Dynamic Models Applied to Value Learning in Artificial Intelligence

Nicholas Kluge Corrêa[1] and Nythamar de Oliveira[2]

[1]Master of Electrical Engineering and PhD student in Philosophy - PUCRS
nicholas.correa@acad.pucrs.br, ORCID: 0000-0002-5633-6094

[2] Ph.D. in Philosophy (State University of New York, 1994), Full Professor - PUCRS
nythamar@yahoo.com, ORCID: 0000-0001-9241-1031

Graduate Program in Philosophy of the Pontifical Catholic University of Rio Grande do Sul – Av. Ipiranga, 6681 - Partenon, Porto Alegre - RS, 90619-900.

Abstract

Experts in Artificial Intelligence (AI) development predict that advances in the development of intelligent systems and agents will reshape vital areas in our society. Nevertheless, if such an advance is not made prudently and critically-reflexively, it can result in negative outcomes for humanity. For this reason, several researchers in the area are trying to develop a robust, beneficial, and safe concept of AI for the preservation of humanity and the environment. Currently, several of the open problems in the field of AI research arise from the difficulty of avoiding unwanted behaviors of intelligent agents and systems, and at the same time specifying what we want such systems to do, especially when we look for the possibility of intelligent agents acting in several domains over the long term. It is of utmost importance that artificial intelligent agents have their values aligned with human values, given the fact that we cannot expect an AI to develop human moral values simply because of its intelligence, as discussed in the Orthogonality Thesis. Perhaps this difficulty comes from the way we are addressing the problem of expressing objectives, values, and ends, using representational cognitive methods. A solution to this problem would be the dynamic approach proposed by Dreyfus, whose phenomenological philosophy shows that the human experience of being-in-the-world in several aspects is not well represented by the symbolic or connectionist cognitive method, especially in regards to the question of learning values. A possible approach to this problem would be to use theoretical models such as SED (situated embodied dynamics) to address the values learning problem in AI.

Keywords: Artificial intelligence, Value learning, Cognitive science, Dynamical cognition.

I. Introduction

Researchers and specialists in Artificial Intelligence (AI) development stipulate that within 10 years many human activities will be surpassed by machines in terms of efficiency. Several aspects of our public policies will need to be modified to accommodate such advances, which promise to reshape areas such as transportation, health, economics, military fighting, lifestyle, etc. (GRACE et al. 2017). There is also concern about the risks that machines with a high level of human or superhuman intelligence may bring to humanity in the coming decades. A survey conducted by Müller and Bostrom (2016) consisted of building a questionnaire to assess progress in the field of AI research and prospects for the future, interviewing various experts in the field. The questionnaire showed that, on average, there is a 50% chance that high-level (human) machine intelligence will be achieved between 2040 and 2050, reaching a 90% probability by 2075. It is also estimated that this intelligence will exceed human performance in 2 years (10% chance) to 30 years (75% chance) after reaching human intelligence levels (MÜLLER; BOSTROM, 2016).

However, in the same survey, 33% of respondents classified this development in AI as “bad” or “extremely bad” for humanity (MÜLLER; BOSTROM, 2016). As there is no guarantee that such systems will be “good” for mankind, we should investigate further the future of superintelligence and the risks it poses to the human race. Some several open questions and problems need to be solved. How will we remedy the economic impacts of AI to avoid negative effects such as mass unemployment (FREY; ORSBORNE, 2013)? How can we prevent the self-motivation of jobs from pushing the distribution of income into law of disproportionate power among classes, genders, and races (BRYNJOLFSSON; MCAFEE, 2014)? Can autonomous lethal weapons be built without changing humanitarian rights, and, should autonomous weapons be completely banned (DOCHERTY, 2012) (CHURCHILL; ULFSTEIN, 2000)? How can we ensure privacy by applying machine learning to confidential data such as medical data sources, phone lines, emails, online behavior patterns (ABADI et al. 2016)? How can we understand what complex AI systems are doing to iteratively classify and reconstruct images from neural networks (MORDVINTSEV, OLAH, TYKA, 2015)?

Some researchers have already created models (ASI-PATH) of how an AI could cause some kind of catastrophe, becoming super-intelligent through recursive self-improvement (BARRET; BAUM, 2017), something known in the AI literature as a Singularity. Such models suggest scenarios where intelligent agents, after obtaining some kind of strategic advantage (DSA - decisive strategic advantage or MSA - major strategic advantage), such as advances in nanotechnology or robotics, could achieve considerable power of domination (BOSTROM; ĆIRKOVIĆ, 2008). The scenarios suggest different types of takeovers by artificially intelligent systems, ranging from fast takeoffs, situations where a drastic takeover by such systems occurs, to slow takeoffs, where gradually the human race becomes more dependent and, to some extent, under the control of AI (SOTALA, 2018).

The development of an AI ethic presupposes, in fact, the intuitive formulations of Isaac Asimov's so-called Three Laws of Robotics (1950), at a time when this theme still seemed relegated to the realm of science fiction - recalling that such ethical-moral codifications were introduced in a 1942 tale, Runaround: “(1) A robot may not harm a human being or, by inaction, allow a human being to be harmed; (2) a robot must obey the orders given by human beings, except where such orders conflict with the First Law; (3) a robot must protect its existence, provided such protection does not conflict with the First or Second Law. In our century, this ethical orientation of not harm mankind was extended not only to robots and robotic artifacts but to machines and intelligent devices generally associated with AI resources.

Thus, Shulman (2010, pp. 2) suggests a model that explains in which situations an AI would abandon cooperation with the human race and take hostile action, in which an artificial agent that believes it has a P probability of being successful, if it initiates aggression, receiving some expected utility $[EU\ (Success)]$, and with a $(1 - P)$ probability of failing, receiving $[EU\ (Failure)]$. If it gives up the aggressive strategy, the agent receives utility $[EU\ (Cooperation)]$. The AI will rationally initiate the aggression only if:

$$P \times EU(success) + (1 - P) \times EU(failure) > EU(cooperation)$$

II. Safety Issues in AI

Ultimately, there is a consensus in the literature: AI development must be done in a safe, beneficial, and robust manner. An article published by Amodei et al. (2016) entitled “Concrete Problems in AI Safety” lists several open problems in the field of AI research that must be addressed if we are to reap the benefits of AI without compromising our safety. These problems are classified into specification and robustness problems and are the current barriers to be overcome in the area (LEIKE et al. 2017).

To better synthesize and develop the content of this study, we will refer briefly only to specification errors. Specification errors occur when the utility function of the AI is poorly specified by programmers, causing unwanted and even harmful results, even if the learning is perfect with explicitly clear data (AMODEI et al. 2016). Some examples of specification errors are negative side effects, reward hacking, and safe interruption (corrigibility).

Negative side effects occur when the maximization of the reward function focuses on achieving a goal while the agent ignores important factors in the environment, causing potential cross effects. In reward hacking, the AI agent finds a solution to its goal that maximizes its reward function, but unexpectedly, perverting the intention of the programmers (AMODEI et al. 2016). The Safe Interruption or Corrigibility concerns how we can be able to interrupt an agent if it is behaving unexpectedly, and in a certain way, correct the detected errors without the agent opposing to interruptions. (SOARES et al. 2015).

Two theses published by Bostrom, (2012), firstly proposed by Omohundro (2008) in his seminal paper "The Basic AI Drives", point out how these problems can present a risk. The Thesis of Instrumental Convergence shows us how a series of self-improvement and preservation goals can be pursued by any intelligent agent with a terminal goal. We can formulate this thesis as follows:

> Several instrumental objectives can be identified, which are convergent in the sense that their attainment would increase the chances of the agent's terminal objective, implying that these instrumental objectives are likely to be pursued by any intelligent agent (BOSTROM, 2012, p. 6).

Without careful engineering of these systems, risks with an "intelligence explosion" (the exponential increase in the cognitive capacity of the agent) can create agents much more powerful than our ability to control them. On the other hand, and correlated to the first thesis, the Orthogonality Thesis proposes that intelligence and final objectives have independent and orthogonal properties. The hypothesis is argued as follows:

> Intelligence and ultimate goals are orthogonal axes along which possible agents can freely vary. In other words, more or less any level of intelligence could, in principle, be combined with more or less any final objective (BOSTROM, 2012, p. 3).

The thought behind the orthogonality thesis is analogous to the so-called Hume's Guillotine (also known in English as Hume's fork or Hume's law), opposing what is factually and empirically verifiable (matters of fact and real existence) to what should be, in rational terms, normative and counterfactual (relations of ideas). Hume observed a significant difference between descriptive statements and prescriptive or normative statements, and therefore, it would not be obvious, self-evident (self-evident) or valid (valid) to derive the latter from the former. The undue passage from being (Is) to being (Ought), which would be one of the seminal problems of research in metaethics, normative ethics and applied ethics in the twentieth century, was noted by the Scottish philosopher in a famous passage in section I of part I of his Treatise of Human Nature:

> In every moral system I have encountered to date, I have always noticed that the author follows for some time the common way of reasoning, establishing the existence of God, or making observations regarding human affairs, when suddenly I am surprised to see that, instead of the usual propositional copulations, as it is and is not, I do not find a single proposition that is not connected to another by one should or should not. This change is imperceptible but of the utmost importance. For as this must or must not express a new relationship or affirmation, it would need to be noted and explained; at the same time, it would need to give a reason for something that seems inconceivable, that is, how this new relationship can be deduced from entirely different ones (HUME, 2009, p. 509).

Just as descriptive, purely factual statements can only bind or imply other descriptive or factual statements and never standards, the problems of orthogonality and value alignment consist in guaranteeing, if an AGI (artificial general intelligence, that is, a hypothetical intelligence of a machine with the capacity to understand or learn any intellectual task that a human being can perform) were to develop enough intelligence to have power over the human species, that such intelligence would do with human beings only what we would wish or accept it to be done.

In this sense, the problem of alignment is identical to what we see in moral philosophy about utilitarianism, in that the maximization of utility by some moral agent can culminate in morally repugnant conclusions, including the violation of the rights of others. Although it may guarantee the resolution of tasks in computational time (polynomial), the mere efficiency or optimization of procedures does not ensure normative universalizability (as it would be, moreover, a basic premise of ethical deontological and non-utilitarian models) and may eventually conflict with the interests or rights of other people. We should also note that the ethics of artificial intelligence is part of the ethics of technology in general and, specifically, for robots, learning machines, and other artifacts and artificially intelligent entities.

In our approach, the AI ethic comprises both robotics (robotic ethics), which is concerned with the moral behavior of human beings when designing, building, using and programming artificially intelligent beings, and a machine ethic, which is concerned with the moral behavior of artificial moral agents themselves. Both bioethics and neuroethics would have much to learn, to teach and to interact with the ethics of artificial intelligence, especially through the interface of artificial life models, genomic editing and neural networks with the ethical-normative challenges of orthogonality, value alignment, and transhumanism, integrating the neurobiological, cultural and technological legacies of the *homo sapiens sapiens.*

Anthropomorphic bias tends to shape the entire spectrum of possible minds and intelligence, but this is a mistake, known as the Fallacy of Mind Projection (JAYNES, 2003). On the contrary, we consider intelligence as a function of optimizing an agent's ability to achieve goals in a wide variety of environments with limited resources (LEGG, 2008). To best exemplify this thinking, we use a quote from Dijkstra, (1984), "*the question of whether a machine can think is as relevant as the question of whether submarines can swim*". The upper limit of brute processing for the whole known universe, imposed by the laws of physics, is 10120 operations in 1090 bits (10120 bits including the degrees of gravitational freedom) (LLOYD, 2002). The human level of information processing is 1011 operations per second (MORAVEC, 1998). This difference between the human level and the highest possible degree of optimization leaves open a wide range of possible levels of superhuman intelligence (SOTALA, 2010). In the Kantian sense, reason can be defined as the ability to obtain logical inferences or, in a systematic way, the ability to synthesize in unity, through comprehensive principles, the concepts provided by the intellect, in that agents use reason to establish and pursue ends (goals, purposes, Zwecken), using the rest of nature as a means to their ends. Humanity is thus considered as an end in itself and a terminal end of nature (ALLISON, 1996).

For these reasons, the alignment of values between AI and humans is an important problem to be solved in the area of machine ethics (SOARES; FALLESTEIN, 2014). Practically all problems of specification, robustness, and value alignment seem to occur at the same point when our representations of values or final objectives (goals) lose their meaning or are misinterpreted. Is the objective-representational approach doomed to error? Would the cognitive models used in the creation of artificially intelligent agents, especially symbolism and connectionism, be incapable of expressing the meaning of human values? If so, would there be any alternative?

III. Cognitive Models: Symbolism and Connectionism

Since the late 1950s, the discussion about cognition and intelligence has been permeated by the computational framework, also known as a symbolic view. This perspective starts from the assumption that cognitive systems are intelligent in that they can encode knowledge into symbolic representations. Symbolists believe that through sets of "if-then" rules and other forms of calculation for symbolic algorithms, all cognition is accomplished by manipulating such representations (THAGARD, 1992).

Newell, (1990), defined the computationalist proposal, which is also referred to as the "Physical Symbol System Hypothesis", as follows: Natural cognitive systems are intelligent by being physical systems that manipulate symbols in such a way as to present intelligent behavior, codifying knowledge about the external world in symbolic structures (NEWELL, 1990, pp. 75-79). Newell has dedicated much of his work to building systems that express his vision of a physical symbol system. His most promising model is known as SOAR. SOAR is a symbolic computational system that formulates its tasks based on symbol and goal hierarchies, thus generating an algorithmic production and decision making system for problem-solving (NEWELL, 1990, p. 39).

of cognition are high-level effects that depend on lower-level phenomena. Thus, the connectionist hypothesis encapsulates the idea that the fact that most determine the cognitive capacity of an agent is not the ability of representative manipulation, but its architecture. Thus, connectionists attack the problem of cognition by performing reverse engineering on the central nervous system, copying its basic processing unit, namely the neuron (Churchland and Sejnowski, 1992, p. 2). Sejnowski (1988, p. 7) notes in his connectionist hypothesis: "*The intuitive processor is a dynamic sub-conceptual connectionist system that does not admit a complete, formal and precise description on a conceptual level*".

Thus, theories of cognition in AI (symbolism, connectionism, and dynamism) can be considered theoretical structures, since they provide us with the filters, analogies, and metaphors by which we try to understand the phenomenon of cognition, and thus create theoretical models that can generate simulations to be tested (BEER, 1998). The symbolism, for example, highlights the internal representations of the system or agent, and the algorithms by which these representations are manipulated. Connectionalism emphasizes the neural network architecture, the learning algorithm, the preparation of training data, and the protocol used (ELIASMITH, 1996).

However, the limitations of the symbolic computational hypothesis, especially in the aspects of time, architecture, computing, and representation, led researchers to consider new theoretical models, such as the dynamic hypothesis (van GELDER, 1998). And as much as the connectionist model is similar to the dynamic model in the aspects pointed out in the symbolic model (time, architecture, computing, and representation), the connectionist model still fails to produce agents that can solve the above-mentioned problems of specification and robustness.

In this article, we do not adopt an anti-representationalist position, as we humans constantly use and manipulate representations, as in language, writing, speech, music, and other forms of abstract thinking. However, we skeptically position ourselves concerning the function of representations in systems that involve value-objective-methods, and therefore goal-oriented behavior. Perhaps, in some cases, the roles played by the internal states of a cognitive agent simply cannot be interpreted as representative (FRANKISH; RAMSEY, 2014).

IV. Criticism of the Symbolic Method

One of the biggest criticisms raised against the symbolic computing model is the difficulty in meeting time constraints. When trying to replicate the phenomenon of cognition van Gelder and Port, (1998, p. 2) states that the symbolists "*leave time out of the picture*". Since the objective of cognitive science is to describe the behavior of natural cognitive agents, and by definition, these agents operate in real-time, a cognitive model that replicates the human experience of cognition must present real-time cognitive processes (in the case of humans: ± 10 milliseconds) (van GELDER; PORT, 1998).

The limits imposed by symbolic architecture are another source of criticism of the computational method. For Newell, (NEWELL, 1990, p. 82), the behavior is determined by a variable content being processed by a fixed structure, which is the architecture. Dynamists criticize this view of the cognitive system as "*a box*" within a body, in turn within a physical environment. However, where do we draw the line that divides the box from your body? And, more controversially, the body with the environment? Van Gelder and Port, (1998, p. 8), analyze the internal architecture in the cognitive agent as not being a fixed structure, where all aspects of cognition, brain-body-environment, as mutually influencing each other continuously.

Consequently, this view of architecture often refers to the symbolic method as a computational method, because it describes the mind as a special type of computer. This characterization is following the architecture proposed by Newell, (1990), and identifies the mental computer with the brain. The body, through the sensory organs, delivers to the cognitive system (brain) representations of the state of its environment; the system on its part calculates an appropriate response and the body carries the action (van GELDER; PORT, 1998, p. 1). However, this system of perceiving-planning-acting ignores important phenomena in decision makings, such as reflex actions, and the speed with which such actions are expressed in real cognitive agents, showing once again that the symbolic computational method has no basis with the biological and physical reality of the cognition phenomenon.

Hubert Dreyfus (1992) was one of the most prominent critics of the symbolic representational approach in the field of AI research. Based on the hermeneutic-existentialist philosophy proposed by Martin Heidegger, Dreyfus indicated in his works that the manipulation of symbols and representations is not enough to generate the non-representational type of existence of a being in the world (*Dasein*). At the bottom of this impasse, there remains a criticism of materialist Cartesian thought and subject-object dualism: materialist Cartesianism that attempts, without success, to replicate the whole world "inside the mind" is doomed to fail according to Dreyfus, because it is impossible to contain the world inside the mind for the simple fact that the world is infinitely complex and we are finite creatures (DREYFUS, 2007). Thus, a self-contained, rigid system is not capable of duplicating the type of cognitive agent we desire. Perhaps this indicates to us that representations and experience must operate together for the former to have meaning.

V. Connectionism and Value Learning

We can see that many of the problems mentioned above come from the difficulty of programmers in expressing the meaning of what is proposed by the language (specification errors) and how this should change when the context of the environment evolves (robustness errors). Be it the representative cognitive model, using rules of behavior (if-then), or the connectionist model, using artificial neural networks with reward functions, we still reach the same impasse. How to express our goals and align the values of artificially intelligent agents with ours?

The related approach encounters several difficulties in this task, which are explored in more detail below. Commonly artificial neural networks are trained in a supervised manner, using labeled training data, however, this method may not be the safest for value learning. Dreyfus and Dreyfus (1992), cite an example where a machine learning system is trained to classify, or not, military ground vehicles hidden among the trees. The classifier during the training was able to identify with great precision the desired vehicles, however, the system had a fuzzy performance with images outside the training group. It was later discovered that the set of photos used for training containing vehicles were taken on a sunny day, while the images without the vehicles were made on a cloudy day. What the classifier was identifying was the brightness of the images. Potentially, learning values by induction is susceptible to this failure (SOARES, 2016).

For this reason, it is expected that artificial intelligent agents possess a property called corrigibility. Such systems must have their reward function or value hierarchy adjusted in case something unwanted happens. However, it is also necessary that the same agents cannot influence their learning environment or reward function, much less prevent it from being modified. There are currently no solutions to this problem (SOARES; FALLENSTEIN, 2015). Besides, both supervised training methods, which use labeled data and reinforcement learning, which use utility functions as a proxy for desirable results, are extremely vulnerable in identifying ambiguities (SOARES, 2016), as evidenced by “Sorcerer's Apprentice” problems and situations where the system, due to divergence in testing environments and new environments, and also goal misspecification, have the opportunity to hack its reward (BOSTROM, 2014). The reward hacking scenario, or “wireheading,” is wrongly compared to humans stimulating their pleasure (e.g. drug use). Human appetite is satiable; an artificial agent with the power to maximize its reward will not stop its “compulsive” behavior. It will even seek ways and means to perpetuate its self-compensating behavior free from interference (OMOHUNDRO, 2009).

The utility function can be explained by the von Neumann-Morgenstern utility theorem (von NEUMANN; MORGENSTERN, 1953). The theorem configures utility functions through preference sorting: A is preferred to B, or B is preferred to A, or both have the same preference value. A utility function allows that, given the state of the agent and the state of the world in general, an agent decision is generated between two or more options. The concept of the utility function is a mathematical formalization for the notion of human values and is widely used in economics and decision theory. However, one of the best-known problems of this model is the empirical fact that humans violate the axioms of utility theory and do not have consistent utility functions (TVERSKY; KAHNEMAN, 1981).

An alternative would be to model the intent of operators using inverse reinforcement learning (NG; RUSSELL, 2000): where one agent tries to identify and maximize the reward function of some other agent in the environment (usually a human operator). However, human preferences cannot necessarily be captured by observations alone, and if they are modeled optimally inverse reinforcement learning demonstrate the problem of learning “errors” or biases of human behavior as valid solutions. Recent advances in the area, such as the CIRL (Cooperative Inverse Reinforcement Learning) training model would solve this problem: instead of estimating and adopting the human being's reward function as its own, the system tries to solve a POMDP (Partially Observable Markov decision process), leading to a cooperative learning behavior, in which the system or agent tries to maximize the operator's reward function, but without knowing what it is (HADFIELD-MENELL, 2016). However, this approach generates problems of interpretation, such as the identification of ambiguity and coordination problems between the agents involved in POMDP.

Moreover, situations where humans are part of the reward system of an AI, also called human-in-the-loop, are not considered safe, as there is strong evidence to believe that artificial intelligent agents would be inclined to manipulate the human part of their reward mechanism if it meant an increase in reward (HIBBARD, 2012; BOSTROM, 2014). In general, our current training methods for the connectionist cognitive model are not appropriate for an AI or IAG (general artificial intelligence) operating in the real world. Possible scenarios of self-improvement, or even an “*intelligence explosion*”, as explained by the Instrumental Convergence Thesis (BOSTROM, 2012), can generate calamitous consequences for humanity (YUDKOWSKY, 2008). The ultimate goal of these agents is to maximize the reward, being our values and goals only instrumental to their ultimate goal. Such agents can learn that human goals are instrumentally useful for high rewards, but replaceable, especially if the intelligence of these agents is superior to ours (DEWEY, 2011).

Whether by symbolic representativeness or by connectionist training, so far value objectives cannot be safely expressed, and given the importance of human value alignment with AI, new methods must be investigated. We propose in this article that the dynamic cognitive model offers a new way of thinking about the problem of alignment. In the following section, we will discuss some of the characteristics of the theoretical dynamic model of cognition.

VI. Dynamic Cognitive Model

It can be said that many theoretical models begin as metaphors or analogies, later becoming theories that can be implemented in models and subsequently simulated. The conceptual structures that we form through this process can have a great impact on the way we conduct our studies, the way we approach the problem, the language we describe the phenomena, and the way we formulate a question and interpret an answer. The theory of dynamic systems invites us to think about the phenomenon of cognition and human experience in a progressive way, as proposed by Van Gelder (1998, p. 4), whose Dynamic Hypothesis postulates: "*Natural cognitive systems are certain types of dynamic systems, and are best understood from the dynamic perspective*". Dynamic systems, in this sense, are systems in which, as they evolve in time, their variables are continuously and simultaneously determining the evolution of one of the others, in other words, they are systems governed by non-linear differential equations (van GELDER; PORT, 1998, p. 6). With this statement, the dynamist puts the agent in a situation of coupling with the environment, turning brain-body-environment into an autonomous cognitive dynamic system where it no longer makes sense to talk about cognition or experience without recognizing the three aspects of this triad (van GELDER; PORT, 1998, p. 23).

A dynamic system is a mathematical abstraction composed of a space of $s_i \in S$ states, a set of time-ordered $t_i \in T$, and an evolution operator φ that transforms one state to another along T. S can be numeric or symbolic, continuous, discrete or hybrid, of any topology or dimension. T is typically expressed by the set of integers or real numbers, and the evolution of the operator φ can be deterministic or stochastic (KUZNETSOV, 2004). The situated activity has its philosophical origins in the phenomenological work of Heidegger (2012), which Dreyfus (1992) applying it to the field of AI, in which it is assumed that the Heideggerian agent cannot be separated from the environment or its interpretative context. Gibson's Ecological Psychology (1979) is also a precursor of situated activity, with its notion of affordances: Gibson emphasizes the environment-organism relationship in the phenomenon of perception as a two-way street, where one perceives to act, and acts to perceive. The idea of situated cognition can be extended to theories such as "*extended mind*" (CLARK & CHALMERS, 1998), also known as ECH (extended cognition hypothesis) (ROCKWELL, 2010), which invites us to think in a different way concerning Cartesian thought that places the imprisoned mind inside the brain. We explain gravity as the relationship between gravitational fields; electromagnetism by electromagnetic fields; the position of subatomic particles is expressed through probabilistic waves using Schrödinger's equation, De Broglie's wavelength, and Heisenberg's uncertainty principle. Thus, it seems likely that a sophisticated theory explaining the consciousness and experience of a cognitive agent involves some kind of theory that refers to the dynamic fluctuation of fields.

The theoretical model we present in this article is SED (situated embodied dynamics), proposed by Beer (2000), which emphasizes how the cognitive experience arises from the dynamic interaction brain-body-environment. In the first place, SED takes into account the situation as being fundamental to cognition, placing concrete action, that is, literally acting in the world, as something more fundamental than the abstract descriptions of this action. Thus, the final work of the intelligent agent is to act, an action that occurs in an environment, which is a central part of the behavior, since it is what gives meaning and context to the action. And the interaction of the agent with the environment is mutual, not being the environment just a source of problems to be solved, but a partner with whom the agent is involved from moment to moment (FRANKISH; RAMSEY, 2014). In the SED approach, the concept of embodiment says that the physical form and its functional and biomechanical aspects are essential aspects for behavior, as well as its biology and physiology, in the case of artificial agents, mechanics, hardware and software. All these factors create the conceptual realization by which we create our experiences and representations.

The thought of embodiment has its origin in the phenomenology worked by Merleau-Ponty (1962), who was moreover one of the forerunners of Gibson's notion of affordance (1979), placing body involvement as crucial to the way we perceive and act with the environment. Also being the biological structure that supports the vital cognition for the cognitive phenomenon, we must think about the implications or possibilities of this phenomenon being duplicated by electronic components, and which concepts and abstractions such formation could generate, given the importance of the embodied experience in the creation of abstract concepts (LAKOFF; JOHNSON, 1999). Thus, the role of language, metaphors, and mental representations in the formulation of concepts used in scientific theories is evident, despite all ontological commitment to a certain scientific realism. The term “naturalized epistemology”, forged by W.V. Quine in his 1969 seminal essay “Epistemology Naturalized”, followed several of the epistemic premises of Hume's skepticism, which, as we pointed out above, solves every platonically inspired foundation, including the dualism of Cartesian rationalism, in its pretension to justify an sure knowledge of the truth of the outside world. According to Quine (1969, p. 75):

> It was sad for epistemologists, Hume and others, to have to agree on the impossibility of strictly deriving the science of the external world from sensory evidence. Two fundamental principles of empiricism remained unassailable, however, and remain so today. One is that any evidence that exists for science is sensory evidence. The other is that any inculcation of word meanings must ultimately rest on sensory evidence (QUINE, 1969, p. 75).

As in Quine, the Humean-inspired empiricism that interests us, from Dreyfus, Rorty, Prinz and neopragmatism, is intersubjective, falsificationist and, interestingly, externalist, that is, a form of social linguistic and historically co-constitutive pragmatism of observer subject and the objective world to be known, experienced, lived. The problem of knowledge, as well as that of giving reasons for moral action, remains the great human problem according to the Humean formulation: in the words of Quine (1969, p. 72), "*the Humean problem is the human predicament*" so that not even induction (such as that which has been adopted by models of reflexive balance in metaethics and philosophy of science) can solve the naturalistic fallacies that arise from the guillotine. The externalism of the naturalists, in the wake of Hume and Quine, would here oppose the internalism of the rationalists and Kant, according to which the epistemic justification for cognition and moral action is found in consciousness (cogito) or a structure of transcendental subjectivity.

Although we cannot develop here the internalist-externalist problem, we believe that the debate between rationalism and empiricism that preceded it authorizes us to assert, as Quine suggested, that Hume's great mistake would have been to reduce analytical judgments to a priori, universal, necessary judgments, as opposed to synthetic ones. In tur, they are reducible to posterior judgments, contingent particularities, without solving the problem of induction but allowing, on the contrary, their return through the back door, as Popper would show, by the self-deception of those who intend to justify the moral action with a transcendental or normativism argument. Our programmatic intuition on AI ethics is, therefore, that neither naturalism seems to be able to reduce the alignment to a utilitarian program, nor the deontological, normative models and their transcendental arguments seem satisfactory to avoid anthropomorphic suspicion.

Computational neuroethology is a distinct area of neuroscience, as it involves the creation of joint models of neural circuits, biomechanics, and ecological niches as relevant parts of a cognitive agent (CHIEL; BEER, 1997). Work in the field of autonomous robotics emphasizes that intelligent behavior is an emerging property of an agent incorporated in an environment with which it must interact continuously. Thus, the symbolic computer vision, which places the brain as the source of commands that are issued to the body, may be incomplete. There may be a cognition or “mind” of the body (or mechanical system), governed by the laws of physics itself. This puts the nervous system not in a position to issue commands, but suggestions, reconciled with the biomechanical and ecological context (RAIBERT; HODGINS, 1993). There is the possibility that an AI that has an understanding of human concepts would require a design very close to that of a human being (SOTALA; YAMPOLSKIY, 2013).

Finally, to understand the SED approach we must analyze the assumed dynamics. We refer to dynamics as a mathematical theory that describes systems that systematically change over time. The dynamic framework also provides us with a different filter to observe the phenomenon in question (FRANKISH; RAMSEY, 2014). Dynamic systems are certainly configured as a body of mathematics, and not as a scientific theory of the natural world. The most common examples of dynamic systems are sets of partial differential equations, used to describe phenomena such as the movement of water, behavior of electromagnetic fields, the position of subatomic particles among other natural phenomena. Thus, the dynamic perspective brings with it a set of concepts and filters that influence the way we think about the phenomenon studied; when approaching any system from the dynamic perspective, we try to identify a set of state variables whose evolution can explain the observed behavior, the dynamic laws by which the values of these variables evolve in time, the dimensional structure of their evolution, possible states and dominant parameters (BEER, 2000).

Finally, the hypothesis of the situated embodied and dynamic structure postulates that brains, bodies, and environments are dynamic systems, governed by dynamic laws, and the dynamics of this triad are coupled, is the study of the behavior of the complete dynamic system, brain-body-environment, the correct object of study for cognition (BEER, 2000). The most crucial conclusion to be drawn from this model is: the behavior is a property of the whole brain-body-environment system and cannot, therefore, be adequately attributed to any subsystem isolated from the others. We propose that such an approach, SED, can be an interesting model for embodied agents and safe AI systems.

VII. Discussion and Conclusion

How can this dynamic approach be useful for the problem of learning values? This has been the guiding question of this study. We have seen in this study the imminent advance of AI technologies, and the importance that such advances are made safely, because we cannot anthropomorphize AI, and expect artificial intelligent agents to have the same terminal objectives (values) as us. Therefore, value learning becomes an area of crucial importance in the field. The limitations present in the representative symbolic method and the connectionist model may be indicating to us that a different approach to the behavioral problems of intelligent agents should be considered. Dynamism certainly approaches the problem differently and unveils new aspects that both the symbolic and the connectionist model leave aside.

How should we understand the nature and role of this inner state within a dynamic agent? The traditional computational interpretation of such states would be as internal representations. Unfortunately, despite the fundamental role that the notion of representation plays in computational approaches, there is very little agreement about what its real function is in controlling and maintaining behavior. We should also remember that symbolism, connectionism, and dynamism are theoretical structures, not scientific theories of the natural world, that is, they cannot be proved or refuted. While symbolism emphasizes the manipulation of internal representations, Connectionism emphasizes the architecture of the network and the training protocol. The SED structure, on the other hand, highlights the trajectory space and the determining influences on the brain-body-environment system. A dynamic approach to the problem of value learning may help us to elucidate some of the problems in value learning. However, as stated above, we do not put ourselves in a position of anti-representationalism. On the contrary, a complete theory of cognition is likely to use all three theoretical structures. We suggest that in certain cases, as in goal-oriented behavior, the internal functioning of a dynamic agent cannot be interpreted as representative unless we refine what a representation really can be or mean.

Gärdenfors (2000) proposes a general theory of representation, where concepts such as values are represented as geometric forms within a multidimensional space. Several brain modeling studies try to understand how the brain creates and manipulates information (KRIEGESKORTE; KIEVIT, 2013), and recent findings using simulations of cortical groups analyzed by algebraic topology show that the brain seems to organize itself in an orderly and geometric way when we analyze its structure as a multidimensional object (REIMANN et al.

2017). It is possible that similar structures, corresponding to the concept of value, are found in the hyperdimensional field that composes the cognitive agent.

The dynamic approach differs from the symbolic and connectionist cognitive models because it places biomechanics and ecology with the same relevance as neural activity in the phenomenon of cognition. Perhaps the difficulties we have encountered in learning values and other problems in the field of AI are due to the fact that we are ignoring two crucial factors of the phenomenon. The implications of the dynamic hypothesis not only bring a new way of thinking but also new problems to the field of AI research, thus nurturing new ideas in areas such as neurophilosophy, neuroscience, metaethics, computational neuroethology, and the interdisciplinary field of cognitive science itself. In conclusion, improvement and a better understanding of dynamic systems concepts is needed, with the promise that such methods can be useful for the problem of value alignment in AI and for the cognitive science community in general.

Acknowledgments
I would like to thank the Academic Excellence Program (PROEX) of CAPES Foundation (Coordination for the Improvement of Higher Education Personnel), and the Graduate Program in Philosophy of the Pontifical Catholic University of Rio Grande do Sul, Brazil.

Funding
Research supported by the Academic Excellence Program (PROEX) of CAPES Foundation (Coordination for the Improvement of Higher Education Personnel), Brazil.

References

ABADI, M. et al. Deep Learning with Differential Privacy. In: arxiv.org/abs/1607.00133, 2016.

ALLISON, H. E. *Idealism and Freedom: Essays on Kant's Theoretical and Practical Philosophy*. Cambridge University Press, 1996.

AMODEI, D. OLAH, C. STEINHARDT, J. CHRISTIANO, P. SCHULMAN, J. MANÉ, D. Concrete problems in AI safety. arXiv preprint arXiv:1606.06565, 2016.

ASIMOV, I. *I, Robot*. New York: Doubleday, 1950.

BARRETT, A. M. BAUM, S. D. A Model of Pathways to Artificial Superintelligence Catastrophe for Risk and Decision Analysis. *Journal of Experimental & Theoretical Artificial Intelligence*, JETAI (29), nº. 2, pp. 397-414, 2017. Disponível em: https://arxiv.org/pdf/1607.07730

BEER, R. D. Computational and dynamical languages for autonomous agents. In: *It's about time: An overview of the dynamical approach to cognition. Mind as motion: Explorations in the dynamics of cognition*. Cambridge, Mass.: MIT Press, 1998, p. 121-147.

BEER, R. D. Dynamical approaches to cognitive science. *Trends in Cognitive Sciences.* 4ª, p. 91-99, 2000.

BEER, R. D. The dynamics of active categorical perception in an evolved model agent (with commentary and response). *Adaptive Behavior*. 11ª ed, p. 209-243, 2003.

BOSTROM, N. The Superintelligent Will: Motivation and Instrumental Rationality in Advance Artificial Agents. *Minds and Machines*, 22ª, p. 71-85, 2012.

BOSTROM, N. *Superintelligence*, Chapter 12. Oxford University Press, 2014.

BOSTROM, N. ĆIRKOVIĆ, M. Introduction. In: *Global Catastrophic Risks*, edited by Bostrom, N. Ćirković, M. New York: Oxford University Press, pp. 1-30, 2008.

BRYNJOLFSSON, E. MCAFEE, A. *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*. New York: W.W. Norton; Company, 2014.

CHIEL, H. J. BEER, R. D. The brain has a body: adaptive behavior emerges from interactions of nervous system, body and environment. *Trends in neurosciences,* Nº 20 V.(12), pp. 553-557, 1997.

CHURCHLAND, P. S. SEJNOWSKI, T. *The computational brain*. Cambridge, MA, MIT Press. 1992.

CHURCHILL, R. R. ULFSTEIN, G. Autonomous Institutional Arrangements in Multilateral Environmental Agreements: A Little-Noticed Phenomenon in International Law. *American Journal of International Law*, 94ª ed, V (4), p. 623-659, 2000. dx.doi.org/10.2307/2589775

CLARK, A & CHALMERS, D. J. The extended mind. *Analysis* 58ª ed, p. 7-19. 1998.

DEWEY, D. Learning what to value. *Artificial General Intelligence*, p. 309-314, 2011.

DIJKSTRA, E. W. The threats to computing science. Paper presented at the ACM 1984 South Central Regional Conference, Austin, TX, Nov. 16-18, 1984.

DOCHERTY, B. L. *Losing Humanity: The Case Against Killer Robots*. New York: Human Rights Watch, 2012.

DREYFUS, H. L. *What Computers Still Can't Do: A critique of Artificial Reason*. Cambridge: MIT Press, 1992.

DREYFUS, H. L. DREYFUS, S. E. *What artificial experts can and cannot do*. AI Society, 6ª ed, V (1),p. 18-26, 1992.

DREYFUS, H. L. Why Heideggerian Artificial Intelligence failed and how fixing it would require making it more Heideggerian. *Philosophical Psychology*. 20ª ed, V (2), 2007.

ELIASMITH, C. The third contender: A critical examination of the dynamicist theory of cognition. *Philosophical Psychology*, 9ª ed, V (4), p. 441-463, 1996.

FRANKISH, K. RAMSEY, W. N. *The Cambridge handbook of artificial intelligence*. Cambridge University Press. 2014.

FREY, C. OSBORNE, M. *The Future of Employment: How Susceptible Are Jobs to Computerisation?* Technical Report, Oxford Martin School, University of Oxford, Oxford, UK, 2013.

GÄRDENFORS, P. *Conceptual Spaces: The Geometry of Thought*. MIT Press. 2000.

GIBSON, J. J. *The Ecological Approach to Visual Perception*. Houghton Mifflin. 1979.

GRACE, K. SALVATIER, J. DAFOE, A. ZHANG, B. EVANS, O. *When will AI exceed human performance? Evidence from AI experts*. arXiv preprint arXiv:1705.08807. 2017.

HADFIELD-MENELL, D. RUSSELL, S. J. ABBEEL, P. DRAGAN, A. *Cooperative inverse reinforcement learning*. Advances in neural information processing systems, p. 3909-3917, 2016.

HEIDEGGER, M. *Ser e Tempo*. Tradução e edição bilíngue com notas de Fausto Castilho. Campinas e Petrópolis: Editora Unicamp e Vozes, 2012. (1927)

HIBBARD, B. *The error in my 2001 VisFiles column*. 2012.

HUME, D. *Tratado da Natureza Humana*. Tradução de Débora Danowiski. 2ª ed. São Paulo: Editora da UNESP, 2009. (1739)

JAYNES, E. T. *Probability theory: The logic of science*. Ed. G. Larry Bretthorst. New York: Cambridge University Press. 2003. doi:10.2277/0521592712.

KRIEGESKORTE, N. KIEVIT, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, nº 17, V.(8), pp. 401-12, 2013.

KUZNETSOV, Y. A. *Elements of Applied Bifurcation Theory.* 3ª ed. Springer. 2004.

LAKOFF, G. JOHNSON, M. *Philosophy in the Flesh*. Basic Books. 1999.

LEGG, S. Is there an elegant universal theory of prediction? In: *Algorithmic learning theory: 17th international conference*, ALT 2006, Barcelona, Spain, October 7–10, 2006. Proceedings, ed. José L. Balcázar, Philip M. Long, and Frank Stephan. Lecture Notes in Computer Science 4264. Berlin: Springer. 2006. doi:10.1007/11894841_23.

LEIKE, J. MARTIC, M. KRAKOVNA, V. ORTEGA, P. EVERITT, T. LEFRANCQ, A. ORSEAU, L. *AI Safety Gridworlds*. arXiv preprint arXiv:1711.09883. 2017.

LLOYD, S. Computational capacity of the universe. *Physical Review Letters* 88 (23): 237901, 2002. doi:10. 1103/PhysRevLett.88.237901.

MERLEAU-PONTY, M. *Phenomenology of Perception*. New York: Humanities Press, 1962. (1945)

MORAVEC, H. P. When will computer hardware match the human brain? *Journal of Evolution and Technology* 1998. Disponível em: http://www.transhumanist.com/volume1/moravec.htm.

MORDVINTSEV, A. OLAH, C. TYKA, M. *Inceptionism: Going deeper into neural networks.* In: Google Research Blog. Retrieved June 20, 2015.

MÜLLER, V. C. BOSTROM, N. Future progress in artificial intelligence: A survey of expert opinion. *Fundamental issues of artificial intelligence*, p. 555-572, 2016.

NEWELL, A. *Unified theories of cognition*. Cambridge, MA, Harvard University Press. 1990.

NG, A. Y. RUSSELL, S. J. *Algorithms for inverse reinforcement learning*. In Pat Langley, editor, 17th International Conference on Machine Learning (ICML-'00), p. 663-670, 2000.

OMOHUNDRO, S. M. The Nature of Self-Improving Artificial Intelligence. Disponível em: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.137.1199&rep=rep1&type=pdf. Acessado em 11 de agosto de 2019.

QUINE, W.V.O. Epistemology Naturalized. In *Ontological Relativity and Other Essays*. New York: Columbia University Press, 1969, pp. 69–90.

RAIBERT, M. H. HODGINS, J.K. *Biological Neural Networks*. Invertebrate Neuroethology and Robotics. Academic Press, p. 319–354, 1993.

REIMANN, M. W. NOLTE, M. SCOLAMIERO, M. TURNER, K. PERIN, R. CHINDEMI, G. DŁOTKO, P. LEVI, R. HESS, K. MARKRAM, H. Cliques of Neurons Bound into Cavities Provide a Missing Link between Structure and Function. *Frontiers in Computational Neuroscience*. Nº 11, V (48), 2017. doi: 10.3389/fncom.2017.0004

ROCKWELL, T. Extended cognition and intrinsic properties. *Philosophical Psychology*, 23ª ed, p.741-757, 2010.

ROCKWELL, T. *Neither Brain nor Ghost: A Non-Dualist Alternative to the Mind-Brain Identity Theory*. Bradford Books, MIT press, 2005.

SHULMAN, C. *Omohundro's Basic AI Drives and Catastrophic Risks*. The Singularity Institute, 2010. Disponível em: https://intelligence.org/files/BasicAIDrives.pdf

SMOLENSKY, P. On the proper treatment of connectionism. *Behavioral and Brain Sciences*. 11ª ed, V(1), p. 1-23, 1988.

SOARES, N. FALLENSTEIN, B. *Aligning Superintelligence with Human Interests: A Technical Research Agenda*. Technical Report, Machine Intelligence Research Institute, Berkeley, CA. 2014.

SOARES, N. FALLENSTEIN, B. YUDKOWSKY, E. ARMSTRONG, S. *Corrigibility*. Artificial Intelligence and Ethics, ed. T. Walsh, AAAI Technical Report WS-15-02. Palo Alto, CA: AAAI Press, 2015.

SOARES, N. Value Learning Problem. In: *Ethics for Artificial Intelligence Workshop at 25th International Joint Conference on Artificial Intelligence* (IJCAI-2016) New York, NY, USA 9-15 July, 2016.

SOTALA, K. From mostly harmless to civilization-threatening: Pathways to dangerous artificial intelligences. In *ECAP10: VIII European Conference on Computing and Philosophy*, edited by Klaus Mainzer. Munich, 2010.

SOTALA, K. *Disjunctive scenarios of catastrophic AI risk*. Artificial Intelligence Safety and Security, pp. 315-337, 2018.

SOTALA, K. YAMPOLSKIY, R. V. *Responses to Catastrophic AGI Risk: A Survey*. *Technical report*. Berkeley, CA: Machine Intelligence Research Institute. 2013.

THAGARD, P. *Conceptual revolutions*. Princeton, Princeton University Press. 1992.

TVERSKY, A. KAHNEMAN, D. The framing of decisions and the psychology of choice. *Science*, Nº. 211 V.(4481), pp. 453-458, 1981.

VAN GELDER, T. The dynamical hypothesis is cognitive science. *Behavioral and Brain Sciences*. 21ª ed, p. 615- 628, 1998.

VAN GELDER, T. PORT, R. *It's about time: An overview of the dynamical approach to cognition. Mind as motion: Explorations in the dynamics of cognition*. Cambridge, MA, MIT. 1998.

VON NEUMANN, J. MORGENSTERN, O. *Theory of games and economic behavior*. Princeton, NJ: Princeton University Press. 1953.

YUDKOWSKY, E. Artificial Intelligence as a Positive and Negative Factor in Global Risk. In: *Global Catastrophic Risks*, edited by N. Bostrom and M. M. Ćirković, 308-45. New York: Oxford University Press, 2008. https://intelligence.org/files/AIPosNegFactor.pdf

About the authors

Corrêa, N. K is a Ph.D. student in philosophy, and a Masterin electrical engineering, currently working in the area of machine ethics at PUCRS University in Brazil. He has an academic interest in the area of machine ethics and the problem of moral alignment between AI and humans.

De Oliveira, N.F. Ph.D. in Philosophy from the State University of New York at Stony Brook. He is currently a full professor at the Pontifical Catholic University of Rio Grande do Sul (PUCRS), Coordinator of Philosophy at CAPES (2018-22), Coordinator of the Neurophilosophy Research Group (Brain Institute, InsCer), Editor of Veritas magazine, member of the Clinical Bioethics Committee and member of the coordinating committee of the Brazilian Center for Research in Democracy, created in 2009.