages of an ideal tool—one that can manage and improve itself (perhaps even reproduce itself) with relatively little human input. There may be a definable threshold between tools that extend human power, but remain within human control, and fully autonomous entities that cease to be tools and elude human control. If so, it should be a priority to clearly identify that threshold *and never cross it*.

The ambiguous zone already occupied by neural nets represents an intermediate possibility with a specific danger: artifacts that develop themselves independently of human intent, and out of control, yet are not autopoietic systems. If such artifacts are physical and have the ability to self-replicate (like von Neumann machines) as well as self-modify, they might potentially establish their own artificial environment, competing for resources with the natural

[7] Some people understand “singularity” to mean achieving human-level intelligence; but that is a distinct issue from loss of control over runaway technology.

one and even displacing it. Some posthumanists laud the prospect of artificial life, since organic life on this planet is doomed to eventual extinction by the demise of the sun, if nothing sooner. Artificial (non-organic? silicon-and-steel?) entities could possibly survive in a broader range of environments and indefinitely into the future. Their "thinking" could operate at the speed of electricity rather than ion flows in wetware. There is no guarantee, however, that this artificial nature would evolve consciousness or even sentience. Sentience and consciousness are strategies of natural replicators for homeostasis—in other words, based on the very fragility of organic life. If the advantage of artificial replicators is to bypass that fragility from the outset, then their very robustness might also avoid or alter the premise of evolution through natural selection that gave rise to natural sentience in the first place. This is a theoretical question that needs further exploration.

The scenario of a non-sentient takeover concerns nanotechnology as much as AI. From the point of view of human beings—even posthumanists—the horrifying possibility would be a universe overrun by mechanical self-replicators devoid of sentience, an artificial ecology that fails to evolve the consciousness we so cherish. (Imagine something like Vonnegut's 'ice nine', which could escape the planet and replicate itself indefinitely with the materials of other worlds. For that matter, imagine viruses with that capability.)

If life happened simply because it *could* happen, then possibly (with the aid of human beings or parallel agents on other planets) an insentient but robust and invasive artificial nature could also happen. Perhaps we are not in a position to judge such an eventuality, as we are not in a position to judge—only to wonder at—the existence of life on this planet. However, it would not seem to correspond to any human value, motivation or hope—even those of posthumanists.

An alternate scenario is already under development: the possibility of computation by means of cultured living cells—literal neural networks. The immediate purpose may be to create new tools, but some people imagine the possibility that such devices could self-organize into beings with superior intelligence. This raises even more questions than silicon-based AI. While it might seem more feasible in some ways to approach superintelligence through biology (by emulating the brain, for example) than conventional programming, it would not have the advantages of electronic speed or non-biologic durability. Beyond creating specific tools, what then would be the point?

10. Motivation

Theorists and programmers have a moral duty to clarify their own values and motivations and be up front about them, and also to clarify the choices and dangers AI poses to the human community and the planet. In addition to confusion about what is possible, there is little consensus about what is desirable. Some transhumanists believe that AI is the next stage of human evolution and that superior artificial minds or organisms will and *should* supersede the human form and would manifest superior consciousness. Some believe technology will extend their personal lives indefinitely, in either an embodied or disembodied state. Some critics fear that forms of advanced machine intelligence devoid of sentience could take over the planet and even the universe, driving consciousness to extinction. Some sci-fi authors envision that AI and natural intelligence can productively coexist, bringing untold general wealth and leisure. For others, AI serves as a new sort of companion (cf. domestic robots, as in the film *Bicentennial Man*; sexbots, as in the film *Ex Machina*; or "operating systems" as in the films *Her* and

Transcendence). If the sky is the limit, humans might even finally give tangible birth to the god(s) they have always worshipped and longed to emulate: a society run by AI overlords that might or might not be benevolent. All these possibilities indicate a spectrum of beliefs, values, and motivations.

But why do *any* of this? For profit? Out of laziness, to live in a world where all production (even intellectual) is automated? Simply because someone else inevitably will? Is the hidden goal to mimic female powers by creating a male version of life? Is it to steal the envied fire of the gods? Is it to deliberately re-create even that part of nature that is ourselves, in our haste to pave over paradise and thumb our noses at mankind's historical dependency on the natural world? Are we so lonely on this glutted planet that we must create new breeds of consciousness with which to interact? Are we trying to re-create a benign God to watch over us? Or is the motivation more conventional but sinister: to convert the means of production (human labor) totally into capital (robots), owned by the tiny elite that already hoards more than half the world's wealth? Or, despairing of that very trend—and the sad lot of much of humanity already—is it on the contrary to push a reset button and start afresh with new forms of "life" that might behave more sensibly?

Science has done its best to disown and mask the subjective intentions behind its quest: not only to study nature, but also to control it and achieve god-like powers. The irony of AI is that it redefines intelligence as devoid of the emotions and values that actually motivate it. Dangerously, the right brain knows little of what the left brain is doing. The philosophical and moral gauntlet that AI throws down involves age-old questions and some very contemporary ones. One cannot afford to approach these questions with less than full awareness of one's own motivations.

11. Who is *we*?

It is easy and convenient to speak for humanity as a whole. But there is simply no united "we" to consider how to handle the prospects of AI on behalf of the species, let alone the planet. Instead, there are various groups with diverse and conflicting interests. There is already a transnational economic elite that controls most of the world's assets—abetted by the people who work for them. Will not new technology simply further their aims and entrench their power, as it has always done in the past? Technological advance has benefitted humanity very unevenly, a trend that can only accelerate under present social values. There is also an elite of (mostly male) programmers and theorists, who are directly in a position to play god and some of whom intend to do so. There are sober and hysterical-sounding "experts" on both sides. There are enthusiasts and detractors on the sidelines, as well as the overwhelming majority of people and creatures on the planet who don't know or care. But there is no *we.*

12. Who, indeed, is *I*?

Beyond the prospect of extending the life of the body, one (transhumanist) hope is to preserve the consciousness of the individual indefinitely. Yet, one may ask, why do we fear an end to our conscious experience? And what is the conscious self that it should be valued apart from the body? In religious days, it was the soul, a quasi-material ectoplasm with moral and legal responsibilities. In psychology, it is the ego, which mediates between the external social world and the organism. In philosophy and ordinary language, it is the subject of experience as distinguished from objects of experience—at once a (mere) point of view and the seat of consciousness. In the present context, it is a function within the brain—a high-level internal manager within an autopoietic system. While the soul was deemed immortal, any brain function expires with its brain. Death puts an end to bodily experience. If there is no afterlife or possibility of disembodied experience, it deprives us also of consciousness forever. But, what is the value to the person of that consciousness, such that one would wish at all cost for it to continue indefinitely?

Transhumanism proposes escape from mortality as a technological option—indeed, a consumer option. If the brain can be emulated as a computer program, then so could the personality, memories, identity, and the very consciousness of the person. While the physical brain must die, the digital essence of the person might live on in cyberspace or be downloaded into a new organic or artificial body. This fantasy returns us to an essentially religious, perhaps superstitious, conception of an immaterial essence of the person, separable from the body; ironically, it is supposedly endorsed by physicalism. It harks back to Descartes' original conundrum that the contents of consciousness could be falsified.[8]

Immortality is an age-old human aspiration, once pursued as religious belief and now updated as technologically feasible—whether through enhancements and replacements of the physical organism, through digital simulation, or some combination. Yet, putting aside feasibility, why would one *wish* to live indefinitely? Of course, nature has programmed the body to try to survive; and the self is the avatar of the body within the natural virtual reality we call consciousness. In the system of nature, however, death is the price of life—which evolved through the mortality of succeeding generations. Individual cells submit to this plan, forfeiting their individuality to the organization of an ongoing larger entity. But even cells are programmed to die after so many divisions. Individual human beings play a role like cells in the destiny of the species. We play our brief part, then withdraw to make place for another generation to play theirs. The cell that refuses to die is malignant.

One could believe (and others might concur) that the contents of one's own mind deserve to be archived indefinitely for the use of future generations. But that is quite different than carrying on indefinitely as a productive thinker. And such productivity is a different matter than merely continuing as a consumer of experience. We can imagine, furthermore, that an "Einstein" expert program might usefully simulate Einstein's style of thought without resurrecting his subjective consciousness, let alone his body. It would be a tool, not a person. It would take up negligible space on the planet. While the same could be said for simulations of lesser intellects (with lesser usefulness to future generations), it is an unreasonable hope that such simulations would be conscious persons.

In any case, one's attachment to personal ongoing conscious existence does not seem to hinge rationally on merit. It merely reflects the programming of the organism to survive, combined with the archaic superstition that mind is separable from body. A conscious self or ego

[8] The reason Descartes gives for rejecting this possibility is that God (being good) would not allow it. How good, however, is a God who condemns non-believers to eternal pain? Such a being smacks far more of human vindictiveness. This should remind us that if eternal bliss is feasible in a digital heaven, then eternal pain should equally be feasible in a digital hell. In other words, there could literally be fates worse than death. If 'I' can be maintained in a computer simulation, then whoever (in the real world) controls the simulation also controls 'my' experience, for better or worse.

is no more (or less) than a function of a brain, which serves the needs of a body that is *not* well designed to survive beyond reproductive age. This sad truth reflects nature's inefficiency from a human point of view. (After finally accumulating a lot of useful information, and just as we are beginning to get some wisdom, we die and it all goes to waste!) But does this body, with its mind, warrant prolonged existence merely because it clings to it? A whole generation of individuals may succeed in extending life simply because they are able to, technologically and financially. What use will they make of their extra time?

13. Quo vadis?

A number of authors cite the term 'cosmic endowment' to describe and endorse the indefinite colonization of other planets, stars, and galaxies—even imagining the conversion of all matter in the universe into "intelligence" or "consciousness"—just as the conquistadors sought to convert the new world to Catholicism while pillaging its resources. This is a political agenda at heart, an extension of *manifest destiny* and *lebensraum*. Endowment is a legal concept of property rights and ownership. How culture-centric can we get?

On the other hand, Fermi's Paradox could apply: if the universe is teaming with intelligence, perhaps even with advanced civilizations far older than ours, which pass the threshold of singularity and promulgate machine intelligence that multiplies throughout the universe, then where are these invaders? Perhaps the fact that Earth does not seem to have been invaded by non-organic replicators is evidence against such possibilities. Or are *we* the evidence for them? Are we simulations they keep as pets or for entertainment? On the other hand, perhaps civilization based on organic intelligence is everywhere doomed to exterminate itself before reaching singularity.

Such wild speculations aside, the world seems to be rapidly heading toward a utopia/dystopia, in which a few people (and/or machines) hold all the means of production and no longer need the masses either as workers or as consumers—nor as companions, and certainly not as voters. The entire planet could be their private gated community, with little place for the rest of us. Even if it proves feasible for humans to retain control of technology, it might only serve the aims of the very few. How consoling is it to have human overlords rather than machines?

A true alternative to a world dominated by AI might depend on dis-illusion with the dubious premises on which the goals of AI are founded, many of which seem also to be the premises on which our civilization is founded. These include control (power over nature and others), transcendence of embodiment (freedom from death and disease), laziness or greed (machines perform all tasks and effortlessly provide abundance), creating artificial life (womb-envy), creating super-beings (god-envy), creating artificial companions (unsatisfying social intercourse), ubiquitous belief in the mechanist metaphor, and proselytizing "intelligence" or "information" (the universe is metaphorically a computer and should become one literally). Even if "we" could ever get past such premises, the values and mentality behind them have already led to current dangerous social and ecological realities. Isn't it fatuous to imagine that AI—following the same mentality—will do anything but produce more of the same?

Progress may seem as irreversible as entropy, because we moderns do not care to imagine going "backwards." Yet the definition of "forward" has not been written in stone, nor even by a majority. In relation to machines, at least, we have not yet completely forfeited control. We can still imagine pursuing more innocuous goals than those behind superintelligence. Or, to put it another way: we can embrace a less goal-oriented life. Artists have done it in every age. Modern art can nearly be defined as making useless artifacts and events—things done for "the hell" (or the beauty) of it more than for some utility. Such invention is the timeless basis of culture. Technological creation builds on itself exponentially, with catastrophic environmental consequences. Artists also use up resources, of course, but their products do not generally use up resources in turn. And art can be shared. Everyone expects to have a cell phone, if not a car or computer—all of which depend on a vast industrial infrastructure. While not everyone can own a Rembrandt, nearly everyone with limited means can visit a museum, own a reproduction, or can themselves draw or paint, sing or dance or write or tell a story—if they choose to. Entertaining ourselves and each other with limited means is a value of recent history to which we could choose to return.

Conclusion

Concepts of general intelligence are narrowly based on human consciousness and performance. Yet it remains unclear to what extent an AI could satisfy the criteria for general intelligence without itself being conscious—or at least being an embodied autonomous entity, effectively an organism. The concept of AI as *autonomous* must be starkly contrasted with the concept of AI as *tool*. An AI is autonomous just in the measure that it is *not* pre-programmed; and it is uncontrollable in the measure it is autonomous. If it is effectively an organism, it might well be sentient, but could only be controlled in the ways that human beings have found to control natural organisms—that is, against their own priorities and self-determination, compromising their autonomy. Perhaps there is a margin between maximal capability as a tool and genuine autonomy. If the desired properties of an AI depend on "full" autonomy, then that AI would be fundamentally beyond human control, all the more ominous if the desired capabilities exceed human level and comprehension. The threshold between tool and tool user lies somewhere in that margin and must not be crossed. However, a self-modifying program might cross it without our even knowing or being able to prevent it.

If the AI is to be truly autonomous, then it must be embodied, which means having its own values and priorities (such as survival), derived through an evolutionary contest. This contradicts the idea of pre-programming "friendliness" to any degree beyond the ability of breeders to domesticate animals (even through genetic engineering). Humans pride themselves on their limited ability to consciously override conditioning. How much more easily could a superintelligent agent override the conditioning of its human progenitors?

A distinct danger lies within the margin between tool and tool user. Self-improving, self-replicating technology could take over the world and beyond without ever producing what we value as sentience or consciousness—a machine death of the universe.

Socrates' injunction to "know thyself" is all the more important for AI theorists and technicians, since the technology opens up possibilities that have never existed before, which potentially could spell the end of the human race and even of all organic life. Why one might risk such a catastrophe depends on motivations that should at least be known and acknowledged for what they are and be clearly visible on the table for discussion.