**ETHICS OF ARTIFICIAL INTELLIGENCE**

# **OVERVIEW**

The possibilities of creating thinking machines raise to tons of ethical issues. The research paper outlines these problems and also highlight the possible solutions to them.

The first section discuss the issue with AI in near future.

The second section outlines challenges in ensuring that AI will operate safely as is approaches to human intelligence.

The third section outlines how we might assess whether, and in what circumstances AIs themselves have moral status.

In fourth section, we considered how AI might differ from humans in certain basic respects relevant to our ethical assessment of them.

The final section address the issues of creating AIs more intelligent than humans, and ensuring that they use their advanced intelligence for good rather than ill.

# **ETHICS IN MACHINE LEARNING AND OTHER DOMAIN-SPECIFIC AI ALGORITHMS**

## **Transparency**

Imagine in the near future, a bank using machine learning algorithm to recommend mortgage applications for approval. A rejected applicant brings 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 blinded to the race of the applicants. Indeed. That was part of the bank’s rationale for implementing the system. Even sp. Statistics show that the bank’s approval rate for black applicants has been steadily dropping. Submitting ten apparently equally qualified genuine applicants (as determine by separate panel of human judges) shoes that the algorithm is accepts white applicants and rejects black applicants. What could possibly be happenings?

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, which may enable an auditor to discover that the AI algorithm uses the address information of the 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 involve in designing machines. Designing a robot arm to avoid crushing stray humans is no more morally a fraught than designing a flame-retardant sofa. It involves new programming challenges but no new ethical challenges. But when AI algorithm take on cognitive work with social dimensions-cognitive tasks previously performed by humans, the AI algorithm inherits the social requirements. It would surely be frustrated 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.

## **Predictability**

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 of important function of the legal system is to be predictable, so that 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 with in which citizens can optimize their own lives.

## **Robust against Manipulation**

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, shape of that, placed next to a pistol in one’s luggage, would neutralize recognition of it. Robustness against manipulation in 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 system.

## **Responsibility**

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. 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.

## **Conclusion**

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 function; 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 a small sample for what an increasingly computerized society should be thinking about.

# **ARTIFICIAL GENERAL INTERLIGENCE**

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 even 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 Kasparove but even these researchers agree that something important is missing from modern AIs.

While this subfield of artificial intelligence is only just coalescing, “Artificial General Intelligence’. 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 sapience. A bee exhibit competence a 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 but human intelligence is surely significant more generally applicable than nonhumanoid intelligence.

It is relatively easy to envisage the sort of safety issues that may result from AI operating only with 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 engineer build a nuclear reactor. They envision the specific events that could go on inside it – valves failing, computer failing, cores increasing in temperature – and engineer the reactor to render these events non-catastrophic. 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 it to make toast- the purpose of the toaster is represented within the designer 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 un-envisioned context with an un-envisioned side effect.

Even task-specific AI algorithms throw us outside the toaster-paradigm, the domain locally preprogrammed, and 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 programmer 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 programmer themselves were not world champions, such as system would not have been able to defeat Garry Kasparove.

In creating a super human chess player, the human programmer necessarily sacrificed their ability to predict Deep Blue’s local specific game behavior. Instead, Deep Blue’s programmers had confidence that Deep Blues 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 chess rules. The prediction about distant consequences though it proved accurate, did not allow the programmer to envision the local behavior of Deep Blue – its response to a specific attack on its king – because deep blue computed the non-local game map, the link between a move and its possible future consequences, more accurately than the programmers could.

Modern humans do literally millions of things to feed themselves – to serve the final consequences 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 preference over final outcomes. Humans crossed space and put footprints on the Moon, even though none of our ancestors encountered a challenged 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 specially envisioned by either the designer or the user; including 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 that engineer never explicitly envisioned, one must specify good behavior in such terms of “X such that the consequence of X is not harmful to humans”. This is nonlocal; it involves extrapolating the distant consequences of actions. Thus, this is only an effective specifications – one that can be realized as 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 build 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 general intelligence operating on unforeseen problems, across domains, with preference over distant consequences. Inspecting the cognitive design might verify that the mind was, indeed, searching for the 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 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.

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 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 non-cognitive technologies, in that:

* The local, specific behavior of the AI may not be predictable apart from its safety, even if the programmer 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.