Computation, Rule Following, and Ethics in AIs

Hyunjin Seo
University of Kansas
hseo@ku.edu

Stuart Thorson
Syracuse University
thorson@syr.edu

Abstract

As interest in developments of artificial intelligence (AI) models has grown, so has concern that they embed unintended, undesirable risks and/or fail to properly align with human values and norms. In the extreme case, it is argued that AI may pose existential risks to the human species. We consider entities satisfying the Extended Church-Turing Thesis and claim these include both human and non-quantum based AI. We then introduce rules, including moral and ethics rules, as linguistic entities and illustrate how they can be encoded and treated computationally. Following Wittgenstein, we show that rules and rule following cannot be purely private. Whether particular rules are being followed in specific instances depends upon ongoing engagement with a language community. However, in situations involving application of ethics rules there may be no widely agreed community to use in evaluating whether rules are being followed properly. Indeed, how are we to determine which are appropriate ethical rules? Every appeal to rule following itself is based upon more rules; it is rules all the way down. Deliberative reasoning is at the core of moral and ethics discourse and issues in conceptualizing rule-following AIs become of particular interest.

1. Introduction

In her entertaining book of essays [1], the playwright Jean Kerr relates her trials raising children in the early 1950's. In the title essay, she describes preparing for a dinner party she was hosting by first instructing her three young boys with a list of behavior rules including, "not to go in the living room, not to use the guest towels in the bathroom, and not to leave the bicycles on the front steps." However, when she checked the dinner table just prior to arrival of guests she found that her floral centerpiece was now just three stems. She had neglected to include as a rule that her sons not to eat the daisies on the dining-room table.

This example illustrates the broader point that we face in behaving in a world where we frequently encounter situations seemingly unlike any we have faced in the past. For Kerr's children, such a situation was the vase of daisies on the dining room table. For a juror, it might be deciding whether to find a person guilty of a crime for which the death penalty was a possibility. For a military officer, it could be determining whether blips on a radar screen constitute early warning of a missile attack. Ethical principles or rules may provide guidance for behavior in such situations but, as with Kerr's children, they generally require not only that the literal text of the rules be known but also that the so-called spirit or intent of those rules be understood. Behind literal principles will be additional rules for determining how and when they should be applied in any specific situation.

That ethical rules alone can be inadequate has long been acknowledged. Kierkegaard [2] famously wrote about Abraham hearing the voice of God instructing him to sacrifice his son, Isaac and introduced his notion of the *teleological suspension of the ethical* to account for Abraham's message from God taking priority over the ethical imperative not to kill one's child. As Schrag [3] notes, "The ethical has here to do, in short, with the universal (or quasi-universal) norms and principles which are accepted by a group and set forth as proper guides or directives for moral action." Ethics are shared within a community of practice; sanctioning involves taking away community membership (e.g. losing medical license, disbarment, or being sent to prison). Of course, different groups of people may differ on ethical principles and their application in particular contexts. Consider, as examples, disagreements over abortion, capital punishment, and even the wearing of face masks during a public health crisis.

The public nature of ethical rules arises from their being expressed in language.[1] Examples of such rules,

[1] Here we invoke Wittgenstein's argument against the existence of a private language. This will be developed in our formal discussion of rules in Section 2.2.

URI: https://hdl.handle.net/10125/79464
978-0-9981331-5-7

written in English, might be "Avoid being violent when possible," "Avoid killing human beings," or "Wear a face covering when around other people." While these rules, as do most ethical rules, take a linguistic form, it is not the case that they can generally be assigned truth values. Rather, they exemplify what C.S. Pierce termed practical reasoning. In his words, they involve "What am I prepared deliberately to accept as the statement of what I want to do? What am I to aim at, What am I after? To What is the force of my will to be directed?" (quoted in [4]).

Ethics rules can help us deliberate and plan for future behavior in as yet to be experienced situations by leaving open bindings of key terms. When is it possible to avoid being violent? What does it mean to be violent in a specific situation? Is it violent to yell at someone? Is it violent to restrain someone? Answers to such questions will depend upon particular contexts or environments in which we consider the rules and which additional rules might also be invoked.

Ethics rules embedded or programmed into an AI, if they are to emulate a key aspect of human ethical rules, should support something like this deliberative consideration. A central purpose in this paper is to offer an account of rules which warrants our considering both human and AI rules, including ethics rules, in computational terms. We consider behaviors which are solely rule governed as well as those which are also rule following. It is rule following behaviors, we argue, that pose serious challenges to aligning ethics rules among intelligent agents.

2. Computation

As mentioned above, discourse about ethics and values entails forms of practical reasoning about possible courses of actions. In distinction, consider intuitions about rule governed systems that are not thought to also engage in practical reasoning. A rock's falling behavior is *rule governed*. Nothing about its history will affect how it falls when dropped off a cliff. On the other hand, some rules are normative in that they derive from conventions or standards. Examples would include representing numbers with arabic numerals as opposed to Roman ones and ethics and moral rules. To follow normative rules we must learn how to apply them in any particular situation. With respect to such rules, our behavior would be characterized as *rule following*. A characteristic of rule following behavior is that it is possible for a rule follower to make mistakes as they follow rules. For example, in calculating a tip when paying a restaurant bill, we may incorrectly compute the percentage. Moreover, what counts as a mistake is not a private matter; proper rule following is, as we argue below, open to community assessment. While examples such as those above are simple to generate, it is not such a simple task to characterize what it is that is being exemplified.

The first half of the 20th century saw articles by Gödel, Turing, Post, and Church [5–8] showing seemingly independent paths to a characterization of what it meant to have a procedure for effectively computing something. This characterization, the Extended Church-Turing thesis (ECT), asserts that anything which is computable in the physical world is computable by a Turing machine. ECT asserts that any process which could naturally be called an effective procedure can be realized by a Turing machine where an effective procedure (algorithm) is, informally, a recipe instructing the machine, from moment to moment, precisely how to behave. Informally, an abstract Turing machine contains a read-write head and an infinitely long tape divided into squares. It has the capacity to read the content (a 1, 0, or blank) of the square under its read-write head, write a 1 or 0 on the square under its read-write head, to erase the content of the square under its read-write head, and to move its read-write head one square to the left or to the right. The formal embedding of logic and mathematics, computability, into binary arithmetic was accomplished independently by Gödel and Turing [5, 6] and has had enormous implications some of which continue to reverberate today.

From the perspective of ECT, a computation is simply a sequence (possibly infinite) of machine configurations. What is intriguing about this is that seemingly independent formulations of "effective computability" proposed by Church, Turing, and Post [5–7] all lead to the same sort of Turing machine type of result. Whatever is computable by a Turing machine can also be computed by a program and is a partial recursive function. The ECT claim then is that *everything that can be computed in the physical world can be computed by a Turing machine*. [2] Moreover, and this is the 'extended' part of ECT, the time it takes to compute something on a given Turing machine will be polynomial in the time it takes on any other. In other words, though simulating a supercomputer on my phone may take polynomial

[2] We thank an anonymous reviewer for pointing out that since ECT assumes that machines have discrete states it may be that analog machines could would be able to evade limits of discrete machines. This is an important observation. However, this seems more a question of physics than of computation. For example, if it is accepted that quantum mechanics undercuts viewing space-time as being fundamentally continuous, then discrete state machines are what is required to simulate precisely a quantum-based reality. In any event, ECT asserts but does not prove that a Turing machine is computationally capable of calculating anything that can exist in the physical world. A main purpose of this paper is to provisionally accept ECT as true and investigate its consequences for aligning and embedding ethical rules into an AI.

more time, at least the amount of time will not explode exponentially. In this paper, we are interested in the question of whether there is some natural association between this notion of computation and the concept of rule. Specifically, what might be significant issues in embedding moral and ethical rules in an AI.

Here we consider whether moral and ethical rules, as well as rules more generally, can be naturally represented as computable procedures in the ECT sense. If, as we argue, the answer to this is affirmative, it then makes sense to view questions about AI ethics and alignment from the perspective of computation. We are assuming that ethical and moral reasoning by humans is fundamentally computational and does not involve some special human-only force or quality. Our ruling out such nomological danglers should not preclude those who disagree from being interested in the ethics issues and challenges posed by AI we discuss in Section 2.2. Indeed, those who hold to a non-materialist understanding of the mind might see the concerns we raise to be reinforcing of those their ontological committments identify.

It is worth reiterating that the ECT formulation is extremely general. To say, for example, that a person is a physical realization of a Turing machine is, by itself, to say very little. To be substantively interesting, it would be necessary to provide details about the architecture of the posited Turing machine. It is also important to note that applications of this approach makes a clear ontological claim that objects, including moral and ethics rules, can be modeled computationally.

Our main argument is comprised of three parts. The first, involving ECT and computation, has been outlined above. The second, that rules, in particular ethics and moral rules, can be represented computationally is developed in the next Section 2.1. Our particular formulation of this point is but one of many ways this might be accomplished. The reader who is already convinced, or willing to assume, that rules can be written in computational terms is urged to go directly to Section 2.2 where we consider how this relates to questions of algorithmic fairness and aligning human values with AI.

2.1. Rules and Computation

Consider a general *rule* as a procedure for doing something. At first glance, a rule might be simple (as in a procedure for placing a mark on a piece of paper) or it might be quite complex (as in a rule for recognizing a collection of marks as a poem). The intelligibility of a rule will often be less an attribute of the rule itself than it is of the context in which the rule is encountered. For example, from the perspective of an English speaker attempting to recognize marks as an "A", the rule (or rules) may appear trivial (just look at the marks and decide!). Yet for AI researchers trying to understand how it is that humans recognize characters, the rules may appear both complex and subtle. An implication of this latter example is that many rules may not be cognitively accessible to the person whose behavior may be partially determined or governed by those rules.

Our argument is straightforward.[3] We first offer a syntactic account of *rule* which is intended to permit the written expression of the largest possible variety of claims about rules. Next we define semantics for a *rule* showing how to computationally evaluate a syntactically correct rule. In discussing both syntax and semantics will first discuss informally what we mean and then offer a more formal definition. Formal definitions may be skipped over without losing the central thread of our argument. The conclusion of this argument is our claim that a general account of rules can be expressed in computational terms.

Rules, in particular ethics rules, while propositional do not generally take on truth values. In this sense, one way of envisioning a rule, R is as a relation between two environments. If H partially describes an environment, and C denotes a partial state of affairs in some possible "future" environment, then R can be interpreted as positing a relation between two environments (or worlds). For example, an ethics rule "When shopping in a store do not remove items without first paying for them" expresses in ordinary language a rule that relates an environment (shopping in a store) with a possible state of affairs (pay for items before removing them from the store). Informally this suggests that the application of ethical rules in particular circumstances requires figuring out what environment we are in and what sort of future environment we would like to be in and then determining which rules properly apply. Each of these steps, we suggest, is fundamentally computational in that it involves the, possibly recursive, application of Turing-like computations until all relevant variables take on concrete values (in computing language, we say these variable are bound). So, for example, applying a rule involving *shopping in a store* requires first recognizing that our current situation does involve shopping in a store. This recognition may itself involve many applications of rules to get from our current visual field to cognitive recognition that we are now in a store. Many, maybe most, of these computations will occur without any cognitive awareness on our part. We simply know we are now in a store.

[3] The discussion of rules in Section 2.1 derives from that in [9] which also contains a taxonomy of rule types and examples based on our definition.

Formally, to account for these many computations we must identify their general form; their syntax. H and C can be constructed from a number of $h_1, h_2, \ldots, h_m$ and $c_1, c_2, \ldots, c_n$, respectively. The elements of H and C may be connected by logical operators or expressed by clauses internal to the body of the rule. Thus a general form of a rule is:

$R < h_1, h_2, \ldots, h_m; c_1, c_2, \ldots, c_n¿$ where the ";" is used to separate elements of H from those of C. An illustration in this notation is: $R < h_1 \wedge (h_2 \vee h_3); c_1 \wedge c_2 >$.

The elements of$H$ identify a partial state of affairs in some environment. A rule is evaluated in a particular environment. It is the function of H to indicate the proper environment by referring to some part or description of it either explicitly or by virtue of where it is activated. For illustration purposes, we will make reference to a Lisp-like denotational semantics though later we will suggest that a reduction semantics, similar to what we see in Mathematica, should also work.

What then are the h_i and c_j? From a computational perspective, these are variables where each argument of R could take on more than one value. Moreover, variables can themselves be rules. Nothing in the syntax of a rule precludes rules utilizing other rules. A statement of the form $R(R(H; C); R(H : C))$ is syntactically valid.

The syntactic form of a rule is not sufficient for our interests. We want also to know something about what the rule refers to; its semantics. Indeed, rules are interesting precisely because they are often applicable to more than a single situation. There can be distinct values for variables in one or more environments and our semantics must be robust enough to incorporate the ways in which rules might be used in, among other things, computational intelligence and ethical reasoning. The basis for our semantics depends upon associating or binding the variables h_i to particular values in a given environment in order to complete the conditions necessary to executing or evaluating C by binding values to the c_i as well. If we have a form which contains a variable, a binding for that variable permits us to rename the variable consistently throughout the form.

Our formal discussion of evaluation derives from the environment model of evaluation described in [10]. The Lisp dialect Scheme is based upon such an evaluation model. In Scheme, an environment is a sequence of frames where each frame is a set of bindings. Moreover, each frame (except the global frame) has a pointer to an enclosing frame. The value of a variable relative to an environment is determined by looking sequentially for the first frame in the environment in which the variable is bound. A variable is said to be unbound in an environment when there is no binding for the variable in any of the environment's frames. Specification of these details permits the instantiation of rules as we define them and thereby connect them with effective computability.

Values to which variables are bound are determined by presence of an environment together with an evaluation model as computation may be required to identify an h_i with some value or partial state of affairs in a given environment. Binding is rule-governed and generally the result of pattern matching algorithms. A critic might then claim that this begs the question as we are still left with rules. Our rejoinder is, while agreeing that rules require rules, that our point is precisely to demonstrate that a common sense understanding of rules renders them equivalent to effective procedures and thus subject to computational analysis. It's rules all the way down.

Specifically, the generality which characterize rules derives from the computational processes required to bind previously free variables. For a rule having at least one unbound variable, its generality stems from the number of ways in which those variables can be bound. Rule generality is associated with (i) the definition of the variable, (ii) the relevant environment, and (iii) the process used to bind. This generality is important since in ordinary language we recognize the existence of rules which are not applicable to the present situation. A rule against stealing does not go out of existence when there are no objects which might be stolen in the vicinity. While this rule may be inappropriate to one's current environment, it does not lose its relevance to others.

An important part of our thesis is that rules are linguistic objects. We cannot speak of a rule absent a formulation of it and the computations involved in evaluating a rule will be specified by its internal structure. Informally, we also want our conception of rule to comport well with our ordinary understanding of rules. We then rely on ECT to warrant that we can, in principle, computationally rewrite rules initially formulated in one language in another in much the same way one computer operating system can be emulated in another. In this manner we might hope to take ethics rules expressed in, say English, and rewrite them computationally in, say, Lisp. However, and this is a central point we will develop below, using ECT to say that we can in principle do this is quite different from knowing whether or not we have, in fact, accomplished it. This, we will argue, is at the center of our understanding of the ethics AI alignment problem.

First, however, we must formally define a semantics for representing rules computationally. We will illustrate this continuing to use elements of the Lisp

programming language. In interpreted Lisp, a top level form is read and evaluated according to Lisp evaluation procedures, and the result of the evaluation is then printed. For example, if we feed Lisp the form `(+ 5 (+ 3 2))`, our Lisp will respond by printing the form `10`.

Here both the original form and the returned form denote the same thing—the number 10. We can say that while the original form was reduced to the simpler form, both `(+ 5 (+ 3 2))` and `10` denote the same object. On the other hand, consider the form `(quote (+ 3 2))`.

The quote special form in Lisp is evaluated meaning to return the form being quoted and `(+ 3 2)` will be returned. Here the original form and the one resulting from its evaluation do not denote the same thing. The original form denotes that to which it evaluates. This illustrates that Lisp evaluation embodies both a reduction semantics (the first example) and a denotational semantics (the second example). Reduction is a replacement process where subforms of a form are rewritten according to specified evaluation rules. In a pure reduction semantics the process is defined for any form so if no rewrite rule exists for subforms of a form, then the form simply reduces to itself. In order for Lisp procedure calls to be evaluated any variables appearing in the call must be bound prior to the call being made. In the rest of this discussion, we will employ a Lisp-like syntax but assume a reduction semantics.

Consider an informal discussion of the value of life. The precise meaning we associate with life will depend upon the context of the discussion; it may mean different things in discussions of abortion policy, criminal justice, or evolutionary theory. Similarly, the value of symbols in Lisp will depend upon the environment in which they are evaluated.

Our semantic notion of a rule then is:

Definition. *For at least one relevant environment E (which must be finite), a rule is a procedure of the form $R < h_1, h_2, \ldots, h_m; c_1, c_2, \ldots, c_n >$ such that*

(i) one or more h_i is initially unbound, and

(ii) one or more c_j is initially unbound, and

(iii) the internal structure of R is such that an attempt to evaluate R in E creates an attempt to locate a binding for at least one h_i.

A rule is defined both by its syntax and by what occurs with evaluation attempts. Note that identifying a rule with a particular procedure is not the same as defining it as being effectively computable. Our definition has a number of implications. First, evaluation of a rule must invoke at least one binding rule to attempt to locate a value (binding) for any unbound variable. The binding rule recursively invokes itself until it succeeds or fails. This ensures that evaluating a rule reduces, or at least does not increase, its generality. Further, the definition permits rules whose evaluation is interrupted as a result of binding to particular values. Attempting to bind at least one h_i is essential for otherwise there will be no action consequences. If all h_i were already bound, then we would have a non-contingent formula. Such a formula could be termed a command if the h_i, if any, are already bound or a commandment if the h_i may be bound differently in different environments. A commandment which applies to all environments would include moral laws intended to apply regardless of context such as "thou shall not kill humans." Evaluation of a rule which results in all of its variables being bound is a formula. Significantly, rules, including ethics rules, need not produce such definitive results. In this sense, our conception of rule is broader than that of a recipe. Most importantly, rules, as we have defined them, are forms for expressing moral or knowledge claims using variables.

2.2. Aligning Ethics Rules

Thus far we have presented the Extended Church-Turing thesis (ECT) asserting that anything that can be computed in the physical world can also be computed by some Turing machine. Importantly, Turing introduced the formal notion of a general machine, now termed a universal Turing machine, which can take the code of any given machine and reproduce the behavior of that machine. A machine is universal in that it can precisely simulate the behavior of any other computing machine given access to that machine's underlying code. This is the idea that gave rise to programmable computers. Such machines can at one moment balance our checkbooks or maintain our calendars and, at another, guide a spaceship to the moon. Some computers may be faster than others, but all are capable of doing the same thing—compute all effective procedures or, equivalently, all partial recursive functions. In the previous section, we offered a precise characterization of the syntax and semantics of a *rule* and suggested that our formalized definition of rules, including ethical rules, are effectively computable.

Taken together, ECT and rules as defined here, support the claim that, in principle, it is possible to embed ethical rules into an AI. Much more importantly, though, we must be concerned with what specific combination of ethical rules are coded into a given AI. How, for example, is an autonomous vehicle to

weigh the implications of various possible courses of action when a collision of some sort is calculated to be inevitable? Or, drastically upping the stakes, what action should an autonomous national security AI initiate when it senses a "significant" probability that missiles are headed its way? And, should such an AI have an off switch permitting human override of its selected course of action? In considering such possibilities we presumably would like the AI rules to be in close harmony with human rules.

Consider the following cases. First, in the early morning of September 1, 1983 a Soviet Su-15 interceptor shot down a Korean Airlines passenger jet heading to Seoul by way of New York City and Anchorage. The flight, KAL 007, had veered considerably off course and was unknowingly flying over restricted Soviet airspace. Precisely why it was so far off course is disputed though a critical navigation beacon near Anchorage was offline for maintenance and either the pilot neglected to notice this and take manual corrective action or for some other reason the plane's guidance system was in the wrong mode and for five hours the flight continued to fly far north of its planned route. In fact, it was so far north that it exceeded the tolerance programmed into the plane's computer whereby it would have reset the navigation system mode to request a new waypoint.

Later that same month on September 26, Soviet Air Defence Forces Lieutenant Colonel Stanislav Petrov was duty officer for the Oko nuclear early-warning system when the system's computers sensed that missiles launched from the United States were directed toward the Soviet Union. Military protocol required Petrov to immediately send reports of any apparent missile launch up to his military and political superiors. The result could well be initiation of a retaliatory strike against the U.S. "The siren howled, but I just sat there for a few seconds, staring at the big, back-lit, red screen with the word 'launch' on it. A minute later the siren went off again. The second missile was launched. Then the third, and the fourth, and the fifth. Computers changed their alerts from *launch* to *missile strike*", Petrov is later reported as telling the BBC [11]. Though the computer generated alert was unambiguous, something bothered Petrov. "There were 28 or 29 security levels. After the target was identified, it had to pass all of those 'checkpoints'. I was not quite sure it was possible, under those circumstances" [11]. Instead of sending up an alert as his orders required, Petrov decided this was most likely a system malfunction and so reported. A possible nuclear exchange was averted. Petrov credited his decision to his civillian education. His colleagues, he claimed, had all been trained in military schools to simply follow orders.

Recall that in our terminology, an order is a rule with no unbound variables. What Petrov did was to question the applicability of that order and do something else. He reports not being sure he did the right thing until about 23 minutes passed and there had been no explosive strikes. While, given ECT, there must exist sets of rules which could emulate Petrov's decision to suspend evaluation of the order rule and do something else, how might an AI coder identify such rules? The situation Petrov faced was extremely rare. He had been told that the computer code reporting a launch had multiple programmed safeguards that had to be passed before categorizing the sensed data as a 'strike.' Still, something triggered a feeling that this was more likely a system malfunction as it did not seem reasonable to him that all those safeguards could have been passed so quickly and also he was not getting any launch reports from Soviet satellite radar operators.

This examples can be seen as involving computers and humans following (or not following) explicit rules. One ended tragically and the other did not. While neither example involves anything approaching an AI, they do suggest the importance of examining what we might expect as more decisions become largely controlled or informed by AI-based systems. Together, the examples illustrate an issue stemming from our analysis. Recall that some behaviors may be *rule following*. However, given our materialist ontology, must not all behavior ultimately be rule governed in the sense that it, at least in principle, can be accounted for by physical laws?

Imagine we are teaching a child how to continue a series of numbers by adding 2 to the previous number. We illustrate what we mean by showing the sequence $0, 2, 4, 6, 8, 10, \ldots$. The child asks, what does $\ldots$ mean and we respond that it means to go on in the same way. The child lights up and says, "Oh, I see," and extends the sequence to $12, 14, 16, 18, 20$. Then, asked to start the series at 1000, the child responds with $1000, 1004, 1008$? Our response might be "No, no; just do what you did before." However, suppose the child answers that she *is* doing what she did before? As a matter of fact, she says her rule was that up until 1000 the rule was to add 2 and after that to add 4 to the previous number.

Wittgenstein [12, 13] argues that nothing in the child's previous answers logically excludes her having been following a rule that adds 4 after 1000. Indeed, an infinite number of distinct rules can be imagined which are consistent with both $12, 14, 16, 18, 20$ and $1000, 1004, 1008$. It does not suffice to say that the rule you intended to convey was an *algorithm* which takes

a natural number as an input and returns a sequence beginning with that number since an algorithm itself is a rule, perhaps comprising a set of rules, and thus recursively subject to the same criticisms.

A response might be to try to end this quibbling by implementing the rule you intend as a computer program and saying that what you mean by the rule is the computer code and what you mean by following the rule is producing the same sequence as does the computer when fed a natural number. However, the same problems arise in a slightly different guise. Your code implementation requires following a set of coding rules, and we are once again left with having to clarify what it means to follow those rules. Further, any actual computer is, of course, a physical device subject to design errors, manufacturing defects, and malfunctions and the code in which you implement the algorithm itself may contain errors.[4]

The central issue here is, as pointed out by Wittgenstein, that following a rule cannot be something purely privately understood. If this were not the case, then there would be no distinction between my believing I am following a rule and my following it. I could not be mistaken about my rule following. Rather, following any particular rule involves shared linguistic understandings of the communities within which the rule is embedded. For example, within a community of basketball players, "shooting a free throw" involves a set of rules or conventions specifying where to stand, when to stand there, and so on. Learning these rules is a part of what it takes to be fully part of a community of basketball players. Importantly, this learning could not be reduced to looking at a printed list of rules and doing what they said. The *doing what they said* part is precisely that which must be learned and may well involve being coached and corrected as to what it means to follow them in specific instances.

Within many linguistic communities, the sequence $0, 2, 4, 6, 8, 10, \ldots$ has a *standard* interpretation and being a participant in those communities involves learning that interpretation. Understanding what is meant by *going on in the same way (...)* does *not* necessarily involve reducing ... to something simpler. Instead, we come to understand what is meant by being shown many examples. We then demonstrate our understanding by completing additional examples in the presence of community members to see if they agree that we are doing it right. That is, have we demonstrated that we appear to be following the standard interpretation? A critical point here is that the test of our understanding is a public linguistic one and not our privately held belief.

Following rules leads to those rules becoming embedded within our brain and, in the process, possibly rewriting rules already there all in a manner fully compatible with ECT and a materialist ontology. This does not, however, mean that we will be aware of all those rules. A ball is thrown to us and we raise our arms and catch it. Could we articulate the rules learned from infanthood that led to our accomplishing this rather amazing feat?

Learning to follow rules requires engagement with a language community. It is a process involving both rules we are aware of (for example, do not steal) as well as evaluation rules which may lurk beneath our immediate cognition (for example, extreme cases where maybe it's okay to steal).

This notion of following rules by engaging with an appropriate shared language community is extremely relevant to the development of machine intelligence. We illustrate this by reference to natural language AI models. A popular approach for developing these models involves taking domains of interest, collecting lots of presumably relevant data and programming the computer to look for patterns in those data in a largely unsupervised manner. Specifically, predictive natural language models of the sort we discuss take sentences or other strings of text and, in varying ways try to leverage the context of the text to predict what text will come next.

For example, OpenAI's predictive text natural language processing (NLP) model was provided 5700 gigabytes of unlabeled text[5] as data from which to produce a neural network. The resultant model, GPT-3, has 175 billion parameters [14] and is capable, with very little training, of accomplishing tasks including summarizing documents, generating recipes, and acting as a chatbot. This is clearly rule governed behavior though it did not result from learning via rule following in the sense we have described it. Training only minimally required examples (few-shot learning) and did not involve ongoing engagement as a participant in a language community.

Google Brain's Switch-C NLP has up to a trillion parameters [15] and was trained using the *Colossal Clean Crawled Corpus* [16]. The text in this dataset came from a 2019 snapshot of the Web filtered to exclude such things as non-English text, non-sentences, offensive words, and so on. Included then is text from Wikipedia, major news sites, open-access publications, etc. The vast majority of included text, 92%, is estimated to have been written between 2011 and 2019 [16]. Switch-C was trained on tasks such as masking

[4] See [12, 13] for a more complete discussion of why coding rules on a computer does not solve the underlying issue.

[5] As mid-2021 the entire Wikipedia consists of a little under 20 gigabytes of (compressed) text.

15% of the words in sentences and then learning to predict the missing text and locating text to answer questions. Importantly, Switch-C has the capacity to determine which parameters to use for a given presented example by routing incoming sample texts to a specific 'expert' layer in the network as computed on the fly thus avoiding the computational and energy overhead of invoking multiple, often unhelpful, expert layers. The basic idea is that an AI model should not always have to reference all of its components each time it is presented a task. Of course, implementing such a notion is no small job. As with GPT-3, while Switch-C's behavior is rule governed, its training did not involve ongoing active engagement with a language community.[6]

Finally, we can look at what consequences this all has for aligning human values with those in an imaginable AI. Already we have autonomous vehicles on the road and in the air. How can we be confident that such systems comport well with human values? That is that they follow human ethics and moral rules. We do not want AIs eating our metaphorical daisies.

In his recent book, Brian Christian [18] takes up the issue of risks associated with AI. He writes, "How to prevent such a catastrophic divergence—how to ensure that these models capture our norms and values, understand what we mean or intend, and, above all, do what we want—has emerged as one of the most central and most urgent scientific questions in the field of computer science. It has a name: the alignment problem." From an ethics perspective, we would like to see AI and human entities operating from similar or identical principles or, at least, knowing where we disagree. Importantly, we would not want there to be significant divergence between our human values and those of an AI. Additionally, as human ethics and values develop and change, we might want an AI to somehow reflect those changes.

Alignment issues with AI large language models as well as AI facial recognition programs have been well documented. The now classic paper by Buolamwini and Gebru [19] provide a detailed account of how sampling bias in images used to build facial recognition programs bleeds directly into ethically repugnant decisions being made when such AI products are commercialized and used in law enforcement [20]. Similar issues arise with the large language models. Given how they are constructed, the rules they learn reflect the text they use to form those rules. "Biases can be encoded in ways that form a continuum from subtle patterns like referring to women doctors as if doctor itself entails not-woman or referring to both genders excluding the possibility of non-binary gender identities', through directly contested framings (e.g. undocumented immigrants vs. illegal immigrants or illegals), to language that is widely recognized to be derogatory (e.g. racial slurs) yet still used by some" [17]. When insidious ethical and moral rules and concepts are embedded in the corpus used in developing a model, it should be no surprise that those same rules are reflected in the AI model itself. Or when images used to train a facial recognition system do not adequately capture the full range and distribution of actual faces in the worlds in which that system is later deployed, it should be expected that the AI model may produce ethically relevant misclassifications.

Our contention is that these sorts of ethical issues cannot be remedied by simply doing better sampling. In many cases, it is the reference population itself that is of ethical consequence. Is the set of people convicted of a crime in the U.S. an appropriate population to sample from if we are interested in characteristics of criminals? Or should we first acknowledge that the U.S. criminal justice system itself has embedded normatively relevant biases? How do we sample from alternative justice systems and over what time period? As Nick Bostrom argues, eliminating reference population bias is fraught with difficulty and potential paradoxes [21].

There are several related senses of *bias* which merit being distinguished here. The first is technical bias resulting from poor sampling technique. This can occur, for example, when a training set does not properly represent the population from which it is drawn. A second sense of bias we term fairness bias. This form of bias can occur when either a training itself set contains attributes that many might consider irrelevant and inequitable. For example, suppose an AI trained to assist in hiring decisions used, among other factors, schools attended by previously successful applicants. Such an AI would exhibit fairness bias if it turned out that applicants attending the same schools as the CEO had previously been favored. Similarly, if society as whole has become organized in an unfair manner then it should be expected that representative training data from that population will bring along with it fairness bias as seen in the language models discussed earlier. What we are terming fairness bias is, we think, related to Benjamin's *New Jim Code* [22] in that fairness bias can be deeply structural and, as such, is largely immune to mediation via technical means such as better sampling. This is extremely troubling for any AI that is built solely using training data based on a snapshot of society. Exposing fairness biases would seem to require the sort of engaged deliberative reasoning associated with rule

[6]Learning to imitate text from a fixed body of text, however large, is not, we contend, the same as generating sentences in contexts and having others, and these others most certainly could be non-human machines, query and correct your usage. Our distinction here is captured well in [17]. More on this below.

following.

From the perspective we have been developing in this paper, there are additional issues which, we believe, pose serious challenges to embedding ethics in an AI. All AIs, as is the case with humans, are rule governed in that their behaviors are compatible with physical principles and ECT. Ethics rules, however, exemplify a particular kind of rule in that learning and following them requires ongoing engagement with a language community in which those rules are maintained and modified.

In other words, rules such as ethics rules are not static; their evaluation and even structure can change as a consequence of discussion among community members. Though rules may be well understood within a language community at any given point in time, these understandings may change over time and thus appear to an outsider as vague. This poses challenges to attempts to encode the rules within an AI. If the AI is not in ongoing interactive engagement with that language community, its encoded ethics rules may fall out of alignment. The kind of AI model exemplified by GPT-3 and Switch-C would appear incapable of interacting with a reference language community for several reasons. First, as constructed, their representation of language is largely based on the past. They try to respond to a given problem by interpreting it, even if it is presented in natural language, in terms of rules based completely on past usage. Further, a query of these systems as to why they were using language the way they were would involve looking at a huge number of parameter values, loss functions and interconnections. We have an AI parroting human language [17] but unable to participate in deliberative reasoning with humans in that human language. It is akin to trying to create a shared map of a room with a bat [23]. Yet, as we have argued, deliberative reasoning is at the core of moral and ethics discourse.

Nothing we have said precludes developing an AI capable of continuing engagement with a human community using a largely shared human language. Humans may have some rules inaccessible to them and thus not open to deliberation and the AI may have rules unintelligible to we humans. All that is required is that deliberative reasoning be carried among humans and the AI. It is not inconceivable that such a rule following AI could be developed. However, now we run up against a different set of problems. First, the AI could almost certainly reason many orders of magnitude faster than most, if not all, humans. Second, the AI itself could be cloned at very low marginal cost. The result would be that the AIs might quickly dominate community discussions both in terms of number of participants and the speed with which those participants reached ethical conclusions. A consequence is that our shared language would veer in the direction of what was computationally efficient for the AIs and, indeed, ethics rules might also shift to reflect the much faster reasoning of the AIs. Once again ethics rule would move away from alignment with human values.

3. Conclusion

With AI being embedded in many aspects of our lives, warnings are being raised about implications for furthering inequality or bias already existing in society [19, 22, 24]. A widely shared concern, for example, is that use of AI models will lead to legal, financial, or health initiatives disproportionately harmful to marginalized populations if these populations are not properly represented in training datasets [19, 25, 26]. In this paper, we argue that addressing this and other related issues requires rethinking beyond the legitimate critiques regarding representation and sampling. Specifically, we propose a framework centered on rule following to examine fundamental issues in ethics, AI alignment, and computational models.

Our argument here is that though all AIs are rule governed, only some can also be described as rule following. A direct implication is that a snapshot of an AI will not be sufficient to determine whether it is, or has been, rule following as the distinction lies in how it has developed over time in connection with its external environment. For many AIs, this may be an anthropomorphic distinction without a difference.

> Requiring that AI systems, like human beings, use language, possess concepts, and reason—that is to say, follow rules of meaning and grammar, categorization, inference, and so on—looks like one more example of anthropocentrism in AI. For if the rhetoric of AI is disregarded, an AI system's (in practice) inability to follow rules, use language, possess concepts, or reason is an "irrelevant disability" (Turing's phrase)—as unimportant as the lack of a face or mother [27].

However, as AIs move into areas involving decisions with ethical implications we contend that our distinction does matter as learning and acting with regard to ethics and morality is simultaneously rule governed and rule following. Much of what matters in politics, for example, involves debate over what is the 'right'

thing to do. These debates are engaged within language communities using phrases which themselves can restrict or open up questions of ethics and morality. Terms such as *Asian American* or *woman doctor* are not simply descriptors but also focus attention on category distinctions capable of carrying considerable moral weight.

Norbert Weiner, an early theoretician in what has become AI, clearly anticipated the concern expressed in this paper when he wrote "the machine like the djinnee, which can learn and can make decisions on the basis of its learning, will in no way be obliged to make such decisions as we should have made, or will be acceptable to us. For the man who is not aware of this, to throw the problem of his responsibility on the machine, whether it can learn or not, is to cast his responsibility to the winds, and to find it coming back seated on the whirlwind" [28].

References

[1] J. Kerr, *Please Don't Eat the Daisies*. Doubleday & Company, Inc., 1953.

[2] S. Kierkegaard, *Fear and trembling and the sickness unto death*. Princeton University Press, 2013.

[3] C. O. Schrag, "Note on Kierkegaard's teleological suspension of the ethical," *Ethics*, vol. 70, no. 1, pp. 66–68, 1959.

[4] J. A. Clark, "The meaning of ethical propositions," *The Philosophical Review*, vol. 56, no. 6, pp. 631–644, 1947.

[5] K. Gödel, "On formally undecidable propositions of Principia Mathematica and related systems I," *Monatshefte für Mathematik und Physik*, 1931.

[6] A. M. Turing, "On computable numbers, with an application to the entscheidungsproblem," *Proceedings of the London mathematical society*, vol. 2, no. 1, pp. 230–265, 1937.

[7] E. L. Post, "Formal reductions of the general combinatorial decision problem," *American journal of mathematics*, vol. 65, no. 2, pp. 197–215, 1943.

[8] A. Church, *The Calculi of Lambda Conversion.(AM-6), Volume 6*. Princeton University Press, 2016.

[9] J. P. Bennett and S. Thorson, "Are all rules effectively computable," Department of Political Science Working Paper, Syracuse University, 1985.

[10] H. Abelson and G. J. Sussman, *Structure and interpretation of computer programs*. The MIT Press, 1996.

[11] P. Aksenov. (2013, September) Stanislav Petrov: The man who may have saved the world. [Online]. Available: https://www.bbc.com/news/world-europe-24280831

[12] L. Wittgenstein, *Philosophical investigations*. John Wiley & Sons, 2010.

[13] S. A. Kripke, *Wittgenstein on rules and private language: An elementary exposition*. Harvard University Press, 1982.

[14] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *arXiv preprint arXiv:2005.14165*, 2020.

[15] W. Fedus, B. Zoph, and N. Shazeer, "Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity," *arXiv preprint arXiv:2101.03961*, 2021.

[16] J. Dodge, M. Sap, A. Marasovic, W. Agnew, G. Ilharco, D. Groeneveld, and M. Gardner, "Documenting the english colossal clean crawled corpus," *arXiv preprint arXiv:2104.08758*, 2021.

[17] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the dangers of stochastic parrots: Can language models be too big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 2021, pp. 610–623.

[18] B. Christian, *The Alignment Problem: Machine Learning and Human Values*. WW Norton & Company, 2020.

[19] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[20] C. Garvie, *The perpetual line-up: Unregulated police face recognition in America*. Georgetown Law, Center on Privacy & Technology, 2016.

[21] N. Bostrom, *Anthropic bias: Observation selection effects in science and philosophy*. Routledge, 2013.

[22] R. Benjamin, "Race after technology: Abolitionist tools for the new jim code," *Social Forces*, 2019.

[23] T. Nagel, "What is it like to be a bat?" *The philosophical review*, pp. 435–450, 1974.

[24] V. Eubanks, *Automating inequality: How high-tech tools profile, police, and punish the poor*. St. Martin's Press, 2018.

[25] H. Seo, H. Britton, M. Ramaswamy, D. Altschwager, M. Blomberg, S. Aromona, B. Schuster, E. Booton, M. Ault, and J. Wickliffe, "Returning to the digital world: Digital technology use and privacy management of women transitioning from incarceration," *new media & society*, p. 1461444820966993, 2020.

[26] M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath, "A review of challenges and opportunities in machine learning for health," *AMIA Summits on Translational Science Proceedings*, vol. 2020, p. 191, 2020.

[27] D. Proudfoot, "Robots and rule-following," in *Alan Turing: life and legacy of a great thinker*. Springer, 2004, pp. 359–379.

[28] N. Wiener, *The human use of human beings: Cybernetics and society*. Da Capo Press, 1988, no. 320.