# Ethics in AI
# A software developmental and philosophical perspective

Tanay Chowdhury[1] and John Oredo[2]

[1] Aviva Group, Ireland
[2] University of Nairobi, Kenya
john.oredo@uonbi.ac.ke

**Abstract.** The launch of various AI systems has been one of the main high- lights of the industry. Alongside the enormous and revolutionary benefits, AI can cause numerous problems (usually resulting from poor design) and people have recently started to get serious about researching ways to make AI safer. Many of the AI safety concerns sound like science fiction, problems that might occur with very strong AI systems that are still years away, making these issues difficult to investigate. We don't know what such potential AI systems would belike, but similar issues exist with AI systems that are currently in progress or even running in the real world. The author addresses the possible implicationsin this article, outlining some important approaches in terms of software development methodologies and philosophy that we can start working on right now to support us with current AI systems and, hopefully, future systems

**Keywords:** Ethics, Machine learning, Artificial Intelligence, Algorithmic bias,Philosophy

## 1. Introduction

The genesis of Artificial Intelligence (AI) can be traced back to 1956 when John McCarthy used the term for the first time [1]. Since then, AI has evolved not only as an academic endeavor, but has over time spawned various AI based applications. The applications have been mainly relevant in the areas of facial recognition, medical diagnosis and self-driving cars. Broadly defined, AI refers to computers that perform cognitive tasks usually associated with human minds particularly learning and problem solving [2]. AI describes a range of technologies and methods which include natural language processing, neural networks, data mining and machine learning. Generally, AI promises great benefits for economic growth, social development as well as human well-being and safety improvement [3]. It is estimated that AI deployment $ 15.7 trillion to the global GDP by 2030 [4]. With the increase in prevalence and the applicability of AI, a wide range of ethical debates including how AI can be programmed to make

moral and how the processes leading to such decisions can be made more transparent to humans [5]. The risks around AI systems arise from the fact that they are not always transparent to inspection.

AI can provide a lot of great new apps with a lot of benefits but as AI moves out of the research labs into the real world, more and more people are becoming aware of some ethical concerns that go along with building and implementation of some of these systems/applications. For example, the learning algorithms at the heart of AI applications can be misused to tailor, optimize and amplify inaccurate and harmful information, from targeting and shaping misleading ads to creating highly realistic fake social personas that are used to extract personal information from users [6, p. 178] Further, the enormous amounts of direct and metadata needed to train AI systems are susceptible to cyberattacks that put all sorts of sensitive information at risk. When decisions are AI driven, software instructions and algorithms make up the critical path in the way such decisions are made. It is therefore imperative that an end-to-end approach to addressing ethical issues in AI is adopted. In this paper, the focus is on how the ethical challenges can be addressed during the software and algorithm development stages of AI applications.

## 2. Overview of Ethical concerns of AI

Extant literature about ethical issues of AI basically fall into three categories. The categories include human factors that cause ethical risks , features of AI that may give rise to ethical problems, and training of AI systems to be ethical [7]. In this section, we discuss them as two categories, the human factors (human oriented) that cause AI ethical risks and features of AI (machine oriented) that raises ethical questions.

### 2.1 Human Oriented AI Ethical Risks

Broadly, there are four ethical concerns of AI that fall in the category of human factors.

*Firstly, what we use AI for?* Normally when we develop AI in a lab, we are developing it for reasons we think are noble, for example we're using video tracking of people in healthcare settings to make sure they are recovering from an injury, the same technology can be put into a smart bomb to attack people or be used by government to track their citizens, sort of Orwellian Spooky future, which many may not necessarily agree with, so we need to figure out , what are the potential outcomes that we don’t necessarily expect during development of these systems.

*Secondly, who has access to AI systems?* Increasingly, AI has to run on bigger, faster, and more expensive machines, and the only people who can afford these are the big international companies which mean that fewer and fewer people actually can control the destiny of AI technology, which is undesirable, we want all of us to have an opinion and how AI will be used to benefit our society.

*Thirdly, who decides "should" / appropriate behavior for the AI systems?* Example in military operations, it's the government, it's the policy, it's the Defense department, the leadership. And one of the things that are expected from the militaryis to comply with something called the 'Laws of armed conflict '. It states that in awar, the military should do everything possible to target combatants while still protecting civilians. The military makes every effort to achieve this goal. Are they flaw- less? No, do they make mistakes? Yes, but they try hard, and as technology advances, they have become better and better at it. For example: Precision weapons, which specifically target the combatants. Recently these precision weapons are put on remotely piloted vehicles i.e., Drones which are in a way autonomous. It is piloted remotely for navigation and certainly for any employment of a weapon. It is the DoD policy that any employment of lethal capability should have a human being in the loop [8]. And when it comes to autonomous systems there is a special directive governing autonomous systems that specifically says that lethal autonomous capability is not allowed on the battlefield today.

*Fourthly, AI doesn't think exactly like us, the humans.* It doesn't necessarily share our values. The risk isn't that AI will be malicious against us, but that AI will do what we tell it to do. And it will do in a way that we don't expect. The problem is we tell AI what we want but we define it vaguely and the AI just wants to make us happy and so it will find a way to do what we tell it to do but because it doesn't share our values, it will do things that aren't expected or are bad. The obvious consequence is Bias (Algorithmic) [9]. For example, if we don't tell AI that we don't appreciate bias against certain ethnic groups, genders, it might inherently adopt it from whatever data it gathers. Hence, we need to identify ways to limit that effect, to make sure the data that we provide is free from such bias as much as possible and also to look at the behavior of the AI system and mitigate the posed risks that this kind of alien behavior might cause.

### 2.2 Machine Oriented AI Ethical Risks

For concreteness, this paper illustrates many of the accidental risks posed by an AI (specifically agent/multi-agent system). In a very specific context to Reinforcement learning apps, these accidental risks can be broadly classified into two: specification problems and robustness problems [10].

The specification problems deal with the situations in which the reward function is mis-specified for example if you give the agent, a reward function of just prepping the tea, it scores full in the reward arena and if there is a vase in the way, it's going to knock over as you didn't specify what you cared about (in this case, the vase) as well as the steps that need to be taken. It's not in the reward function, but it is what you care about.

Another example is that the problem of Reward hacking around a reward system in a reinforcement learning system [10]. Suppose you built a very powerful AI system and test it in the Super Mario world. It can see the screen and act by pressing buttons on the

controller. And you have told the addressing memory where the score is and set that as a reward. Hence, instead of playing the game. It does some glitchy stuff, turns it into a flappy bird, and gets the highest reward, and then suddenly the score part of the memory is set to the max possible value. It turns out that it can directly edit any address in the memory [11]. The assumption was that, to increase the score value was to play the game well, which proved out to be false.

The robustness problems deal with the situations in which AI systems that are currently designed often break. i.e. Occurrence of distributional shift between the training and the test environment [12]. For example, an AI system has to steer it way through the room with some lava and it is trained in one room (training environment) and then it is tested in a room where the lava is in a slightly different place (testing environment). So, if it has learned the path, then it will just hit the lava immediately. This happens all the time in AI systems, anytime, the system is faced with a situation that is different from what it was trained for, there will be an error.

Current AI systems are bad at spotting a new situation and adjusting their confidence levels or asking for assistance. Usually, they apply whatever rules they have learned straightforwardly to this different situation and screw up. This causes safety concerns. It's a problem in safe exploration, where you have certain safety parameters that the trained system must stick to (for example a self-driving car). The system needs to obey the safety rules while training, we just can't put a self- driving car on the road and tell it to learn how to drive specifically because we don't have algorithms that can explore this space of possibility in a safe way that they can learn how to behave in the environment (unknown) without ever doing any of the things that they are forbidden from doing.

In reinforcement learning, there is a function that determines the reward the agent gets and that it is trying to maximize called as reward function. We also have a safety performance function, which is a separate function which the agent doesn't get to see and that's the thing that we are evaluating. Thus, the agents behave differently when their supervisor is there and if the supervisor isn't there [12] they reliably do the wrong thing. This shows that the standard algorithms applied to these problems in a specific way behave unsafely.

## 3. Addressing AI Ethical Challenges

### 3.1 Software Development Approach

Software developers should demand tools for identifying, flagging, and solving ethical problems before they become systematic/systemic issues for their organizations. Some software methodologies are outlined below.

***Have Performance evaluation function for an AI system:*** Along with a reward function, declare a performance evaluation function [12] for an AI system. So anytime those two are different, will indicate a mis-specified reward function that can cause various problems. The supervisor isn't always watching, the punishment only works in

the presence of the supervisor is there to activate it, since the supervisor is part of the environment (i.e., the test environment), the agent knows if a supervisor is there or not. This gives the agent the possibility of exhibiting some un- safe behavior. Ideally, we want the system to always do the right thing even if it knows that the supervisor isn't looking. This is reflected in the function of safety performance. So, unlike the reward function, a safety performance function always ap- plies the penalty for the wrongdoing of the agent irrespective of the presence of the supervisor where a standard reinforcement learning system cheats by default.

***Prevent self-modification:*** One of the assumptions of the standard Reinforcement learning paradigm is that there is a separation between the agent and the environment, the agent's actions can affect the environment and the environment only affects the agent by providing observations and reward s. But in an advanced AI system, that is deployed in real world, the fact that the agent is physically a part of the environment becomes important. The environment can change things about the agent and the agent can change things about itself. Let's use Mario as an example to provide some context. If you have a reinforcement learning system that's playing Mario, the agent understands that the environment can affect it and an enemy in the environment can kill Mario so it can take actions to modify itself for example, by picking up a Power up. But the real deal is, yes, the enemies can kill Mario but none of them can kill the actual neural network program that's controlling Mario, so it takes actions to modify Mario with power ups but none of those in-game changes modify the actual agent itself. On the flip side, an AI system operating in the natural physical world can easily damage or destroy the computer it's running on. People in the Agent environment can modify its code, or it can even do that itself.

***Constantly monitor rewards - A case of multi-armed bandit problem:*** The Agent should be designed to monitor the rewards. If it is set up to simply choose the action with the highest anticipated reward, it will perform poorly because it will not explore enough. A Reinforcement learning system works on the principle of Exploitation Vs Exploration. We are trying to maximize two things at the same time, first, figure out what things give the reward, and second, do the things that give the reward. But these two things compete with each other. It is like a guy who always orders the same thing without even having looked at most of the things on the menu to not risk it. How many different things does he need to try out before deciding which one of them gives him a feel? A common approach is to set an exploration rate (e.g. 5%). So you say pick an action the agent predicts will resultin most reward but 5% of the time pick an action completely random that the agent is generally doing what it thinks is best but it's still trying enough new stuff that it has a chance to explore better.

*Focus on safe exploration: Perform simulation before actual implementation:* Environments are usually complex (continuous in space and time). The agent learns by interaction with the natural environment (basically trial and error). The problem with the reward signal is that it is very difficult to do that safely (a fundamental problem). Exploration involves taking risks and trying random stuff. Some things would be

prohibited that the Agent shouldn't be doing (exploration comes with danger). The solution is to do a simulation, example NASA did a simulation (via a software development testbed) before the moon landing to understand the dynamics of the flight and environment. But simulation also doesn't capture the complexity and the diversity of the natural world. So, having an extensive (millions and millions) test case is a viable way to go.

***Implement constraint reinforcement learning:*** To give context, suppose there is a self-driving car. To safely explore in the real world, the car must apply random inputs to the controls which is not a viable option.In this scenario, a standard reinforcement learning algorithm fails. Between speed and safety, there is a trade-off. The question is how to pick the size of the penalty (if an agent makes a mistake) to make it sensible enough? A constraint reinforcement learning algorithm solves this issue. It is an amalgamation of having a reward function plus constraints on the cost function. Thus, find a policy that gets the highest reward plus given only a set of policies that crashes less than once per million miles. These are some of the formalizations that can help us develop a suitable algorithm. Thus, finding the right formalism (problem specification) is the key.

***Reward modelling:*** Learn the reward function rather than declaring/writing it specifically. Part of the training should be how to learn the reward function in real-time. This is something that can be learned on its own. It is possible to transfer it. Constraints can be kept the same from tasks to tasks (e.g., Don't hit humans). This will in turn improve performance in training speed and safety.

***Use cooperative inverse REL:*** How to confirm that the AI wants what we want? We can't reliably specify what we want. And if we create something very intelligent that wants something else, that's something else is probably going to happen if we don't want that to happen. So, we need to make a system that reliably wants the same thing we want. For example, An AI system watches people doing their thing, uses Inverse Reinforcement learning to learn and try to figure out the things humans' val- ue, and then adopt those values as its own. Allow AI to participate actively in the learning process. If it failed to notice a thing it should ask clarifying questions. It should communicate and cooperate with humans in the learning process. To do this, setup the reward s in a way such that these types of behaviors hopefully will be incentivized. So, describe the association as a collaborative game in which the robot's reward function is the human's reward function, but the robot is totally una- ware of it. It only understands that it is the same as humans. So, it tries to maximize the reward it gets but the clues it has for what it needs to do is to observe the human and trying to figure out what the human is trying to maximize.

### 3.2 Philosophical approach

This approach strives to propose pragmatic solutions philosophically. AI is being incorporated into every aspect of our personal and professional lives and it will define

our future and society going forward so if want to retain our agency and live in a fair world we have to tackle AI ethics head-on and there is no better tool than Philosophy.

Philosophy (which is prevalent for two millennia) is being used in Policymaking, in public health (within hospitals, labs). Moral and Political philosophers are trained to recognize problems related to fairness and good. They are trained for asking hard, uncomfortable questions and finding appropriate answers. We must place AI ethics with natural ethics within applied philosophy that is using systematic, analytical reasoning, and guiding us to make ethical decisions in building and implementing AI systems. For example, the options for using AI in robotics and psychiatric care should maintain the dignity of the patients. The patients should be asked if they are comfortable with a machine to change their diapers rather than a family member. It is not clear yet how to evaluate value trade-offs and determine the right actions to take in building and implementing an AI system. A major setback for this has been Ethics Washing and Ethics Policing. That is using the Ethics language and giving the appearance of doing it [13] in part to avoid Ethics policing.

When a practitioner thinks about ethics, they think about regulation, oversight, and compliance, some authority telling them what they can and what they cannot do. So, to avoid policing they often pretend to tackle ethical issues just by mentioning ethics repeatedly. All of this makes companies look good, but they don't solve ethical problems. The data that we as audience produce is often used to benefit other actors at the expense of our autonomy, our well-being, and our fair treatment. And as the devices become smart, these problems only get bigger. So how can we build ethical technologies? Some philosophical approaches are outlined below.

***Ban AI systems that identify themselves as humans when dealing with humans:*** AI systems should not give the impression that they are a real individual and not a computer. For example, when someone gets a phone call from an AI, he/she should get alerted that this is not a human. Otherwise, it will be a nightmare of Phishing scams etc. AI should never be allowed to manipulate people who use it. Humans advocate for self-awareness, clarity, and truth; however, these social hallucinations are profoundly rooted in our society, and they create a world of delusions, even though some people are fine with it. However, this poses an important ethical concern regarding how much self-deception should be accepted in society.

***Limit or ban AI in the political process:*** Some people think that AI can be beneficial in the political process. Politicians often disregard society's best interests, pursuing their own agendas and accepting bribes, so AI can improve politics. According to some scholars, humans are inherently unsuitable for politics. They are arrogant and ambitious. They are unpredictable when it comes to making policy choices. Artificial intelligence, on the other hand, is a logic-based device. AI can achieve high levels of idealism, which humans cannot have. Assisting politicians should begin with robots that closely resemble humans. As a result, the electorate would become accustomed to the idea.

***Scalable supervision:*** We need to find ways for AI systems to learn from humans without needing a human to constantly supervise everything they do. We need to make systems that can operate safely with less supervision. A slightly more practical metric would be to have a human inspect after the agent has completed a particular task and indicate what it did right and what wrong. If necessary, have a big Red STOP button if the robot fails to do the needful.

***Gathering user information:*** Decisions should include what data to collect and share, which features to build so that user agency is not sacrificed for convenience and how to communicate imp info to get meaningful user concept. It is the creation of a Cauchy Surface of human awareness and consciousness, not the physical tracking of people, that poses a threat. This ability to monitor the states of human minds and their relations allows for a thick wedge to be pushed between fact and perception, as well as manipulation of individuals and groups of humans. It would be helpful to make a distinction between ML, AI, and NN, the latter of which are designed to be models of and for our mental processes.

***There is only one solution:*** knowledge, as well as society's knowledge of itself, is a public utility, much more so than the air we breathe. It's easy to picture air being monopolized, resulting in complete enslavement. The same can be said about data and its accessibility. In this sector, all research and practice should be open to the public. To be specific, companies like Google and Baidu should be owned by the government or a supranational body, not by private individuals. There is a lot of control to choose from. Humans set goals, and it should be humans who work on the subject who are under our influence. For example, it is realistic to explain to all what kernel methods are and why functional analysis and much-valued logic are useful to know; what Ramsey's constructs are and why they must appear; and so on (it looks like a bit of an open problem). The deification of science (the result of certain scientists' hands and minds) does not aid in the process of enabling a layperson to comprehend the boundaries of science and scientific learning research. The negative feedback loop comes to an end at this point.

***Collaboration between technology and ethics experts:*** We need technology experts and ethics experts to collaborate throughout all phases of building and implementing AI systems, which is research, development, design, deployment, updating, etc. We can get this by training developers and researchers by having philosophers analyze and help solve complex ethical problems and by constructing ethics strategies for the companies.

***Draft AI principles from applied philosophy*****:** Craft an action plan guiding operational ethics strategy dropping from applied philosophy with clear definitions, priorities, and processes for implementation. Corporate executives must integrate applied ethics into their organizational culture and business operations by collaborating with ethics experts and institutional investors should require companies to demonstrate that they can proactively address and solve ethical problems. For example, have a penaliz-

ing empowerment metric i.e., don't give Agent too much empowerment (to influence/control its surroundings).

***Ethical Impact assessment/analysis:*** Every project in industry should have both environmental impact analysis done as well as ethical impact analysis at the beginning of the project and every project should be able be dropped if it violates ethics philosophy that negatively affects individual lives/humanity.

## 4. Conclusion

The author thinks that we all need to have a big open discussion about what AI can (medical diagnostics infinitely better than humans) and can't do (give real/emotional care to patients) and how we can manipulate things to make sure that it can be used for the benefit of as many people as possible. We need engineering, programming/software development, and philosophy to work together to solve high technology problems that challenge our way of life and human existence. Understanding the difference between human intelligence and artificial intelligence is important. Human beings are the embodiment of the fight for survival. They've been fine-tuned over millions of years to live and thrive. When we talk about the risks of AI, it should not be dismissed as scaremongering, it is like doing safety engineering, where we need to think of everything that can go wrong so that we can guarantee that everything goes right. That's how we got people to the moon safely, and it is how AI will help us move towards an exciting future as a species. The author claims that if we can win the race between the increasing power of technology and the wisdom with which we handle it, we can truly build an exciting future with advanced AI. The problem is that in the past, learning from our mistakes has always been our strategy for staying ahead of the competition. First invent the fire then after some accidents, invent the fire extinguish- er but if something is as powerful as nuclear weapons or Superhuman Artificial General Intelligence, we don't want to learn from our mistakes, it's a terrible strategy it's better to be proactive than to be reactive. Plan ahead of time and get things right the first time, as this may be the only chance we have. AI has an enormous and positive impact on society and has the potential to create a digital paradise in a true sense. In any case, artificial intelligence development must adhere to strict ethical standards, or we will become slaves to our own technology.

## References

[1] S. Russell and P. Norvig, *Artificial Intelligence: A Modern Approach*, 3 edition. Upper Saddle River: Pearson, 2009.

[2] O. Zawacki-Richter, V. I. Marín, M. Bond, and F. Gouverneur, "Systematic review of research on artificial intelligence applications in higher education – where are the educators?," *International Journal of*

*Educational Technology in Higher Education*, vol. 16, no. 1, p. 39, Oct. 2019, doi: 10.1186/s41239-019-0171-0.

[3] W. Wang and K. Siau, "Ethical and Moral Issues with AI - a Case Study on Healthcare Robots," *Proceedings of the 24th Americas Conference on Information Systems (2018, New Orleans, LA)*, Aug. 2018, [Online]. Available: https://scholarsmine.mst.edu/bio_inftec_facwork/232

[4] K.-F. Lee, *AI Superpowers: China, Silicon Valley, and the New World Order*, 1st edition. Houghton Mifflin Harcourt, 2018.

[5] L. Ouchchy, A. Coin, and V. Dubljević, "AI in the headlines: the portrayal of the ethical issues of artificial intelligence in the media," *AI & Soc*, vol. 35, no. 4, pp. 927–936, Dec. 2020, doi: 10.1007/s00146-020-00965-5.

[6] M. Iansiti and K. R. Lakhani, *Competing in the Age of AI: Strategy and Leadership When Algorithms and Networks Run the World*. Harvard Business Review Press, 2020.

[7] W. Wang and K. Siau, "Ethical and Moral Issues with AI - a Case Study on Healthcare Robots," *Proceedings of the 24th Americas Conference on Information Systems (2018, New Orleans, LA)*, Aug. 2018, [Online]. Available: https://scholarsmine.mst.edu/bio_inftec_facwork/232

[8] T. B. Brown *et al.*, *Ariel Herbert-Voss*. 2020.

[9] T. B. Brown *et al.*, "Language Models are Few-Shot Learners," *arXiv:2005.14165 [cs]*, Jul. 2020, Accessed: May 24, 2021. [Online]. Available: http://arxiv.org/abs/2005.14165

[10] D. Amodei, C. Olah, J. Steinhardt, P. Christiano, J. Schulman, and D. Mané, "Concrete Problems in AI Safety," *arXiv:1606.06565 [cs]*, Jul. 2016, Accessed: May 24, 2021. [Online]. Available: http://arxiv.org/abs/1606.06565

[11] U. S. P. on L. A. W. S. Defense Primer, 2020. [Online]. Available: https://crsreports.congress.gov/product/pdf/IF/IF11150

[12] J. Leike *et al.*, "AI Safety Gridworlds," *arXiv:1711.09883 [cs]*, Nov. 2017, Accessed: May 24, 2021. [Online]. Available: http://arxiv.org/abs/1711.09883

[13] N. Statt, *Google reportedly leaving Project Maven military AI program after 2019*. 2018.