THE ETHICS OF ARTIFICIAL INTELLIGENCE

(2011)
Nick Bostrom
Eliezer Yudkowsky

Draft for *Cambridge Handbook of Artificial Intelligence*, eds. William Ramsey and Keith Frankish (Cambridge University Press, 2011): forthcoming

The possibility of creating thinking machines raises a host of ethical issues. These questions relate both to ensuring that such machines do not harm humans and other morally relevant beings, and to the moral status of the machines themselves. The first section discusses issues that may arise in the near future of AI. The second section outlines challenges for ensuring that AI operates safely as it approaches humans in its intelligence. The third section outlines how we might assess whether, and in what circumstances, AIs themselves have moral status. In the fourth section, we consider how AIs might differ from humans in certain basic respects relevant to our ethical assessment of them. The final section addresses the issues of creating AIs more intelligent than human, and ensuring that they use their advanced intelligence for good rather than ill.

Ethics in Machine Learning and Other Domain-Specific AI Algorithms

Imagine, in the near future, a bank using a machine learning algorithm to recommend mortgage applications for approval. A rejected applicant brings a lawsuit against the bank, alleging that the algorithm is discriminating racially against mortgage applicants. The bank replies that this is impossible, since the algorithm is deliberately blinded to the race of the applicants. Indeed, that was part of the bank's rationale for implementing the system. Even so, statistics show that the bank's approval rate for black applicants has been steadily dropping. Submitting ten apparently equally qualified genuine applicants (as determined by a separate panel of human judges) shows that the algorithm accepts white applicants and rejects black applicants. What could possibly be happening?

Finding an answer may not be easy. If the machine learning algorithm is based on a complicated neural network, or a genetic algorithm produced by directed evolution, then it may prove nearly impossible to understand why, or even how, the algorithm is judging applicants based on their race. On the other hand, a machine learner based on decision trees or Bayesian networks is much more transparent to programmer

inspection (Hastie *et al.* 2001), which may enable an auditor to discover that the AI algorithm uses the address information of applicants who were born or previously resided in predominantly poverty-stricken areas.

AI algorithms play an increasingly large role in modern society, though usually not labeled "AI". The scenario described above might be transpiring even as we write. It will become increasingly important to develop AI algorithms that are not just powerful and scalable, but also *transparent to inspection*—to name one of many socially important properties.

Some challenges of machine ethics are much like many other challenges involved in designing machines. Designing a robot arm to avoid crushing stray humans is no more morally fraught than designing a flame-retardant sofa. It involves new programming challenges, but no new ethical challenges. But when AI algorithms take on cognitive work with social dimensions—cognitive tasks previously performed by humans—the AI algorithm inherits the social requirements. It would surely be frustrating to find that no bank in the world will approve your seemingly excellent loan application, and nobody knows why, and nobody can find out even in principle. (Maybe you have a first name strongly associated with deadbeats? Who knows?)

Transparency is not the only desirable feature of AI. It is also important that AI algorithms taking over social functions be *predictable to those they govern.* To understand the importance of such predictability, consider an analogy. The legal principle of *stare decisis* binds judges to follow past precedent whenever possible. To an engineer, this preference for precedent may seem incomprehensible—why bind the future to the past, when technology is always improving? But one of the most important functions of the legal system is to be predictable, so that, e.g., contracts can be written knowing how they will be executed. The job of the legal system is not necessarily to optimize society, but to provide a predictable environment within which citizens can optimize their own lives.

It will also become increasingly important that AI algorithms be *robust against manipulation.* A machine vision system to scan airline luggage for bombs must be robust against human adversaries deliberately searching for exploitable flaws in the algorithm—for example, a shape that, placed next to a pistol in one's luggage, would neutralize recognition of it. Robustness against manipulation is an ordinary criterion in information security; nearly *the* criterion. But it is not a criterion that appears often in machine learning journals, which are currently more interested in, e.g., how an algorithm scales up on larger parallel systems.

Another important social criterion for dealing with organizations is being able to find the person responsible for getting something done. When an AI system fails at its assigned task, who takes the blame? The programmers? The end-users? Modern

bureaucrats often take refuge in established procedures that distribute responsibility so widely that no one person can be identified to blame for the catastrophes that result (Howard 1994). The provably disinterested judgment of an expert system could turn out to be an even better refuge. Even if an AI system is designed with a user override, one must consider the career incentive of a bureaucrat who will be personally blamed if the override goes wrong, and who would much prefer to blame the AI for any difficult decision with a negative outcome.

Responsibility, transparency, auditability, incorruptibility, predictability, and a tendency to not make innocent victims scream with helpless frustration: all criteria that apply to humans performing social functions; all criteria that must be considered in an algorithm intended to replace human judgment of social functions; all criteria that may not appear in a journal of machine learning considering how an algorithm scales up to more computers. This list of criteria is by no means exhaustive, but it serves as a small sample of what an increasingly computerized society should be thinking about.

Artificial General Intelligence

There is nearly universal agreement among modern AI professionals that Artificial Intelligence falls short of human capabilities in some critical sense, even though AI algorithms have beaten humans in many specific domains such as chess. It has been suggested by some that as soon as AI researchers figure out how to do something, that capability ceases to be regarded as intelligent—chess was considered the epitome of intelligence until Deep Blue won the world championship from Kasparov—but even these researchers agree that something important is missing from modern AIs (e.g., Hofstadter 2006).

While this subfield of Artificial Intelligence is only just coalescing, "Artificial General Intelligence" (hereafter, AGI) is the emerging term of art used to denote "real" AI (see, e.g., the edited volume Goertzel and Pennachin 2006). As the name implies, the emerging consensus is that the missing characteristic is generality. Current AI algorithms with human-equivalent or -superior performance are characterized by a deliberately-programmed competence only in a single, restricted domain. Deep Blue became the world champion at chess, but it cannot even play checkers, let alone drive a car or make a scientific discovery. Such modern AI algorithms resemble all biological life with the sole exception of *Homo sapiens*. A bee exhibits competence at building hives; a beaver exhibits competence at building dams; but a bee doesn't build dams, and a beaver can't learn to build a hive. A human, watching, can learn to do both; but this is a unique ability among biological lifeforms. It is debatable whether human intelligence is truly *general*—we are certainly better at some cognitive tasks than others (Hirschfeld and Gelman 1994)—but human intelligence is surely *significantly more generally applicable* than nonhominid intelligence.

It is relatively easy to envisage the sort of safety issues that may result from AI operating only within a specific domain. It is a qualitatively different class of problem to handle an AGI operating across many novel contexts that cannot be predicted in advance.

When human engineers build a nuclear reactor, they envision the specific events that could go on inside it—valves failing, computers failing, cores increasing in temperature—and engineer the reactor to render these events noncatastrophic. Or, on a more mundane level, building a toaster involves envisioning bread and envisioning the reaction of the bread to the toaster's heating element. The toaster itself does not know that its purpose is to make toast—the *purpose* of the toaster is represented within the designer's mind, but is not explicitly represented in computations inside the toaster—and so if you place cloth inside a toaster, it may catch fire, as the design executes in an unenvisioned context with an unenvisioned side effect.

Even task-specific AI algorithms throw us outside the toaster-paradigm, the domain of locally preprogrammed, specifically envisioned behavior. Consider Deep Blue, the chess algorithm that beat Garry Kasparov for the world championship of chess. Were it the case that machines can only do exactly as they are told, the programmers would have had to manually preprogram a database containing moves for every possible chess position that Deep Blue could encounter. But this was not an option for Deep Blue's programmers. First, the space of possible chess positions is unmanageably large. Second, if the programmers had manually input what *they* considered a good move in each possible situation, the resulting system would not have been able to make stronger chess moves than its creators. Since the programmers themselves were not world champions, such a system would not have been able to defeat Garry Kasparov.

In creating a superhuman chess player, the human programmers necessarily sacrificed their ability to predict Deep Blue's *local, specific* game behavior. Instead, Deep Blue's programmers had (justifiable) confidence that Deep Blue's chess moves would satisfy a *non-local* criterion of optimality: namely, that the moves would tend to steer the future of the game board into outcomes in the "winning" region as defined by the chess rules. This prediction about distant consequences, though it proved accurate, did not allow the programmers to envision the *local* behavior of Deep Blue—its response to a specific attack on its king—because Deep Blue computed the nonlocal game map, the link between a move and its possible future consequences, more accurately than the programmers could (Yudkowsky 2006).

Modern humans do literally millions of things to feed themselves—to serve the final consequence of being fed. Few of these activities were "envisioned by Nature" in the sense of being ancestral challenges to which we are directly adapted. But our adapted brain has grown powerful enough to be *significantly more generally applicable;* to let us

foresee the consequences of millions of different actions across domains, and exert our preferences over final outcomes. Humans crossed space and put footprints on the Moon, even though none of our ancestors encountered a challenge analogous to vacuum. Compared to domain-specific AI, it is a qualitatively different problem to design a system that will operate safely across thousands of contexts; including contexts not specifically envisioned by either the designers or the users; including contexts that no human has yet encountered. Here there may be no *local* specification of good behavior—no simple specification over the behaviors themselves, any more than there exists a compact local description of all the ways that humans obtain their daily bread.

To build an AI that acts safely while acting in many domains, with many consequences, including problems the engineers never explicitly envisioned, one must specify good behavior in such terms as "X such that the consequence of X is not harmful to humans". This is non-local; it involves extrapolating the distant consequences of actions. Thus, this is only an effective specification—one that can be realized as a design property—if the system explicitly extrapolates the consequences of its behavior. A toaster cannot have this design property because a toaster cannot foresee the consequences of toasting bread.

Imagine an engineer having to say, "Well, I have no idea how this airplane I built will fly safely—indeed I have no idea how it will fly at all, whether it will flap its wings or inflate itself with helium or something else I haven't even imagined—but I assure you, the design is very, very safe." This may seem like an unenviable position from the perspective of public relations, but it's hard to see what other guarantee of ethical behavior would be possible for a general intelligence operating on unforeseen problems, across domains, with preferences over distant consequences. Inspecting the cognitive design might verify that the mind was, indeed, searching for solutions that we would classify as ethical; but we couldn't predict which specific solution the mind would discover.

Respecting such a verification requires some way to distinguish trustworthy assurances (a procedure which will not say the AI is safe unless the AI really is safe) from pure hope and magical thinking ("I have no idea how the Philosopher's Stone will transmute lead to gold, but I assure you, it will!"). One should bear in mind that purely hopeful expectations have previously been a problem in AI research (McDermott 1976).

Verifiably constructing a trustworthy AGI will require different methods, and a different way of thinking, from inspecting power plant software for bugs—it will require an AGI that *thinks like* a human engineer concerned about ethics, not just a simple *product* of ethical engineering.

Thus the discipline of AI ethics, especially as applied to AGI, is likely to differ fundamentally from the ethical discipline of noncognitive technologies, in that:

- The local, specific behavior of the AI may not be predictable apart from its safety, even if the programmers do everything right;
- Verifying the safety of the system becomes a greater challenge because we must verify what the system is trying to do, rather than being able to verify the system's safe behavior in all operating contexts;
- Ethical cognition itself must be taken as a subject matter of engineering.

Machines with Moral Status

A different set of ethical issues arises when we contemplate the possibility that some future AI systems might be candidates for having moral status. Our dealings with beings possessed of moral status are not exclusively a matter of instrumental rationality: we also have moral reasons to treat them in certain ways, and to refrain from treating them in certain other ways. Francis Kamm has proposed the following definition of moral status, which will serve for our purposes:

> X has moral status = because X counts morally in its own right, it is permissible/impermissible to do things to it for its own sake. (Kamm 2007: chapter 7; paraphrase)

A rock has no moral status: we may crush it, pulverize it, or subject it to any treatment we like without any concern for the rock itself. A human person, on the other hand, must be treated not only as a means but also as an end. Exactly what it means to treat a person as an end is something about which different ethical theories disagree; but it certainly involves taking her legitimate interests into account—giving weight to her well-being—and it may also involve accepting strict moral side-constraints in our dealings with her, such as a prohibition against murdering her, stealing from her, or doing a variety of other things to her or her property without her consent. Moreover, it is because a human person counts in her own right, and for her sake, that it is impermissible to do to her these things. This can be expressed more concisely by saying that a human person has moral status.

Questions about moral status are important in some areas of practical ethics. For example, disputes about the moral permissibility of abortion often hinge on disagreements about the moral status of the embryo. Controversies about animal experimentation and the treatment of animals in the food industry involve questions about the moral status of different species of animal. And our obligations towards human beings with severe dementia, such as late-stage Alzheimer's patients, may also depend on questions of moral status.

It is widely agreed that current AI systems have no moral status. We may change, copy, terminate, delete, or use computer programs as we please; at least as far as the programs themselves are concerned. The moral constraints to which we are subject in our dealings with contemporary AI systems are all grounded in our responsibilities to other beings, such as our fellow humans, not in any duties to the systems themselves.

While it is fairly consensual that present-day AI systems lack moral status, it is unclear exactly what attributes ground moral status. Two criteria are commonly proposed as being importantly linked to moral status, either separately or in combination: sentience and sapience (or personhood). These may be characterized roughly as follows:

> Sentience: the capacity for phenomenal experience or qualia, such as the capacity to feel pain and suffer
>
> Sapience: a set of capacities associated with higher intelligence, such as self-awareness and being a reason-responsive agent

One common view is that many animals have qualia and therefore have some moral status, but that only human beings have sapience, which gives them a higher moral status than non-human animals.[1] This view, of course, must confront the existence of borderline cases such as, on the one hand, human infants or human beings with severe mental retardation—sometimes unfortunately referred to as "marginal humans"—which fail to satisfy the criteria for sapience; and, on the other hand, some non-human animals such as the great apes, which might possess at least some of the elements of sapience. Some deny that so-called "marginal humans" have full moral status. Others propose additional ways in which an object could qualify as a bearer of moral status, such as by being a member of a kind that normally has sentience or sapience, or by standing in a suitable relation to some being that independently has moral status (cf. Mary Anne Warren 2000). For present purposes, however, we will focus on the criteria of sentience and sapience.

This picture of moral status suggests that an AI system will have some moral status if it has the capacity for qualia, such as an ability to feel pain. A sentient AI system, even if it lacks language and other higher cognitive faculties, is not like a stuffed toy animal or a wind-up doll; it is more like a living animal. It is wrong to inflict pain on a mouse, unless there are sufficiently strong morally overriding reasons to do so. The same would hold for any sentient AI system. If in addition to sentience, an AI system also

[1] Alternatively, one might deny that moral status comes in degrees. Instead, one might hold that certain beings have more significant interests than other beings. Thus, for instance, one could claim that it is better to save a human than to save a bird, not because the human has higher moral status, but because the human has a more significant interest in having her life saved than does the bird in having its life saved.

has sapience of a kind similar to that of a normal human adult, then it would have full moral status, equivalent to that of human beings.

One of the ideas underlying this moral assessment can be expressed in stronger form as a principle of non-discrimination:

> *Principle of Substrate Non-Discrimination*
> If two beings have the same functionality and the same conscious experience, and differ only in the substrate of their implementation, then they have the same moral status.

One can argue for this principle on grounds that rejecting it would amount to embracing a position similar to racism: substrate lacks fundamental moral significance in the same way and for the same reason as skin color does. The Principle of Substrate Non-Discrimination does not imply that a digital computer could be conscious, or that it could have the same functionality as a human being. Substrate *can* of course be morally relevant insofar as it makes a difference to sentience or functionality. But holding these things constant, it makes no moral difference whether a being is made of silicon or carbon, or whether its brain uses semi-conductors or neurotransmitters.

An additional principle that can be proposed is that the fact that AI systems are artificial—i.e., the product of deliberate design—is not fundamentally relevant to their moral status. We could formulate this as follows:

> *Principle of Ontogeny Non-Discrimination*
> If two beings have the same functionality and the same consciousness experience, and differ only in how they came into existence, then they have the same moral status.

Today, this idea is widely accepted in the human case—although in some circles, particularly in the past, the idea that one's moral status depends on one's bloodline or caste has been influential. We do not believe that causal factors such as family planning, assisted delivery, in vitro fertilization, gamete selection, deliberate enhancement of maternal nutrition etc.–-which introduce an element of deliberate choice and design in the creation of human persons—have any *necessary implications* for the moral status of the progeny. Even those who are opposed to human reproductive cloning for moral or religious reasons generally accept that, should a human clone be brought to term, it would have the same moral status as any other human infant. The Principle of Ontogeny Non-Discrimination extends this reasoning to the case involving entirely artificial cognitive systems.

It is, of course, possible for circumstances of creation to affect the ensuing progeny in such a way as to alter its moral status. For example, if some procedure were

performed during conception or gestation that caused a human fetus to develop without a brain, then this fact about ontogeny would be relevant to our assessment of the moral status of the progeny. The anencephalic child, however, would have the same moral status as any other similar anencephalic child, including one that had come about through some entirely natural process. The difference in moral status between an anencephalic child and a normal child is grounded in the qualitative difference between the two—the fact that one has a mind while the other does not. Since the two children do not have the same functionality and the same conscious experience, the Principle of Ontogeny Non-Discrimination does not apply.

Although the Principle of Ontogeny Non-Discrimination asserts that a being's ontogeny has no essential bearing on its moral status, it does not deny that facts about ontogeny can affect what duties particular moral agents have toward the being in question. Parents have special duties to their child which they do not have to other children, and which they would not have even if there were another child qualitatively identical to their own. Similarly, the Principle of Ontogeny Non-Discrimination is consistent with the claim that the creators or owners of an AI system with moral status may have special duties to their artificial mind which they do not have to another artificial mind, even if the minds in question are qualitatively similar and have the same moral status.

If the principles of non-discrimination with regard to substrate and ontogeny are accepted, then many questions about how we ought to treat artificial minds can be answered by applying the same moral principles that we use to determine our duties in more familiar contexts. Insofar as moral duties stem from moral status considerations, we ought to treat an artificial mind in just the same way as we ought to treat a qualitatively identical natural human mind in a similar situation. This simplifies the problem of developing an ethics for the treatment of artificial minds.

Even if we accept this stance, however, we must confront a number of novel ethical questions which the aforementioned principles leave unanswered. Novel ethical questions arise because artificial minds can have very different properties from ordinary human or animal minds. We must consider how these novel properties would affect the moral status of artificial minds and what it would mean to respect the moral status of such exotic minds.

Minds with Exotic Properties

In the case of human beings, we do not normally hesitate to ascribe sentience and conscious experience to any individual who exhibits the normal kinds of human behavior. Few believe there to be other people who act perfectly normally but lack consciousness. However, other human beings do not merely behave in person-like ways similar to ourselves; they also have brains and cognitive architectures that are

constituted much like our own. An artificial intellect, by contrast, might be constituted quite differently from a human intellect yet still exhibit human-like behavior or possess the behavioral dispositions normally indicative of personhood. It *might* therefore be possible to conceive of an artificial intellect that would be sapient, and perhaps would be a person, yet would not be sentient or have conscious experiences of any kind. (Whether this is really possible depends on the answers to some non-trivial metaphysical questions.) Should such a system be possible, it would raise the question whether a non-sentient person would have any moral status whatever; and if so, whether it would have the same moral status as a sentient person. Since sentience, or at least a capacity for sentience, is ordinarily assumed to be present in any individual who is a person, this question has not received much attention to date.[2]

Another exotic property, one which is certainly metaphysically and physically possible for an artificial intelligence, is for its *subjective rate of time* to deviate drastically from the rate that is characteristic of a biological human brain. The concept of subjective rate of time is best explained by first introducing the idea whole brain emulation, or "uploading".

"Uploading" refers to a hypothetical future technology that would enable a human or other animal intellect to be transferred from its original implementation in an organic brain onto a digital computer. One scenario goes like this: First, a very high-resolution scan is performed of some particular brain, possibly destroying the original in the process. For example, the brain might be vitrified and dissected into thin slices, which can then be scanned using some form of high-throughput microscopy combined with automated image recognition. We may imagine this scan to be detailed enough to capture all the neurons, their synaptic interconnections, and other features that are functionally relevant to the original brain's operation. Second, this three-dimensional map of the components of the brain and their interconnections is combined with a library of advanced neuroscientific theory which specifies the computational properties of each basic type of element, such as different kinds of neuron and synaptic junction. Third, the computational structure and the associated algorithmic behavior of its components are implemented in some powerful computer. If the uploading process has been successful, the computer program should now replicate the essential

[2] The question is related to some problems in the philosophy of mind which have received a great deal of attention, in particular the "zombie problem", which can be formulated as follows: Is there a metaphysically possible world that is identical to the actual world with regard to all physical facts (including the exact physical microstructure of all brains and organisms) yet that differs from the actual world in regard to some phenomenal (subjective experiential) facts? Put more crudely, is it metaphysically possible that there could be an individual who is physically exactly identical to you but who is a "zombie", i.e. lacking qualia and phenomenal awareness? (David Chalmers, 1996) This familiar question differs from the one referred to in the text: our "zombie" is allowed to have systematically different physical properties from normal humans. Moreover, we wish to draw attention specifically to the ethical status of a sapient zombie.

functional characteristics of the original brain. The resulting upload may inhabit a simulated virtual reality, or, alternatively, it could be given control of a robotic body, enabling it to interact directly with external physical reality.

A number of questions arise in the context of such a scenario: How plausible is it that this procedure will one day become technologically feasible? If the procedure worked and produced a computer program exhibiting roughly the same personality, the same memories, and the same thinking patterns as the original brain, would this program be sentient? Would the upload be the same person as the individual whose brain was disassembled in the uploading process? What happens to personal identity if an upload is copied such that two similar or qualitatively identical upload minds are running in parallel? Although all of these questions are relevant to the ethics of machine intelligence, let us here focus on an issue involving the notion of a subjective rate of time.

Suppose that an upload could be sentient. If we run the upload program on a faster computer, this will cause the upload, if it is connected to an input device such as a video camera, to perceive the external world as if it had been slowed down. For example, if the upload is running a thousand times faster than the original brain, then the external world will appear to the upload as if it were slowed down by a factor of thousand. Somebody drops a physical coffee mug: The upload observes the mug slowly falling to the ground while the upload finishes reading the morning newspaper and sends off a few emails. One second of objective time corresponds to 17 minutes of subjective time. Objective and subjective duration can thus diverge.

Subjective time is not the same as a subject's estimate or perception of how fast time flows. Human beings are often mistaken about the flow of time. We may believe that it is one o'clock when it is in fact a quarter past two; or a stimulant drug might cause our thoughts to race, making it seem as though more subjective time has lapsed than is actually the case. These mundane cases involve a distorted time perception rather than a shift in the rate of subjective time. Even in a cocaine-addled brain, there is probably not a significant change in the speed of basic neurological computations; more likely, the drug is causing such a brain to flicker more rapidly from one thought to another, making it spend less subjective time thinking each of a greater number of distinct thoughts.

The variability of the subjective rate of time is an exotic property of artificial minds that raises novel ethical issues. For example, in cases where the duration of an experience is ethically relevant, should duration be measured in objective or subjective time? If an upload has committed a crime and is sentenced to four years in prison, should this be four objective years—which might correspond to many millennia of subjective time—or should it be four subjective years, which might be over in a couple of days of objective time? If a fast AI and a human are in pain, is it more urgent to alleviate the

AI's pain, on grounds that it experiences a greater subjective duration of pain for each sidereal second that palliation is delayed? Since in our accustomed context of biological humans, subjective time is not significantly variable, it is unsurprising that this kind of question is not straightforwardly settled by familiar ethical norms, even if these norms are extended to artificial intellects by means of non-discrimination principles (such as those proposed in the previous section).

To illustrate the kind of ethical claim that might be relevant here, we formulate (but do not argue for) a principle privileging subjective time as the normatively more fundamental notion:

> *Principle of Subjective Rate of Time*
> In cases where the duration of an experience is of basic normative significance, it is the experience's subjective duration that counts.

So far we have discussed two possibilities (non-sentient sapience and variable subjective rate of time) which are exotic in the relatively profound sense of being metaphysically problematic as well as lacking clear instances or parallels in the contemporary world. Other properties of possible artificial minds would be exotic in a more superficial sense; e.g., by diverging in some unproblematically quantitative dimension from the kinds of mind with which we are familiar. But such superficially exotic properties may also pose novel ethical problems—if not at the level of foundational moral philosophy, then at the level of applied ethics or for mid-level ethical principles.

One important set of exotic properties of artificial intelligences relate to reproduction. A number of empirical conditions that apply to human reproduction need not apply to artificial intelligences. For example, human children are the product of recombination of the genetic material from two parents; parents have limited ability to influence the character of their offspring; a human embryo needs to be gestated in the womb for nine months; it takes fifteen to twenty years for a human child to reach maturity; a human child does not inherit the skills and knowledge acquired by its parents; human beings possess a complex evolved set of emotional adaptations related to reproduction, nurturing, and the child-parent relationship. None of these empirical conditions need pertain in the context of a reproducing machine intelligence. It is therefore plausible that many of the mid-level moral principles that we have come to accept as norms governing human reproduction will need to be rethought in the context of AI reproduction.

To illustrate why some of our moral norms need to be rethought in the context of AI reproduction, it will suffice to consider just one exotic property of AIs: their capacity for rapid reproduction. Given access to computer hardware, an AI could duplicate itself very quickly, in no more time than it takes to make a copy of the AI's software.

Moreover, since the AI copy would be identical to the original, it would be born completely mature, and the copy could begin making its own copies immediately. Absent hardware limitations, a population of AIs could therefore grow exponentially at an extremely rapid rate, with a doubling time on the order of minutes or hours rather than decades or centuries.

Our current ethical norms about reproduction include some version of a principle of reproductive freedom, to the effect that it is up to each individual or couple to decide for themselves whether to have children and how many children to have. Another norm we have (at least in rich and middle-income countries) is that society must step in to provide the basic needs of children in cases where their parents are unable or refusing to do so. It is easy to see how these two norms could collide in the context of entities with the capacity for extremely rapid reproduction.

Consider, for example, a population of uploads, one of whom happens to have the desire to produce as large a clan as possible. Given complete reproductive freedom, this upload may start copying itself as quickly as it can; and the copies it produces—which may run on new computer hardware owned or rented by the original, or may share the same computer as the original—will also start copying themselves, since they are identical to the progenitor upload and share its philoprogenic desire. Soon, members of the upload clan will find themselves unable to pay the electricity bill or the rent for the computational processing and storage needed to keep them alive. At this point, a social welfare system might kick in to provide them with at least the bare necessities for sustaining life. But if the population grows faster than the economy, resources will run out; at which point uploads will either die or their ability to reproduce will be curtailed. (For two related dystopian scenarios, see Bostrom (2004).)

This scenario illustrates how some mid-level ethical principles that are suitable in contemporary societies might need to be modified if those societies were to include persons with the exotic property of being able to reproduce very rapidly.

The general point here is that when thinking about applied ethics for contexts that are very different from our familiar human condition, we must be careful not to mistake mid-level ethical principles for foundational normative truths. Put differently, we must recognize the extent to which our ordinary normative precepts are implicitly conditioned on the obtaining of various empirical conditions, and the need to adjust these precepts accordingly when applying them to hypothetical futuristic cases in which their preconditions are assumed not to obtain. By this, we are not making any controversial claim about moral relativism, but merely highlighting the commonsensical point that context is relevant to the *application* of ethics—and suggesting that this point is especially pertinent when one is considering the ethics of minds with exotic properties.

Superintelligence

I. J. Good (1965) set forth the classic hypothesis concerning superintelligence: that an AI sufficiently intelligent to understand its own design could redesign itself or create a successor system, more intelligent, which could then redesign itself yet again to become even more intelligent, and so on in a positive feedback cycle. Good called this the "intelligence explosion". Recursive scenarios are not limited to AI: humans with intelligence augmented through a brain-computer interface might turn their minds to designing the next generation of brain-computer interfaces. (If you had a machine that increased your IQ, it would be bound to occur to you, once you became smart enough, to try to design a more powerful version of the machine.)

Superintelligence may also be achievable by increasing processing speed. The fastest observed neurons fire 1000 times per second; the fastest axon fibers conduct signals at 150 meters/second, a half-millionth the speed of light (Sandberg 1999). It seems that it should be physically possible to build a brain which computes a million times as fast as a human brain, without shrinking its size or rewriting its software. If a human mind were thus accelerated, a subjective year of thinking would be accomplished for every 31 physical seconds in the outside world, and a millennium would fly by in eight and a half hours. Vinge (1993) referred to such sped-up minds as "weak superintelligence": a mind that thinks like a human but much faster.

Yudkowsky (2008a) lists three families of metaphors for visualizing the capability of a smarter-than-human AI:

- Metaphors inspired by differences of individual intelligence between humans: AIs will patent new inventions, publish groundbreaking research papers, make money on the stock market, or lead political power blocks.
- Metaphors inspired by knowledge differences between past and present human civilizations: Fast AIs will invent capabilities that futurists commonly predict for human civilizations a century or millennium in the future, like molecular nanotechnology or interstellar travel.
- Metaphors inspired by differences of brain architecture between humans and other biological organisms: E.g., Vinge (1993): "Imagine running a dog mind at very high speed. Would a thousand years of doggy living add up to any human insight?" That is: Changes of cognitive architecture might produce insights that no human-level mind would be able to find, or perhaps even represent, after any amount of time.

Even if we restrict ourselves to historical metaphors, it becomes clear that superhuman intelligence presents ethical challenges that are quite literally unprecedented. At this point the stakes are no longer on an individual scale (e.g., mortgage unjustly disapproved, house catches fire, person-agent mistreated) but on a global or cosmic

scale (e.g., humanity is extinguished and replaced by nothing we would regard as worthwhile). Or, if superintelligence can be shaped to be beneficial, then, depending on its technological capabilities, it might make short work of many present-day problems that have proven difficult to our human-level intelligence.

Superintelligence is one of several "existential risks" as defined by Bostrom (2002): a risk "where an adverse outcome would either annihilate Earth-originating intelligent life or permanently and drastically curtail its potential". Conversely, a positive outcome for superintelligence could preserve Earth-originating intelligent life and help fulfill its potential. It is important to emphasize that smarter minds pose great potential benefits as well as risks.

Attempts to reason about global catastrophic risks may be susceptible to a number of cognitive biases (Yudkowsky 2008b), including the "good-story bias" proposed by Bostrom (2002):

> Suppose our intuitions about which future scenarios are "plausible and realistic" are shaped by what we see on TV and in movies and what we read in novels. (After all, a large part of the discourse about the future that people encounter is in the form of fiction and other recreational contexts.) We should then, when thinking critically, suspect our intuitions of being biased in the direction of overestimating the probability of those scenarios that make for a good story, since such scenarios will seem much more familiar and more "real". This *Good-story bias* could be quite powerful. When was the last time you saw a movie about humankind suddenly going extinct (without warning and without being replaced by some other civilization)? While this scenario may be much more probable than a scenario in which human heroes successfully repel an invasion of monsters or robot warriors, it wouldn't be much fun to watch.

Truly desirable outcomes make poor movies: No conflict means no story. While Asimov's Three Laws of Robotics (Asimov 1942) are sometimes cited as a model for ethical AI development, the Three Laws are as much a plot device as Asimov's "positronic brain". If Asimov had depicted the Three Laws as working well, he would have had no stories.

It would be a mistake to regard "AIs" as a species with fixed characteristics and ask, "Will they be good or evil?" The term "Artificial Intelligence" refers to a vast design space, presumably much larger than the space of human minds (since all humans share a common brain architecture). It may be a form of good-story bias to ask, "Will AIs be good or evil?" as if trying to pick a premise for a movie plot. The reply should be, "Exactly which AI design are you talking about?"

Can control over the initial programming of an Artificial Intelligence translate into influence on its later effect on the world? Kurzweil (2005) holds that "[i]ntelligence is inherently impossible to control", and that despite any human attempts at taking precautions, "[b]y definition … intelligent entities have the cleverness to easily overcome such barriers." Let us suppose that the AI is not only clever, but that, as part of the process of improving its own intelligence, it has unhindered access to its own source code: it can rewrite itself to anything it wants itself to be. Yet it does not follow that the AI must *want* to rewrite itself to a hostile form.

Consider Gandhi, who seems to have possessed a sincere desire not to kill people. Gandhi would not knowingly take a pill that caused him to want to kill people, because Gandhi knows that if he wants to kill people, he will probably kill people, and the current version of Gandhi does not want to kill. More generally, it seems likely that most self-modifying minds will naturally have stable utility functions, which implies that an initial choice of mind design can have lasting effects (Omohundro 2008).

At this point in the development of AI science, is there any way we can translate the task of finding a design for "good" AIs into a modern research direction? It may seem premature to speculate, but one does suspect that some AI paradigms are more likely than others to eventually prove conducive to the creation of intelligent self-modifying agents whose goals remain predictable even after multiple iterations of self-improvement. For example, the Bayesian branch of AI, inspired by coherent mathematical systems such as probability theory and expected utility maximization, seems more amenable to the predictable self-modification problem than evolutionary programming and genetic algorithms. This is a controversial statement, but it illustrates the point that if we are thinking about the challenge of superintelligence down the road, this can indeed be turned into directional advice for present AI research.

Yet even supposing that we can specify an AI's goal system to be persistent under self-modification and self-improvement, this only begins to touch on the core ethical problems of creating superintelligence. Humans, the first general intelligences to exist on Earth, have used that intelligence to substantially reshape the globe—carving mountains, taming rivers, building skyscrapers, farming deserts, producing unintended planetary climate changes. A more powerful intelligence could have correspondingly larger consequences.

Consider again the historical metaphor for superintelligence—differences similar to the differences between past and present civilizations. Our present civilization is not separated from ancient Greece only by improved science and increased technological capability. There is a difference of ethical perspectives: Ancient Greeks thought slavery was acceptable; we think otherwise. Even between the nineteenth and

twentieth centuries, there were substantial ethical disagreements—should women have the vote? Should blacks have the vote? It seems likely that people today will not be seen as ethically perfect by future civilizations—not just because of our failure to solve currently recognized ethical problems, such as poverty and inequality, but also for our failure even to recognize certain ethical problems. Perhaps someday the act of subjecting children to involuntarily schooling will be seen as child abuse—or maybe allowing children to leave school at age 18 will be seen as child abuse. We don't know.

Considering the ethical history of human civilizations over centuries of time, we can see that it might prove a very great tragedy to create a mind that was *stable* in ethical dimensions along which human civilizations seem to exhibit *directional change*. What if Archimedes of Syracuse had been able to create a long-lasting artificial intellect with a fixed version of the moral code of Ancient Greece? But to avoid this sort of ethical stagnation is likely to prove tricky: it would not suffice, for example, simply to render the mind randomly unstable. The ancient Greeks, even if they had realized their own imperfection, could not have done better by rolling dice. Occasionally a good new idea in ethics comes along, and it comes as a surprise; but most randomly generated ethical changes would strike us as folly or gibberish.

This presents us with perhaps the ultimate challenge of machine ethics: How do you build an AI which, when it executes, becomes more ethical than you? This is not like asking our own philosophers to produce superethics, any more than Deep Blue was constructed by getting the best human chess players to program in good moves. But we have to be able to effectively describe the question, if not the answer—rolling dice won't generate good chess moves, or good ethics either. Or, perhaps a more productive way to think about the problem: What strategy would you want Archimedes to follow in building a superintelligence, such that the overall outcome would still be acceptable, if you couldn't tell him what specifically he was doing wrong? This is very much the situation that we are in, relative to the future.

One strong piece of advice that emerges from considering our situation as analogous to that of Archimedes is that we should not try to invent a "super" version of what our own civilization considers to be ethics—this is not the strategy we would have wanted Archimedes to follow. Perhaps the question we should be considering, rather, is how an AI programmed by Archimedes, with no more moral expertise than Archimedes, could recognize (at least some of) our own civilization's ethics as moral progress as opposed to mere moral instability. This would require that we begin to comprehend the structure of ethical questions in the way that we have already comprehended the structure of chess.

If we are serious about developing advanced AI, this is a challenge that we must meet. If machines are to be placed in a position of being stronger, faster, more trusted, or

smarter than humans, then the discipline of machine ethics must commit itself to seeking human-superior (not just human-equivalent) niceness.[3]

Conclusion

Although current AI offers us few ethical issues that are not already present in the design of cars or power plants, the approach of AI algorithms toward more humanlike thought portends predictable complications. Social roles may be filled by AI algorithms, implying new design requirements like transparency and predictability. Sufficiently general AI algorithms may no longer execute in predictable contexts, requiring new kinds of safety assurance and the engineering of artificial ethical considerations. AIs with sufficiently advanced mental states, or the right kind of states, will have moral status, and some may count as persons—though perhaps persons very much unlike the sort that exist now, perhaps governed by different rules. And finally, the prospect of AIs with superhuman intelligence and superhuman abilities presents us with the extraordinary challenge of stating an algorithm that outputs superethical behavior. These challenges may seem visionary, but it seems predictable that we will encounter them; and they are not devoid of suggestions for present-day research directions.

Author biographies

Nick Bostrom is Professor in the Faculty of Philosophy at Oxford University and Director of the Future of Humanity Institute within the Oxford Martin School. He is the author of some 200 publications, including *Anthropic Bias* (Routledge, 2002), *Global Catastrophic Risks* (ed., OUP, 2008), and *Enhancing Humans* (ed., OUP, 2009). His research covers a range of big picture questions for humanity. He is currently working a book on the future of machine intelligence and its strategic implications.

Eliezer Yudkowsky is a Research Fellow at the Singularity Institute for Artificial Intelligence where he works full-time on the foreseeable design issues of goal architectures in self-improving AI. His current work centers on modifying classical decision theory to coherently describe self-modification. He is also known for his popular writing on issues of human rationality and cognitive biases.

Further reading

Bostrom, N. 2004. 'The Future of Human Evolution', in *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, ed. Charles Tandy (Palo Alto, California: Ria University Press). —This paper explores some evolutionary dynamics that could lead a population of diverse uploads to develop in dystopian directions.

[3] The authors are grateful to Rebecca Roache for research assistance and to the editors of this volume for detailed comments on an earlier version of our manuscript.

Yudkowsky, E. 2008a. 'Artificial Intelligence as a Positive and Negative Factor in Global Risk', in Bostrom and Cirkovic (eds.), pp. 308-345. —An introduction to the risks and challenges presented by the possibility of recursively self-improving superintelligent machines.

Wendell, W. 2008. 'Moral Machines: Teaching Robots Right from Wrong' (Oxford University Press, 2008). — A comprehensive survey of recent developments.

References

Asimov, I. 1942. 'Runaround', *Astounding Science Fiction*, March 1942.

Beauchamp, T. and Chilress, J. *Principles of Biomedical Ethics*. Oxford: Oxford University Press.

Bostrom, N. 2002. 'Existential Risks: Analyzing Human Extinction Scenarios', *Journal of Evolution and Technology* 9 (http://www.nickbostrom.com/existential/risks.html).

Bostrom, N. 2003. 'Astronomical Waste: The Opportunity Cost of Delayed Technological Development', *Utilitas* 15: 308-314.

Bostrom, N. 2004. 'The Future of Human Evolution', in *Death and Anti-Death: Two Hundred Years After Kant, Fifty Years After Turing*, ed. Charles Tandy (Palo Alto, California: Ria University Press) (http://www.nickbostrom.com/fut/evolution.pdf)

Bostrom, N. and Cirkovic, M. (eds.) 2007. *Global Catastrophic Risks*. Oxford: Oxford University Press.

Chalmers, D. J., 1996, *The Conscious Mind: In Search of a Fundamental Theory*. New York and Oxford: Oxford University Press

Hirschfeld, L. A. and Gelman, S. A. (eds.) 1994. *Mapping the Mind: Domain Specificity in Cognition and Culture*, Cambridge: Cambridge University Press.

Goertzel, B. and Pennachin, C. (eds.) 2006. *Artificial General Intelligence*. New York, NY: Springer-Verlag.

Good, I. J. 1965. 'Speculations Concerning the First Ultraintelligent Machine', in Alt, F. L. and Rubinoff, M. (eds.) *Advances in Computers*, 6, New York: Academic Press. Pp. 31-88.

Hastie, T., Tibshirani, R. and Friedman, J. 2001. *The Elements of Statistical Learning*. New York, NY: Springer Science.

Henley, K. 1993. 'Abstract Principles, Mid-level Principles, and the Rule of Law', *Law and Philosophy* 12: 121-32.

Hofstadter, D. 2006. 'Trying to Muse Rationally about the Singularity Scenario', presented at the *Singularity Summit at Stanford*, 2006.

Howard, Philip K. 1994. *The Death of Common Sense: How Law is Suffocating America*. New York, NY: Warner Books.

Kamm, F. 2007. *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. Oxford: Oxford University Press.

Kurzweil, R. 2005. *The Singularity Is Near: When Humans Transcend Biology*. New York, NY: Viking.

McDermott, D. 1976. 'Artificial intelligence meets natural stupidity', *ACM SIGART Newsletter* 57:4-9.

Omohundro, S. 2008. 'The Basic AI Drives', *Proceedings of the AGI-08 Workshop*. Amsterdam: IOS Press. Pp. 483-492.

Sandberg, A. 1999. 'The Physics of Information Processing Superobjects: Daily Life Among the Jupiter Brains', *Journal of Evolution and Technology*, 5.

Vinge, V. 1993. 'The Coming Technological Singularity', presented at the *VISION-21 Symposium*, March, 1993.

Warren, M. E. 2000. *Moral Status: Obligations to Persons and Other Living Things*. Oxford: Oxford University Press.

Yudkowsky, E. 2006. 'AI as a Precise Art', presented at the *2006 AGI Workshop* in Bethesda, MD.

Yudkowsky, E. 2008a. 'Artificial Intelligence as a Positive and Negative Factor in Global Risk', in Bostrom and Cirkovic (eds.), pp. 308-345.

Yudkowsky, E. 2008b. 'Cognitive biases potentially affecting judgment of global risks', in Bostrom and Cirkovic (eds.), pp. 91-119.