

ON MACHINE ETHICS

J. STORRS HALL

ABSTRACT. Although moral concerns not particularly crucial for today's AI programs, designed as they are entirely by humans on whom the accountability rests, future AIs which learn and grow on their own are a different matter. The notion of providing a moral compass to a self-modifying mind is problematic; even if the moral module is exempt from modification, the rest of the system could change enough to render it irrelevant. Appropriate moral design requires finding properties that are invariant over the expected processes of change. We consider some possibilities and their implications.

STICK-BUILT AI

No existing artificial intellect has equalled, or even approached, human-level intelligence. Indeed it is fair to say that no AI program to date has exhibited intelligence at all, in the sense that the term is commonly used to apply to humans. A human exhibiting a notable skill, such as driving a car or playing grandmaster-level chess, will have learned the skill by a process of practice and internal concept formation. All computer programs currently exhibiting such skills have done no such thing. The hard, intelligence-requiring part of the formation of such skills has always been done by the human programmers.

Borrowing a term from the construction industry, we can refer to such programs as stick-built AIs. Since all the concepts used by a stick-built AI will have been provided to it by its designers, the problem of morality in a stick-built AI is simply one of competent design. All the appropriate rules can be formulated in terms of the concepts from which the program is built. As with its ontology, its goal structures and constraints come built-in.

AUTOGENOUS AI

An artificial intellect that was as intelligent as a human would be able to learn and grow mentally. It would create its own new concepts as a result of experience and its own skills as a result of practice.

With autogeny, however, comes the problem of moral constraint. We, the creators of the autogenous AI, cannot know what concepts it will use to understand the world, and thus we cannot write our rules in terms of them.

We can illustrate the problem with a simplified sketch of a cognitive architecture. Suppose an AI is composed of two major modules, a world model with the ability to do predictions, and an utility function over states of the world. In this sketch, the AI uses the model to predict the effects of its possible actions and uses the utility function to decide which to do.

As the AI learns and improves its understanding, it is improving its world model. In particular it is creating a richer ontology as it discovers more subtle regularities in reality. It can straightforwardly improve its model by means of experience – it

can evaluate possible modifications to the model by comparing their predictions to actual events.

The problem comes when it tries to improve its utility function. There is no obvious meta-function to guide modifications. Even so, the utility function may need to be updated as understanding of the world deepens.

Suppose, for example, the AI once believed that burnt offerings to the appropriate god would prevent lightning from striking its house. Subsequent events showed that lightning rods were a more efficacious prophylactic. Its utility function need not be updated, since the survival of the house remains the motivation. On the other hand, if the old utility function valued temples built to the gods in and of themselves, it would need to be revised.

(As an aside, we note that the only actual programs that operate by the “purely rational” method of the sketch are those dealing with extremely simplified domains, such as chess-playing. In practice, the rational ideal is usually computationally intractable; the vast majority of programs operate on heuristic “reflexes” which are generalizations from a model that, in most cases, the program itself does not explicitly contain. (This is of course true of humans as well.) Even so, we will ask the reader to assume for the sake of the argument that the heuristic forms are an approximation to the rational ideal.)

AI EVOLUTION

In the long run, AIs will evolve. Evolution in current stick-built AI is almost entirely the same human cultural evolution seen with any engineered artifact: the only thing a current, stick-built AI program can do to increase its likelihood of being copied in future generations is to work well.

However, once AIs become intelligent enough to have a significant role in the production of future AIs themselves, the logic of natural selection will take hold and the evolutionary pressure towards self- (and kin-) interest can be expected to become significant.

EXISTENTIAL RISK

In evolutionary history, the emergence of a new, more intelligent species is not without danger to the old one whose niche it desires. For example, from about 40 to 30 thousand years ago, anatomically modern humans appear to have systematically replaced the Neanderthals, leaving only fossils behind.

Among the treatments of the possibility of a similar replacement of humanity by AIs are Vinge [1993] and Bostrom [2003]. Vinge writes (after considering an optimistic scenario):

But in this brightest and kindest world, the philosophical problems themselves become intimidating. A mind that stays at the same capacity cannot live forever; after a few thousand years it would look more like a repeating tape loop than a person. ... To live indefinitely long, the mind itself must grow ... and when it becomes great enough, and looks back ... what fellow-feeling can it have with the soul that it was originally?

Bostrom is more sanguine:

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

The notion of humanity, much less “the good of humanity,” is however notoriously fuzzy at the edges (consider the abortion debate). Isaac Asimov began his career as an enthusiastic technophile, writing about robots and his Three Laws; by the end of his canon the robots had reinterpreted the part about not allowing humans to come to harm as requiring them, the robots, to prevent technological progress as a kind of secret police. If a single individual can change his ideas so drastically in a lifetime (and who among us has not?), the invariance of such an idea as “the good of humanity” is probably a frail reed to lean on as our sole support in the future.

Bostrom continues:

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

However, our “complicated mental ecology” was forged by evolution and it seems likely that the same thing will happen to AIs. As a starting point, while the evolution is still in a purely human-cultural phase, note that there are many proto-AIs with widely varying architectures, few of which could be said to correspond to Bostrom’s prescription. Existing selection pressures, i.e. criteria by which researchers choose the next architectures, are almost entirely pragmatic; the key issue today is to get the program to work at all.

AN INVARIANT

A more solid foundation for beneficence in future AIs might be found if we can discover some property that is invariant across the evolutionary process and base our design prescriptions on that.

The one principle that we can expect to find in any coherent utility function regards the operation of the AI itself. The system as a whole works only as well as the predicting ability of the world model. Maintaining or improving the grasp and range of the world model would be a necessary subgoal of virtually any other higher-level desideratum embodied by the utility function. Perhaps it was in a similar spirit that Socrates supposedly claimed that there is only one good, namely, knowledge; and only one evil, namely, ignorance.

Note that a goal of improving the knowledge in the world model does not have the same problems, or at least to the same degree, as some possible utility functions, because it can be couched in terms of the AI’s internal structure and operations instead of concepts relating to the outside world. A thorough ontology of its own

structure should be available to the AI, as an engineered construct, in a way that it is not available to us as (naturally) evolved humans.

Knowledge requires experience and processing power to digest it. These in turn require material resources. Thus the goal of knowledge seems to imply a certain amount of self-interest on the part of the AI. This isn't surprising: virtually any utility function that allows for states of the world better than the current one would motivate the AI to obtain the resources necessary to effect the change.

That, coupled with the fact that evolution itself selects for effective self-interest, is a strong indication that future AIs will, like humans, have self-interest as a core motivation.

THE MORAL LADDER

The reader will be familiar with the results of Axelrod [1984] in game theory involving the evolution of cooperation in iterated prisoner's dilemma tournaments. The TIT-FOR-TAT strategy predominated in an environment of strategies designed by human theorists. Subsequent work in evolutionary game theory has shown that this is part of a broader phenomenon: "nicer" strategies can predominate in environments of less nice ones, but only in small stages. In an environment where all the other strategies are effectively random, ALWAYS DEFECT is optimal. In an environment of all the two-state automata (of which TIT-FOR-TAT is one), the strategy GRIM, which retaliates endlessly if its opponent ever defects, is optimal. Once the environment has shifted enough toward niceness, strategies like PAVLOV or more forgiving versions of TIT-FOR-TAT can predominate over TIT-FOR-TAT itself.

The key mechanism in the emergence of cooperation is that cooperators form self-reinforcing groups. The individuals in these groups benefit from the cooperation. The bigger the groups become, the more beneficial it is for any other individual to become a cooperator and be able to join in the groups.

The reason this is germane to discussions of morality is that human morality, too, evolved. As with most natural evolution, earlier forms continue to co-exist with the newly evolved ones, so the structure of the resulting moral landscape resembles a wedding cake. Cooperative groups arise within an environment of non-cooperators. Slowly they grow and sometimes merge.

Autonomous agents, even ones as single-mindedly self-interested as game-theory tournament participants, are better off if they cooperate. The implication is one that appears to be at least compatible with most major formulations of metaethical theory. For example, application of the Categorical Imperative reduces the Prisoner's Dilemma to a choice between "both cooperate" and "both defect," which even a purely self-interested agent will make correctly.

LEVIATHAN

In the paragraph following the famous "nasty, brutish, and short" passage, Hobbes wrote,

It may seem strange to some man that has not well weighed these things that Nature should thus dissociate and render men apt to invade and destroy one another[.] Let him therefore consider with himself: when taking a journey, he arms himself and seeks to go well accompanied; when going to sleep, he locks his doors; when even in his house he locks his chests; ... what opinion he has of his

fellow subjects, when he rides armed; of his fellow citizens, when he locks his doors; and of his children, and servants, when he locks his chests. Does he not there as much accuse mankind by his actions as I do by my words?

We can see, perhaps, one example of a general improvement of the moral environment in that modern citizens do not generally consider it necessary to arm themselves when going on a journey. They are still well advised to lock their houses, however.

Over approximately the same period in human history, slavery has gone out of style. Over a longer period, using the Amazonian Yanomamo as a proxy for pre-historical hunter/gatherer cultures, we have come to where we are from a situation in which the average adult male had killed at least one other man.

OPEN SOURCE

It seems perfectly reasonable to suppose that stick-built AI will be such as to alleviate or at least minimize these concerns. Your robot vacuum cleaner will not be programmed to pilfer your household goods. As noted, at that level this is simply an aspect of competent design.

Now consider the situation of an intelligent, autogenous – and self-interested – AI. How can we ensure that it, even after extensive learning and self-modification, remains as trustworthy as the vacuum cleaner?

Essentially what we have to do is to give it an understanding of the evolution of morality. (Vacuum cleaners that learn to steal are unlikely to be a popular consumer item.) Self-selecting groups of cooperators drive the dynamic of the moral ladder. The key to joining a group of cooperators is not only being trustworthy, but being guaranteeably trustworthy. For humans, this is an unattainable ideal. For machines, it may not be. Protocols might well be developed that would allow an AI to guarantee certain aspects of its behavior; for example, releasing the source code to its utility function.

An AI that understood the evolutionary process it was part of could be reasonably expected to maintain invariants in its makeup that were evolutionary advantages or stable strategies in its milieu. Thus we should expect not only curiosity and self-interest, but the ability to be trustworthy, to be strongly conserved in future generations of AIs.

THE HORIZON EFFECT

The other key to beneficence is also a conservable trait in the appropriate environment. It is a long planning horizon for benefits – one is as concerned about the future as the present.

The horizon effect was first remarked as a weakness of chess-playing programs. If the program is looking ahead in the game a fixed number of moves, it can evaluate, say, taking a pawn with a queen just before the limit, and miss the fact that its queen will be captured just after the limit. Useful chess programs have heuristics to avoid this kind of blindness.

The more intelligent a creature becomes, and the more intelligent other creatures in its environment are, the more likely that “What goes around, comes around” is a valid generalization over the results of its actions.

Humans come with a built-in horizon effect; we don't care so much if our misdeeds will be discovered after we die. What's more, in the environment of ancestral adaptation, life was shorter and less certain than it is now, so the horizon effect was exacerbated.

Software does not have a fixed life span, and in an environment increasingly filled with other superintelligent programs, payback time, for good or ill, might well decrease from the human norm. What is more, an AI should expect that its peers, no less than itself, will be getting smarter as time goes by, and that it will be spending its future in a world more complex than its present. The probability of being able to cheat and get away with it indefinitely should become vanishingly small.

Having a long planning horizon should be a reasonably stable trait. Assuming it is backed up by enough intelligence that long-range plans actually work better than a series of short-term ones over the same period, it will be a clear evolutionary advantage in the long run. Furthermore, the plans need not be complete: the prediction that being honest will keep you out of certain kinds of trouble is more likely to be true than any detailed, specific scenario.

THE EXTENDED FAMILY

Throughout much of history, families tended to be power centers because innate kin-driven altruism tended to make them units of mutual cooperation. In the past few centuries the effect has declined because the culture has evolved to support reciprocal altruism-based cooperation in artificial groupings such as corporations and nations.

Even so, the present world contains significant conflict, much of it between groups of cooperating individuals. In contrast, the optimal unit of cooperation would seem likely to be all intelligent beings. It seems likely that, in time, superintelligent AIs would evolve out of any sectarian identification. There does not appear to be any reason that AIs would harbor a "machine-vs-human" tribalism, *except* inasmuch as some AIs would be intrinsically more trustworthy than humans and thus more worth dealing with. Even humans would find this to be so.

SUMMARY AND CONCLUSIONS

- Stick-built AIs, virtually all existing programs of note, should be carefully designed and built, but to a great extent, accountability for proper behavior on their part rests with their (human) designers.
- Autogenous programs, particularly ones approximating the rational ideal, stand in need of meta-functions to guide the development of their utility functions as they grow and evolve.
- Structural and evolutionary invariants are a promising approach to this problem.
- Curiosity and self-interest appear to be structurally conservable invariants. Self-interest is a strongly conserved evolutionary invariant, and trustability and long planning horizons seem likely to be conservable in the appropriate evolutionary environment.
- Thus if an AI is built with these traits, they are likely to remain present in their essence, even though improved in detail beyond human comprehension.

- Even though not all AIs will initially be built with this moral core (and humans are certainly not), it seems likely that if enough of them are, the existing human moral environment is good enough for them to form the cooperating cadres necessary to carry through the next step up the ladder of moral evolution.

The final conclusion is at once heartening and poignant. It seems likely that we are on the verge of creating beings who are as good as we like to pretend to be but never really are. We will be second-class citizens, not only intellectually, but morally, never able to join the cadres of truly, guaranteeably, un-self-deceivingly honest AIs. Even so, that world will be a better place than ours, and we can take pride in creating it.