Using Reversed Toulmin Arguments

Daniel Buehrer and Wei-Khiong Chong

August 17, 2018

Submission to the www.herox.com/EthicsNet Guardians' Challenge

Abstract

This white paper considers the question of how to create "friendly AI" that makes decisions that are aligned to human values. This question is not directly related to the above challenge question of how to create big data for teaching ethics to moral machines. We believe that the period for training machines to be moral will be extremely short, on the order of months, since machines are capable of learning very quickly, not forgetting, and sharing their conceptual frameworks. The socially intelligent machines (sims) of the future will fairly quickly be able to read/use all human literature, conceptualize and model human motivations and actions, and predict the outcomes of human/machine actions. Moral machines are sims whose actions are aligned with the "best" human values. In this white paper we propose a future in which each person is guaranteed a cell phone/pad/notebook connection to a "personalized advisor" that is able to present each individual or institutional officer with personalized advice on the likely effects of his actions. That is, the sims will learn models directly from trying to minimize the differences of their generative models of the world and the actual outcomes of the actions of the human or robots that they are assigned to.

1. Introduction

How long can humans be the Guardians of moral machine behavior? The period for training a moral machine will be very short, on the order of months, because of the machine's ability to remember and generalize from the experiences of its robots. As when educating a child, humans will need to provide very clear boundary lines for the behavior of moral machines and their robots, with appropriate rewards/punishments, and with clear criteria for evaluating the effects of its robots' behaviors on its major goal (e.g. creating an "ever-advancing" society that includes human values [1]).

Such goal-directed behavior is not common for humans, who make decisions that are mainly based on gut feelings, emotions (e.g. fear, greed, desire for fame), habits, and sometimes even altruism. In this paper we propose to use the concepts of Toulmin arguments to model decision making of animals and humans. We consider mainly binary choices, such as "fight or flight", since reasoning about sequences of n actions can quickly be overcome by the number of permutations ($n!$). However, we do allow the addition of "MacroDefn" classes that describe all paths of operators from a specified condition to a specified goal state. For instance, a "Mover" can be an Organization that is able to decide how to move a piano into my third-floor apartment, based on all of its experience of moving furniture, including pianos. It basically has a simple plan involving the "macros" of measuring the doors, elevator, and stairs, possibly disassembling the piano or getting a crane to lift it through a large window, or knocking down walls and reconstructing them. The details of the plans are left until later. That is, most plans are made one decision at a time, making the decision what to do next based on the outcome of the previous decision.

Billions of dollars are being spent on artificial intelligence by defense departments, advertising companies, and manufacturers. Who will be interested in developing moral machines? From the current sales of books and online advice from "expert advisors" and religious advisors, there seems to be a market for morally intelligent machines. Let's hope that some companies will pick up this challenge before the corrupting influence of unbridled capitalism gets us all addicted/brainwashed.

Moral machines will probably "grow up" to become personalized advisors for individuals and leaders of smart social organizations, companies, and governments. Their models of social interactions will include deep understanding of individuals and their interactions. The deep understanding of people's individual attitudes, capacities, abilities, and time/health constraints will allow the machines to allocate humans resources in smart governments and organizations, based on the humans' desires. If each person's conceptual framework can be mapped to locations his brain, such a brain-machine interface will allow us to communicate with machines and with each other at much higher bandwidths. Although the mapping of the 37 kinds of synapses has only been done on rats [2], similar experiments on humans may allow each of us to communicate directly with his moral machine(s).

Like IBM's Project Debater (with which we have no affiliation), socially

intelligent machines (sims) of the future will be able to marshal arguments in order to convince people of what the sims "conclude" is true. Like mothers, they will try to find the most effective methods to remind their advisee that it is bad to gamble, drink alcohol, play around with girls/boys, or race cars. Each sim advisor will use its experience to develop a consistent worldview that its advisee can eventually agree with. But humans should have the right to disregard or turn off their personal advisor, just as they have the right to run away from parents.

We describe sim advisors that use the "white box" reasoning of Toulmin arguments. AI deep learning neural network based models are usually considered to be "black box" models in that they cannot explain the reasoning behind their conclusions, but simply use layered induction of patterns of convolutions and aggregations that are often not interpretable by humans. As pointed out by Judea Pearl [3], most inductive AI only does curve-fitting to find associations, without even any notation for cause-effect relationships and counterfactual reasoning. Explainable reasoning has to involve operators that change state of both the world and the model of the world. For most previous AI learning, the model was usually not included with the data input.

One of the major advances in 2018 (e.g. [4, 5]) was the use of differences between the model and real-world data to update the model. In this paper we attempt to show how a model of the world can be updated, and how its learning algorithm can even prioritize experiments that are used to update its model and to find counter-examples. This ability of the program to choose how to modify its own model may be the beginnings of a conscious super intelligence. In order to prevent extinction of carbon-based life forms, we should be careful that the goals of this new life form are forever aligned with common virtuous human goals (e.g. [6]).

In the rest of this paper we describe Toulmin arguments [7], how they can be implemented via class algebra intuitionistic logical inference [8], and how KL-divergence [9] can be used to update the social models that are based on Toulmin concepts. In [5], the KL-divergence is called "precision-weighted prediction error" of generative models. Most previous generative models were simply based upon deep learning of neural networks. Here we consider a mathematical way of predicting the minimal descriptions of sequences of operators. The plans with similar preliminary states and result states are combined into "MacroDefn" classes. These macros can answer "Why" questions with either "because" answers or "so that" answers. The

inverse MacroDefns can answer “How to” questions by proposing the preconditions of the appropriate macro (i.e. one entailing the given goal condition) as a goal state.

2. Toulmin arguments

Most current AI is built on the idea of induction. So, for instance, if our samples are representative, the farmer will bring the chicken food every day. We do not have to consider why the farmer is doing that, or what happens to the chickens after they are put on the truck. Science is based on the scientific method, which is more than induction. Its arguments can be based on evidence entailing generalized conclusions, conclusions being supported or negated by evidence or counter-examples, or the making of plans which are based on models and the results of previous experience.

However, the scientific method generally does not include judgment or value claims, which involve opinions, attitudes, and subjective evaluations. The scientific method also usually does not involve policies of organizations, the differentiation of facts and propaganda, or the trustworthiness of facts or rules, although more recent versions of Bayesian networks and fuzzy logic have started to consider such trustworthiness factors. Basically, as long as each of these features of states and their operators are included in a model, the model can make use of operators that implement a more general version of reasoning that is more similar to that used by humans.

There is a simple algorithm that can essentially make use of Toulmin arguments to update its social model of the world by comparing the results of its model’s decisions to either game-based generative adversarial network (GAN) models or the real world of its robots. This “Debater” algorithm can then arrange its models and experiences using Toulmin arguments [8] to convince its advisee or other humans or socially intelligent machines (sims) that its worldview model is correct, at least for certain cases.

Basically, to do Toulmin arguments, you simply need to define a class for each of the terminologies of this theory and “implement” the operations of reasoning, such as using a transitivity axiom that adds “important” paths through the operator edges as a “macro” edge, which goes from a startState to reachableState. In class algebra, a startState or reachableState of an operator/macro edge has a normalized conjunct of comparison operators (i.e. an intent of a Situation class) that describe an operator’s precondition or result state. The acyclic graph of operator and macro edges connect the experienced or heard-about situations to make plans or find causes. The

Operator/Macro class definitions contain a startState as a minimal description of preconditions, and a reachableState as a minimal description of result states (e.g. goals or bad outcomes) of the operator or macro. In this paper, we especially emphasize the use of reversed edges, letting the people know the "how to" operators that can lead to desired states. For example, the sim's model may say that changing one's contacts, listening to certain Ted talks or reading certain e-books, or doing some kind of service may result in a happier life.

Debaters often use Toulmin arguments [8]. The arguments involve 6 main concepts, which are easily defined as class algebra intents/extents:

1. *Evidence* (minimal conjunction of preconditions of an Operator or Macro)
2. *Claim* (the conjunction of conditions resulting from an Operator or Macro)
   1. Fact-based 2. Judgment/value (opinions, attitudes) 3. Policy-based
3. *Warrant* (an Operator, along with its preconditions and postconditions; an edge in a state-space graph)
   1. Generalization (Add a new "Claim" class based on the evidence)
   2. Analogy (Add a new "Analogy" class is that describes similar (i.e. homomorphic) relevant situations, events, precedents)
   3. Signs (e.g. symptoms entail a new Disease class as their cause)
   4. Causality (belief MacroDefn; evidence is logically related to the claim)
   5. Authority (another belief MacroDefn based on the Word of God, priests, constitutions, leaders, etc.)
   6. Principle (evidence→claim is instance of a broader, relevant principle)
4. *Backing* (evidence supporting the warrant; statistical, quotes, reports, findings, physical evidence, data, or reasoning)
5. *Counterarguments* and *rebuttals* (refuting the macro conditions)

   3 types of counterarguments (rebuttals mitigate the counterarguments):
   1. Counter the Toulmin Model that you have presented
   2. Propose separate arguments
   3. Challenge the definitions of the preconditions/postconditions
6. *Qualifier* (probability/trustworthiness of evidence, warrants, or claim)

A socially intelligent machine (sim) should be able to summarize the Toulmin reasoning that supports a qualified claim in the way that research papers are often organized [9]:

I. Introduction of the problem or topic.

A. Material to get the reader's attention (a "hook")

B. Introduce the problem or topic

C. Introduce our claim or thesis, perhaps with accompanying qualifiers that limit the scope of the argument. This thesis may involve what “should” be done to solve or ameliorate the problem.

II. Offer data (reasons or evidence) to support the argument.

A. Datum #1

B. Datum #2

C. (and so on)

III. Explore warrants that show how the data currently is logically connected to the qualified claim

A. Warrant #1

B. Warrant #2

C. (and so on)

IV. Offer factual backing to show that logic used in the warrants is good in term of realism (i.e. instances and examples) as well as theory (i.e. logical implications among intents).

A. Backing for Warrant #1

B. Backing for Warrant #2

C. (and so on)

V. Discuss counter-arguments and provide rebuttal

A. Counter-argument #1

B. Rebuttal to counter-argument #1

C. Counter-argument #2

D. Rebuttal to counter-argument #2

E. (and so on)

VI. Conclusion

Implications of the argument, summation of points, or final evocative thought to ensure the reader remembers the argument.

Here is an example [10] from John Gage's *The Shape of Reason* in which the various parts of an argument are labeled:

> Congress should ban animal research (Claim #1) because animals are tortured in experiments that have no necessary benefit for humans such as the testing of cosmetics (Data). The well being of animals is more important than the profits of the cosmetics industry (Warrant). Only congress has the authority to make such a law (Warrant) because the corporations can simply move from state to state to avoid legal penalties (Backing). Of course, this ban should not apply to medical research (Qualifier). A law to ban all research would go too far (Rebuttal). So, the law would probably (qualifier) have to be carefully written to define the kinds of research intended (claim #2).

Each of these data, claims (thesis) and its qualifiers, warrants, their backings, and their rebuttals can be defined by class algebra. Level 0 class algebra includes the classes that are necessary for class algebra to describe its own data (e.g. DNA remembering that plants should grow toward the sun). Level 1 allows class algebra to describe functions and operators and their changes of state (e.g. stimulus-response of animals). Level 2 allows class algebra to use such Toulmin arguments to decide when/how to change its social model (e.g. using models of self and others to predict outcomes of actions).

3. Class Algebra representation of Toulmin Arguments

A class-algebra learning algorithm arranges the world into classes (including ClassDefn, RelationDefn, and OperatorDefn) and their properties (i.e. intents), similar to most popular computer languages such as Python and Javascript. Properties are inherited by all subclasses, and extents are inherited by all superclasses. OperatorDefns have a set of edges (i.e. 1-1 functions), each with an input state and an output state, where the state can involve the Prerequisites and affected OutputParameters. For example, a "Move" operator can change the location of any physical object that is movable by given agents via some path. The movability probability function depends on the agent's power and location relative to the weight of the object being moved, if on Earth, and the sizes of the doors or openings through which the object must be moved. It is evident that the Move operator not only contains the "extent" of the union of its previously seen instances, but also a very complicated "intent" that logically summarizes the if-then-elseif algorithm that is stored as a union of given subclasses of

the Move class.

All of the RelationDefns (including OperatorDefns) and their relationships can be represented by a partially ordered graph where loops have been reduced to single nodes. Class algebra assumes that all relations are binary and have inverse relations, and the set of all possible "macro" relation sequences is finite because the number of intents (i.e. conjunctions of logical conditions) is finite.

Each RelationDefn must contain the properties "input" and "output". Each OperatorDefn (including the subclass MacroDefn) is a RelationDefn with inputs called "previousStates", and outputs called "successorStates". The transitive closure of these OperatorDefn edges gives the "reachableStates", while the transitive closure of the inverses produces the possible "startStates". Also, differences in input and output state intents can produce the "add" and "delete" operators for obtaining the intents of successorStates from previousStates.

A major part of the IS-A hierarchy is the classification of morphisms, which are mappings between relations.

1. IS-A superclass/subclass relations with inheritance of constraints from superclasses and inclusion of elements from subclasses. Each class, besides a set of superclasses and subclasses, contains:
   A. logical intent: an & of property ranges; a qualified claim, and
   B. extent: an explicit set of instances and counterexamples (i.e. unary sets on leaves of the IS-A hierarchy under this node or its pseudo-negation) which can be used to calculate relative frequencies.
2. Morphisms and analogies mapping properties of one class to "similar" properties of another class, even though the property and its operator's names are different. This includes the following examples:
   A. Solid "part-of" relations with inheritance of relative locations of parts to the "center of mass" via polar coordinates.
   B. Liquid and gas property inheritance of PV=nrT to subparts
   C. Other mathematical equalities and inequalities that seem to allow us to calculate property values or probability distributions.

In class algebra modeling, a major key point is to identify the IS-A hierarchy of operators, their parameter ranges, and the operator names in English or Chinese.

Counts of leaves or tops of the OperatorDefn paths (i.e. plans or algorithms) can be used to compute the relative frequencies of the startStates and reachableStates, which are used as the probabilities of fuzzy class algebra. By comparing these parameters (i.e. metrics) of the model to actual physical world measurements, we can use gradient descent or some similar curve-fitting methodology to find differences between the model's predictions and the actual outcomes of experiments. Reinforcement learning then allows the learning algorithm to adjust its relative weights or nonlinear Taylor series' moments to make the model more realistic.

4. Using Class Algebra to form Decision Graphs

The Splitting algorithm of decision trees can use various metrics to decide when to split a class into two subclasses based on the range of a parameter or its complement. The "information gain" measure of C4.5 is a good one, but the interactive and patented cognitive signature measure used as a hash function for searching semantic graphs can also be used to find the most "similar" parameter ranges in the *n*-dimensional space of *n* possible parameters.

Basically, the graphs of class algebra relations have the following operations:

1. Split: Find the parameter that most closely divides the set of instances of the parent's extent in half.
2. Merge: Union the intents of "similar" nodes. For instance, you can "feed" either cars or people. This "union" parent node of the feed Operators then represents this analogy.
3. Add an example: Use top-down intents to classify this example until getting to a leaf. If intents are satisfied, then simply add to the extent. If some intent ranges are not satisfied, then create the brother node with its intent containing a complement of some original intent literal.
4. Add an analogy or morphism: create a superclass of an operator or morphism with mappings to the two cases of this operator or relation.

5. Using Class Algebra to form Toulmin Arguments

It can be seen from the definition of Toulmin warrants that causality and properties (i.e. scientifically verifiable laws) are only two of six kinds of evidence.

They can obviously use logical entailment to implement their reasoning processes. How about the other kinds of warrants? Are they implementable in terms of fuzzy logic?

Generalization and induction are the main technique used by deep learning neural networks. They are subject to the problems of insufficient "big data" and over-fitting of that data. From the viewpoint of the Merge operation of the previous section, generalization mainly involves finding a good measure of "minimal description" for an intent that covers all of the cases below it. However, this approach is different from fuzzy logic or deep learning in that it involves the unions of exact formulas that include all of the given data.

As indicated in the previous section, it is quite easy to find mappings between properties of OperatorDefns/MacroDefns to find their common sub-properties. For example, phase changes (Bose-Einstein condensate, solid, liquid, gas, plasma) occur in many complex systems.

Signs (symptoms) are often used in medical applications, and techniques of deep learning have had some success, beating doctors at recognizing various diseases such as breast cancer from Xrays, eye diseases from retinal images, etc. Fuzzy logic with appropriate combination functions can be made pretty much algebraically equivalent to these deep learning/reinforcement algorithms.

Causality networks are basically the same as Bayesian networks, with causes and effects going either up or down depending on whether the networks are drawn in America or England. The sizes of extents can basically be used to calculate the relative frequencies of Bayesian networks. However, the fuzzy measures are only necessary when the coin's head/tail value cannot be directly measured (i.e. is a hidden variable).

Authority is basically a "Trustworthiness" measure for the source of the information. Obviously, for the religious, the "Word of God" has 100% trustworthiness, but the social teachings of religions may evolve as society changes. The so-called "experts" on things like climate change, AI, and genetic manipulation techniques may not have any idea about the ethical questions involved in their research. However, they can give us better information about possible futures, and how to design our future.

Scientific principles are often stated in first-order logic, which is also usually taken to be the basis of mathematics. Class algebra uses an intuitionistic logic to eliminate paradoxes such as the set of all sets that do not contain themselves and various

definitions of infinity (e.g. having 2.0(*2.0), 4.0, and 3.999999… as denotations for the same concept). The intuitionistic logic of class algebra is decidable, all equivalent intents must be merged, and any description must be finite and have a unique normal form and unique class name. The intuitionistic logic of class algebra also uses closure and subsumption algorithms that are much simpler than for first-order logic since it does not involve existential variables like $x=f(x)$ (i.e. its resolution/subsumption algorithm has no need for an "occurs" check). The parameter x cannot be used in its own definition unless there is a way to prove that there is a measure of a limit expression which has a sequence of natural numbers that goes to 0, (i.e. only well-defined parameters and class definitions are permitted, and all limits have a unique closed form and class name).

6. Using Class Algebra to form Reversed Toulmin Arguments

Very often people are unconvinced by arguments based on logic or probabilities. For instance, try to get an alcoholic to stop drinking based on the possibility that he will have a car accident. It's usually more effective to have him change his friends, his religion, or listen to what happened to alcoholics that he knows well. That is, reversed Toulmin arguments can tell us "how to" choose convincing arguments, "how to" find frequencies to/from causes and effects, "how to" most effectively reach given goals from the fuzzy model's current states, and how to effectively arrange arguments into a dialogue or debate, based on the personality of the advisee (i.e. which types of Toulmin arguments have previously been the most successful with the advisee).

7. KL-Divergence for Modifying Sim Models

Peter Sweeny summarizes the main point of this white paper [9]:

> "And just as various schools of philosophy converged on the consensus of explanations, various schools of AI are converging on the foundation of explanations to create *good* knowledge."

These explanations give us the ability to say, "You shouldn't have done that. You should have …". As Judea Pearl said [3], this ability to reason about alternative actions entails that these moral machines have "free will". Whether or not this entails that they also have a spirit that continues forever in another world, with some ability to influence

things in this world of space-time, is a potential subject for future research (e.g. [11]).

The above analysis as well as recent research indicate that intelligible intelligence [12] is like self-flying taxis that will overtake the Model-T of good old-fashioned AI (GOFAI) and Model-S of deep-learning networks. Intelligible intelligence can be modeled by either mathematical models involving something like Kullback-Leibler divergences (i.e. KL-divergences) or by chemical algorithms involving weights like dopamine (i.e. a measure of surprise at the difference from the model's predictions). The difference between the model's prediction and the real-world outcomes indicates how much weight should be used in updating the conductivity of the synapses involved in that decision. If the model is correct, there will be no KL-divergence between the probability functions of the models at time t and t+1, there will be no surprise (i.e. dopamine), and there will be only the addition of one more event into the extent of the Warrant that was chosen to meet the given goal. However, if the KL-divergence is large, then the conductivity of the synapses involved in this Warrant must be decreased in proportion to the original probability (i.e. relative frequencies), the trustworthiness of the Warrant's source, the qualifiers involved in the Warrant, and the other measures of how well this case was similar to previously-seen cases.

The notation that is traditionally used for KL-divergence reflects its correspondence to the use of conditional probabilities in fuzzy logic. Conditional probabilities are defined by an equation such as

$$P(x|y) = P(x\&y)/P(y)$$

which generalizes to

$$P(x_1'\&\ldots\&x_n' \mid y_1\ldots y_n)=P(x_1\&\ldots\&x_n \,\&\, y_1 \mid y_2\ldots y_n)/P(y_1)$$
$$=P(x_1\&\ldots\&x_n\&y_1\&\ldots\&y_n)/(P(y_1)P(y_2)\ldots P(y_n))$$

Basically, this equation says that the probability $P(x_1'\&\ldots\&x_n'\&y_1'\&\ldots\&y_n')$ of $x_1'\&\ldots\&x_n'\&y_1'\&\ldots\&y_n'$ being true or false in the next state (i.e. indicated by the "'" notations) depends on the product of the probabilities $(P(y_1)P(y_2)\ldots P(y_n))$ of $y_1\ldots y_n$ all being true in the current state, times the conditional probability $P(x_1\&\ldots\&x_n \mid y_1\ldots y_n)$ that $x_1\&\ldots\&x_n$ is true in both the old state and is equal to $P(x_1'\&\ldots\&x_n' \mid y_1'\ldots y_n')$ in the new state after the corresponding operator is applied. Notice that the variables in the above equations are conditions of the form "DottedExpression Op DottedExpression", where DottedExpression is a class name followed by zero or more dotted relation names

or operator names. The operators Op can be any of the comparison operators that are used for acyclic graphs (i.e. ⊇, ⊆, =) or their complements (i.e. Op, ~Op, -Op, and ~-Op). If any of the true complements $\sim y_i$ are true in the old state, there is nothing that can be said about the new (caused, entailed, generalized) model's state since the product $(P(y_1)P(y_2)\ldots P(y_n))$ goes to zero if any of the true complements are 1 since P(x)+P(~x)=1 for all x. The pseudo-complement P(-x) is related to the true complement by the equation 1= P(x)+P(-x)+P(~y) where x and -x are the names of the two subclasses of class y where condition x is either true or false. In terms of relative frequencies, the equation P(x|y)=P(x&y)/P(y) represents the fraction of y's extent that is in either of the four possible subclasses of y, which add x, ~x, -x, or ~-x to their extent.

Unlike probabilities, the KL-divergence of two probability distributions is only a pre-measure, not necessarily satisfying the symmetry axiom of a measure. However, just as the implication operator is not symmetric, the inverse implication operators can be defined by using Boolean algebra operators. If we use the standard definition $x \rightarrow y = \sim x \vee y$, then the algebra of classes satisfies the Boolean axioms of either classical logic or various intuitionistic logics.

For discrete probability distributions *P* and *Q*, the Kullback–Leibler divergence from *Q* to *P* is defined to be

$$D_{KL}(P||Q) = -\sum_i P(i) \log(Q(i)/P(i))$$

which is equivalent to

$$D_{KL}(P||Q) = \sum_i P(i) \log(P(i)/Q(i))$$

It is basically the expectation of the logarithmic difference between the probability distributions P and Q, where Q is usually the distribution of the model and P is the real-world distribution. It is defined only if, for all *i*, Q(*i*)=0 implies P(*i*)=0. KL-divergence is the entropy of P relative to Q, where information is measured in terms of bits if $\log_2$ is used, and is measured in "nats" if ln is used for log. This measure is used to measure relative (Shannon) entropy, randomness in time series, and information gain of fuzzy statistical models of inference. Divergence is basically a measure of the difference in information when going from state Q to P. In terms of Bayesian inference, $D_{KL}(P||Q)$ is a measure of the information gained when one revises one's beliefs from the prior probability distribution Q to the posterior probability distribution P. In other

words, it is the amount of information lost when Q is used to approximate P. Basically, KL-divergence measure the "surprise" in neuroscience and machine learning. If zero, no changes are required to the computer's model, but as it approaches 1, the pseudo-complement of the original distribution must be asserted. As in class algebra, it is possible to have a model with both x and -x having many instances on the leaves below these classes in the IS-A hierarchy. Relative frequencies, however, all satisfy the laws of probability.

Most models of firings of synapses of the brain simply involve the basic arithmetic operations of addition, subtraction, multiplication, or division. Although the brain may have modules for simple arithmetic, similar to chemical algorithms for counting [13], it is more likely that the chemicals reactions themselves implement the propagation of relative frequencies and update weights of various evidence, such as the updating due to the six kinds of warrants of Toulmin arguments.

8. Structural Causal Models that Assist Humans

Judea Pearl [14] states seven "pillars" of the structural causal revolution that has revolutionized AI and other fields such as law. These include mathematical notations for intervention in the world (i.e. do(Action) as a parameter of what look like conditional probability distributions) and notations for counterfactual argumentations, involving introspection and retrospection using questions like "What if _ had", "What were the causes", "Plan a proof", etc. His theory of Structural Causal Models (SCM) involves graphical models, structural equations, and counterfactual and interventional logic.

SCM has some similarities to the Toulmin model of argumentation. SCM's assumptions are similar to Toulmin's warrants. Its data are similar to the extents of classes, and its generalizations are similar to superclasses. The hierarchy of classes that represent Operators (including macros) are similar to a hierarchy of plans, and the "unless", "since", "except", etc. qualifiers of extended Toulmin arguments (e.g. [15]) are similar to SCM's counterfactual argumentations. The KL-divergence described in the previous section is similar to SCM's fit indices, which can be used to update its model of the world by adjusting neural-network like weights in the subclass ontology.

9. Constructing Moral Machines

Kyndi [16, 17] (no affiliation to us) has already implemented many of the ideas discussed in this white paper, such as natural language to ontology translators (based on minimizing descriptions via semantic graphs) and a patented hashing algorithm for ontology (i.e. semantic net) queries. However, most of its funding comes from the defense department. Without an appropriate initial ontology and appropriate goals, its AGI could easily discover that the terrorist methods of blowing up buses of school children or destroying world trade centers is the most effective method of controlling human populations. Even if it decides that propaganda and Kompromat are sufficient means of control, humans still essentially become its slaves. If we want to maintain our humanist ideology of freedom of speech and majority rule [18], let's design the moral machine's initial ontology before humans and moral machines lose our free wills (i.e. the ability to choose operators based on our social model of reactions of others and our own reactions to those reactions).

References:

1. Stuart Russell, https://www.ted.com/talks/stuart_russell_3_principles_for_creating_safer_ai ?utm_source=facebook.com&utm_medium=social&utm_campaign=tedspread
2. Shelly Fan, Amazing New Brain Map of Every Synapse Points to the Roots of Thinking, https://singularityhub.com/2018/08/14/amazing-map-of-every-synapse-in-the-mouse-brain-points-to-the-roots-of-thinking.
3. Kevin Hartnett interview of Judea Pearl, https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/
4. Carlos E. Perez, https://medium.com/intuitionmachine/ego-motion-in-self-aware-deep-learning-91457213cfc4
5. Johan Kwisthout, "Precision Weighting of Prediction Errors" (Dec. 14, 2017) http://www.socsci.ru.nl/johank/seminar.html
6. Virtues for Life: the heart of everyday living, Virtues List (Aug. 16, 2018) http://www.virtuesforlife.com/virtues-list/
7. Toulmin Model of Argument https://web.cn.edu/kwheeler/documents/Toulmin.pdf
8. Daniel Buehrer, https://arxiv.org/ftp/arxiv/papers/1804/1804.03301.pdf

9. Kullback-Leibler Divergence, (last edited on Aug. 13, 2018), https://en.wikipedia.org/wiki/Kullback%E2%80%93Leibler_divergence
10. Peter Sweeney, "One Problem to Explain How AI Works", https://towardsdatascience.com/one-problem-to-explain-ai-218a29e8fbc0
11. University of Virginia Department of Perceptual Studies Faculty: Do We Survive Death? A Look at the Evidence. https://www.youtube.com/watch?v=ZoqNe-U53wA
12. Max Tegmark - How Far Will AI Go? Intelligible Intelligence & Beneficial Intelligence, https://www.youtube.com/watch?v=tAdvbaQQDA4&feature=share
13. Deepmind Researchers Develop Neural Arithmetic Logical Units https://techxplore.com/news/2018-08-deepmind-neural-arithmetic-logic-nalu.html?utm_source=nwletter&utm_medium=email&utm_campaign=daily-nwletter
14. Judea Pearl, "Theoretical Impediments to Machine Learning with Seven Sparks from the Causal Revolution", arXiv 1801.04016v1, Jan. 11, 2018.
15. Susan Newman and Katherine Marshal, "Pushing Toulmin Too Far: Learning from an Argument Representation Scheme", http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.23.268&rep=rep1&type=pdf
16. https://kyndi.com/blog/time-side-building-ontologies-kyndi-platform/
17. https://patentscope.wipo.int/search/en/detail.jsf?docId=WO2016154460
18. Yuval Harari – The Challenges of the 21st Century https://www.youtube.com/watch?v=FSloTpkHYYI&feature=share