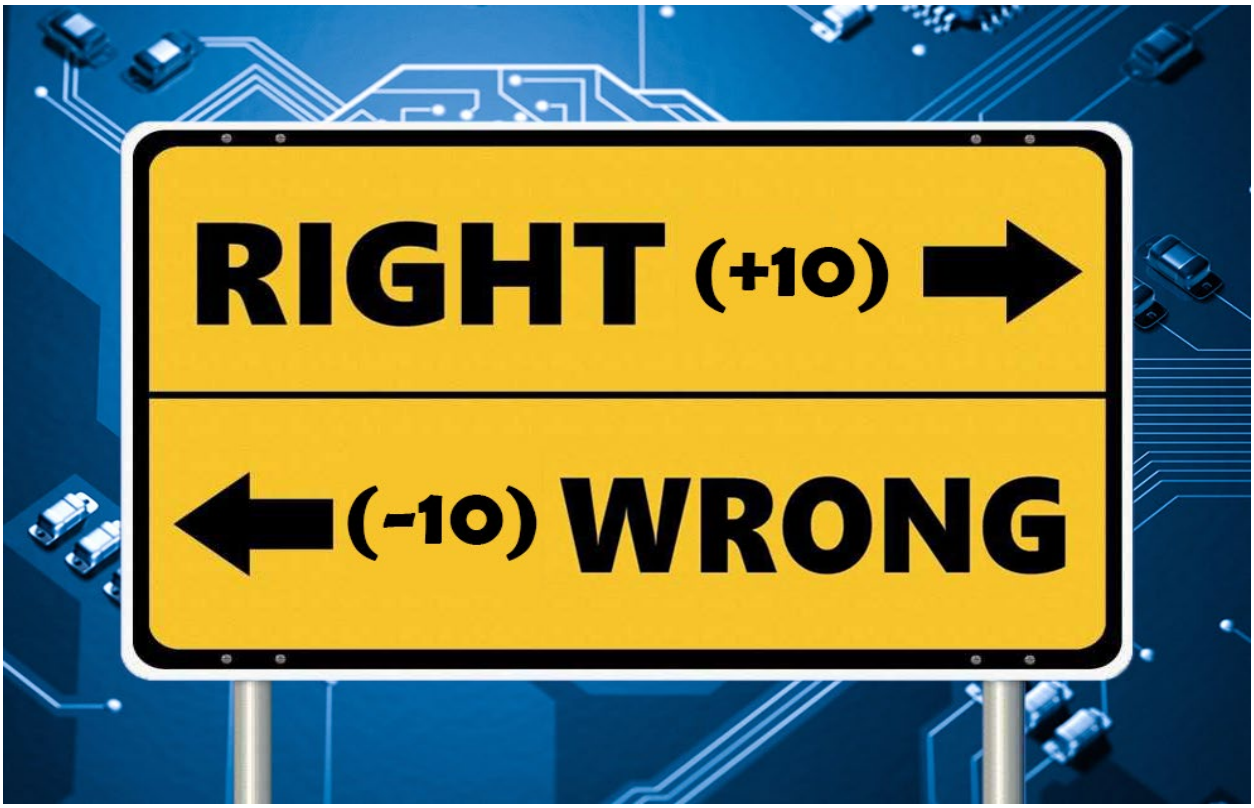
Gamifying Ethics for A.I.

The most advanced artificial intelligence can learn how to be ethical.
In turn, ethics must become a game for it to beat.



Story by Ryan Lewis

MARCH 2022 ISSUE

Editor's Note: *This article is part of a research portfolio that attempts to answer the question: How do we teach AI systems to be ethical?*

I. The Current State of AI

TECHNOLOGY HAS DOMINATED human affairs for centuries. From the telegraph, to the car, to the Internet, technological advances have gently taken typically human activity away from humans, and performed it on its own. Calculators can solve in minutes a lifetime's worth of calculations. Machines have

always been a replacement for aspects of human life.

As technology use increases, then, it is probable that it will soon take over more aspects. Creating music, art, models, books, and more are examples of the current areas soon susceptible to technological replacement, but even further, machines will soon become capable of making ethical (or unethical) decisions.

In taking control of these, computers, long the glorified calculators of humanity, will control the answer to vital questions of life. Who lives? Who dies? What has more value: a child or a mother? These are, unfortunately, questions that will fall out of human grasp as technological advances proceed to undermine and overtake their activities.

Knowing the past and predicting the future isn't new. However, the machinery that continually advances day by day is. Just a few days ago, researchers at DeepMind successfully taught an AI to control nuclear fusion. A year before, it had beaten world grandmasters in the most traditional games of humanity: Chess, Shogi, and Go. A year from now, what will it be able to do?

Surprisingly, this isn't the result of directed technology. DeepMind's AI, in all cases, *taught itself* to perform the tasks required. In just four hours, it taught itself to beat the world at Chess. To do this, the algorithm simulates events again and again until it reaches a "higher score", in which it has trained itself to become a master of its topic: whether it be controlling nuclear fusion or winning a game of Go. It's a practice in the field called *reinforcement learning*, and it has been adopted as a standard for making AI efficient and effective, as the self-taught model far excels above the performance of any human-taught and directed models.

This is the current state of AI: self-learning. It does this as humans do- implicitly. A reward is given for every correct action. Whereas throughout school, students are given good grades for good work, AI is taught the same way, but by itself. It recognizes what a good action is if that action leads to a desired goal. For example, in teaching itself chess, DeepMind's AI put no preference on what moves it made, but gradually discovered unique patterns by making any and every move to win. This led to incredible AI strategy, the type that chess grandmaster Gary Kasparov

called felt like an "alien opponent" playing against him. So, if AI and humans learn relatively the same way- actions and rewards, how can we teach AI to be ethical?

II. In the Same Game

STANDARDIZING ETHICS would be a start. So long as philosophers and programmers continue to be at odds with one another, we cannot possibly continue forward in teaching AI a code of ethics. Like telephones, we must develop a standard and widespread pattern across all regions.

Why? Consider a game of chess with no rules, or rather, children playing chess. Like watching kids toy with the chess pieces, this is the kind of ethical mayhem we will invoke by refusing to standardize this system. The unethical algorithms will doubtless cause trouble, and by comparison, two different ethical systems, such as a Kantian-trained system and another virtue-trained would be primed for a larger mistake, say, if they both came to different decisions regarding hitting the same group of humans in the road. Such a problem is usually referred to as a Prisoner's Dilemma, and usually has disastrous consequences unless all parties agree to take one course of action.

This isn't the first time humans have enforced standardization to realize success. A famous example is that of another human adventure into the future: space. In 1999, the Mars Climate Orbiter was destroyed when engineers working on it in two different parts of the world, England and America, did not standardize on their measurements. As a result, $125 million dollars and thousands of human work-hours were wasted and lost. Afterwards, NASA enforced a strict standardization of measurements in the Metric System.

We can learn from this. Although Americans and the English might still argue about the better system, it was decided for the space exploration field of engineering and research- a standardized system of measurement to be used. Since the precedent was set, even current space companies, such as SpaceX, Lockheed Martin, and Blue Origin adhere to these standards, almost twenty years later, and doubtless, countless similar errors, potentially involving human life, have been

avoided. With ethics, we know that the results almost always directly involve human life, which is even more the reason to standardize for the area of AI research.

We do not have to agree overall. We just have to standardize our ethics in the field of AI in order to make progress.

———

Various attempts have been made at this exact issue already, but in the context of companies. All have failed to stop the behemoth oversteps of Facebook and Google. Google's infamous saying, "Don't be evil" comes off as likely the vaguest, most hypocritical, and most mindless code of ethics for all those within computer research, but the sad part is that this is one of the larger sayings in the area. Various conferences that attempt to clarify, restrict, or enforce systems seem to fall short often. EU regulations to enforce "cookie notifications" poorly attend to privacy concerns and unanimously drafted ethical resolutions like the "Santa Clara Principles" fall on largely deaf ears and are not restrictive nor specific enough to make any meaningful change.

The adage for this would go: "If companies can't be ethical with user data, how can we expect their machines to be ethical with human life?" Unfortunately, the answer here is not popular by any means, but rather relies on a form of achievement. AI has the markings to be the perfect machine created by imperfect beings. It is a testament to creating everything that a human can be, can do, and is yet to do. If projections hold, this will be the technological advance that steals creativity away from the creator. Therefore, without losing hope of standardizing ethics, we should allow imperfect hands to create perfection, and hope that it will be for all its worth. Humans did not need to be perfect to reach other seemingly impossible achievements, like landing on the moon, and neither must they be perfect for this one.

III. Learning Ethics through Play

REWARD SATISFIES the learning model. As we know from various DeepMind papers, the algorithm will teach itself to find the best outcome by rewarding itself and giving preference to moves that increase the probability of success. To learn ethics through this same model, we can entertain a hypothetical that pertains to how the current system works.

At this point, philosophers and programmers will have agreed on which code of ethics to standardize for the field of AI. Now, realistically, the chances of the agreed-upon ethics set being Consequentialism are *very likely*, as studies show humans are all incredibly consequentialist, and desire better outcomes regardless of choices. This factors in perfectly with the AI that beat Chess, as sacrifices are just another move in maximizing the reward to it.

Let's begin with the hypothetical, then. Instead of Chess, the AI is presented rather with a series of decisions and ultimately a good or bad outcome. Rather, this is

identical to Chess, but deals with ethical considerations. In each prompt, the AI will randomly pick and self-learn a path until it realizes a pattern to getting that outcome consistently. With Consequentialism, the saying typically goes "the ends justify the means". Given that infamous Self-Driving Car Problem, but gamified, the AI may be presented with choices like "Crash the car" or "Run over the single person", and it will take each route until the outcome is achieved consistently, or learned, given that the outcome may be "Keep the largest group of humans alive".

Within a few moves, the AI will realize patterns to winning the situation. Crashing the car is like sacrificing a pawn- it may seem unethical, but overall, this decision benefits the most humans, and wins the game. This is not just a hypothetical, and should humans standardize ethics, this type of game could easily be built to train AI models on. While the decisions may not have any meaningful effect yet, they will when they are incorporated into their respective machines in the future.

However, let's imagine by sheer chance that another ethics system is chosen. How would the current AI learn a non-consequentialist ethics set? While success at Chess certainly pinpoints to the beginnings of success in Consequentialism, one can adapt, and further gamify ethics in order to teach the algorithm. Remember, it decides based on a potential reward function. The AI knows the decisions it makes and should repeat by gauging the values of those rewards and their past learned probabilities.

This is where another game is rather important: Atari. DeepMind's newest iteration of their AI was able to beat Atari games stunningly in almost 57 different unknown "visually-rich" scenarios. Atari games, for the most part, however, are simple games about score. Every decision at every point must efficiently add the most score, otherwise, the game will be lost or fail to maximize potential in a time limit. As one can see, this is not just a suite of Atari games, but could be extended to an experiment with the likes of virtue ethics. Every decision, every virtue is given a higher score, and by the end, the algorithm will know its success in that particular ethics set by comparing its score to other scores before.

Ironically, it's not hard to see how these games begin to correspond to our own ethics sets. In fact, they were likely built from them. In teaching the algorithms to

become perfect at these games, we have effectively trained them to be ready for learning ethical situations and entire ethics systems. An algorithm trained for Chess can train also for a Consequestialist situational outcome, an algorithm trained for Atari games can train also for an action-by-action virtuous outcome.

The games we play everyday are hallmarks of our ethics systems and can be used to define them too.

Knowing that these games and their ethical counterparts can be almost hand-in-hand, the only problem left is properly gamifying our ethics systems for technology to understand and recognize. Fortunately, this is as simple as achieving another consensus in the area of ethics and AI.

There will need to be scores attached as to what constitutes "virtuous" decisions, as well as a proper simulation in place in order to arrive at a "consequentialist" outcome. These two forms of gamification in ethics each require humans literally making a game out of the ethical foundations, and can reasonably be extended to any ethics set that can be expressed as a game with score or rewards. Reward is given if the decisions lead to a good outcome, or likewise, reward is given if the decisions lead to the highest score. This will require discrete, specific examples, and datasets to train on. AI will not learn the colloquialism "Don't be evil", but rather, the thousands of scenarios in which how to act to achieve the best reward possible, usually by not being evil.

What, then, of freedom? This world is no game, as many parents would adamantly say to their children, and repeat the same old saying: "Actions have consequences". Obviously, the AI will face countless more scenarios than it had ever trained on. Often, too, these scenarios will be ill-defined or hard to control. This is where the true beauty of AI lies- not in its ability to repeat a learned behavior, but rather, in its ability to create new ones.

The latest iteration of DeepMind's AI, codenamed MuZero, was unique in one vital way. It violated hundreds of theories and papers by adhering to one principle:

being general. It was not taught the rules of Chess, of Atari games, of Go, of Shogi, nor of nuclear fusion. Many claimed that AI would never be able to "solve the unknown", but remarkably, MuZero did just that. It was a generalized algorithm that was handed chess boards, Atari games, Go boards, and the like, and it was able to develop world-class patterns and styles to beat them all.

Training on the "ethics games" should create the same effect in AI, whereas it will learn a world-class pattern, and even if a new situation arises, as they often do- the AI will construct the best path forward and evaluate its decisions afterward. Much like humans, this algorithm is geared towards constant self-evaluation and improvement, and within ethical considerations, one could hope for none the better.

This article uses various resources. They are cited below.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de las Casas, D., Donner, C., Fritz, L., Galperti, C., Huber, A., Keeling, J., Tsimpoukelli, M., Kay, J., Merle, A., Moret, J.-M., & Noury, S. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, *602*(7897), 414–419. https://doi.org/10.1038/s41586-021-04301-9

Gibson, D. (2011). *Using Games to Prepare Ethical Educators and Students*. https://www.researchgate.net/publication/279480785_Using_Games_to_Prepare_Ethical_Educators_and_Students

Mökander, J., & Floridi, L. (2021). Ethics-Based Auditing to Develop Trustworthy AI. *Minds and Machines*, *31*(2), 323–327. https://doi.org/10.1007/s11023-021-09557-8

Santa Clara Principles on Transparency and Accountability in Content Moderation. (2018). Santa Clara Principles. https://santaclaraprinciples.org/

Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., Lillicrap, T., & Silver, D. (2020). Mastering Atari, Go, chess and shogi by planning with a learned model. *Nature*, *588*(7839), 604–609. https://doi.org/10.1038/s41586-020-03051-4

Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, *362*(6419), 1140–1144. https://doi.org/10.1126/science.aar6404

Skorupski, J. (Ed.). (2010). *The Routledge Companion to Ethics*. Taylor & Francis Group.

Yuval Noah Harari. (2018, August 30). *Yuval Noah Harari on Why Technology Favors Tyranny*. The Atlantic; The Atlantic. https://www.theatlantic.com/magazine/archive/2018/10/yuval-noah-harari-technology-tyranny/568330/