Journal of Politics and Ethics in New Technologies and AI

Vol 5, No 1 (2026)

Journal of Politics and Ethics in New Technologies and AI

Formal Assessment of the Moral Worth of Artificial Intelligence

Flavio Soares Correa da Silva

doi: 10.12681/jpentai.43702

Journal of Politics and Ethics in New Technologies and AI
Volume 5, Issue 1 (2026)
e-ISSN: 2944-9243

https://doi.org/10.12681/jpentai.43702

RESEARCH ARTICLE

Formal Assessment of the Moral Worth of Artificial Intelligence

Flavio S. Correa da Silva
Institute of Advanced Studies, University of Sao Paulo – Cidade Universitaria ASO, Sao Paulo, Brazil.

Abstract
In recent years, the name *Artificial Intelligence (AI)* has admitted interpretations beyond the borders of Science and Technology, reaching out to the realms of commercial and social phenomena. Ethical issues have resulted from this polysemy, leading to the development of policies and methodologies to define, assess and safeguard ethical standards in the development and adoption of AI products. In the present article, we unfold the different interpretations of AI and introduce a formal framework based on which computational tools can be built to support the definition and utilisation of ethical standards. The framework is grounded on *virtue ethics* and the *moral worth appraisal* of AI tools, grounding the formulation of declarative logic programs which can be computed to indicate the moral worth of the development and adoption of AI products.

Keywords: History of AI, Moral Worth, Responsible AI, Virtue Ethics

1. Introduction

Artificial Intelligence (AI) in 2025 is turning 70 years old as a scientific endeavour. In recent years, AI has moved beyond the realms of Science and Technology to attract the interest of corporate businesses as a strategic element for competitive advantage, triggering a race for market leadership and leveraged revenues, often at the cost of spreading unrealistic expectations about AI technologies (Barrow, 2024; Floridi, 2024; LaGrandeur, 2024; Markelius et al., 2024; Placani, 2024; Widder & Hicks, 2024). This scenario has raised ethical concerns about business practices and, as a response, the preparation of policies and methodologies for the responsible development and adoption of AI (Cath et al., 2018; Gambelin, 2018; McLarney et al., 2021; Shahriari & Shahriari, 2017; UNESCO, 2022). Bringing these policies and methodologies to action has proven to be challenging, considering the requirements of well-stated action plans and clear business advantages for corporations.

In this article, our goal is to complement existing policies and methodologies with a formal framework to address ethical concerns, aligned with characterisations of ethics, virtue, and the moral worth of actions proposed by Hellenistic philosophers following the tenets of Stoicism, which stands out among

Correa da Silva, F. S. (2026). Formal Assessment of the Moral Worth of Artificial Intelligence. *Journal of Politics and Ethics in New Technologies and AI, 5*(1), e43702. https://doi.org/10.12681/jpentai.43702

Western philosophical schools for focusing on the practice of philosophy and the importance of clear and logical reasoning.

A commitment to the development and adoption of morally praiseworthy products is clearly advantageous to society, but it may not be embraced by corporate business without a clear offer of competitive advantages. For this reason, we also argue that morally praiseworthy product development and adoption can contribute to business longevity, heretofore considered as a competitive business advantage. In Section 2, we review the perspectives on ethics, virtue, and the moral worth of actions adopted in this work. In Section 3, we review the field of AI, with a particular focus on recent developments and the business of AI. In Section 4, we present a formal framework to help assess the moral worth of decisions related to the development and adoption of AI products. Finally, in Section 5, we present some discussion and conclusions.

2. Ethics and the Moral Worth of Actions

Ethics refers to *"a system of accepted beliefs that control behaviour, especially such a system based on morals"*, while morality refers to *"the standards of good or bad behaviour, fairness, honesty, etc. that each person believes in, rather than to laws"*.[1] Combining these two concepts, we can interpret that ethics refers to beliefs that structure a code of conduct based on established standards of good or bad behaviour. These standards are named *morality*.

Ethical studies are commonly structured based on three complementary perspectives (Vallor & Rewark, 2018):

1. **Duty Ethics**, based on moral principles typically established in norms, regulations, and laws, indicating individual and collective conduct leading to social good. Documents establishing moral principles can be as old as the Babylonian *Code of Hammurabi* (c. 1800 BC) and the Egyptian *Laws of Maat* (c. 1200 BC) (O'Regan, 2024).
2. **Consequence Ethics**, explicitly articulated during the late British Enlightenment by Jeremy Bentham and John Stuart Mill, suggesting that actions should maximise pleasure and minimise pain. Consequence Ethics depends on the quantification of the consequences of actions, so that a balance can be calculated to judge the morality of actions.
3. **Virtue Ethics**, based on the concept of *good life* developed by ancient Greek philosophers such as Socrates, Plato, Aristotle, and the Stoics: a good life is a life devoted to Virtue, which is

[1] Cambridge Dictionary, Cambridge University Press, n.d., http://dictionary.cambridge.org/dictionary/english. Retrieved on 10 July 2025.

defined as participation in the natural order of things (Brennan, 2015; Jedan, 2009; Snow, 2017). Virtue is observable through a set of *cardinal* virtues, indicating rules of conduct that can be subject of ethical appraisal. Socrates, Plato, and the Stoics identified four cardinal virtues: *prudence*, *perseverance*, *temperance*, and *justice*. Aristotle further unfolded the cardinal virtues, adding refinements to the four original ones. This characterisation of good life and virtues has been echoed in other regions and times and can be identified in sacred writings of Jewish, Christian, Islamic, Buddhist, and Taoist traditions.

A common attribute of the three perspectives is that they all deal with the appraisal of the moral worth of actions (respectively, based on moral obligations, the balance between entailed pleasure and pain, and cardinal virtues). In the present article, we focus on Virtue Ethics (Vallor, 2016; Vallor, 2024), based on which we introduce a formal framework for the appraisal of the moral worth of actions, which are judged based on the extent to which they support virtuous behaviour. The appraisal of the moral worth of actions is based on *justifications* and *motivations* for acting (Markovits, 2010; Markovits, 2012; Sliwa, 2016): If an individual acts for justified reasons and based on the right motivations, the corresponding actions are considered *moral*, otherwise they are *immoral*.

A framework to assist in the appraisal of the moral worth of actions must provide the means to judge actions objectively and consistently, in order to be a practical guideline for actions. Stoicism is an approach to philosophy that started in Hellenistic Greece and stands out among other Greek philosophical schools by focusing on the practice of philosophy in everyday life and the importance of clear and logical reasoning. Given our interest in operational guidelines for the responsible development and adoption of AI, these attributes seem appropriate to build a formal framework to support the assessment of actions.

The proposed framework is built using a non-monotonic first-order logic known as the logic of *stratified normal clauses* (Apt et al., 1988; Fitting & Ben-Jacob, 1990; Kolaitis, 1991; Kunen, 1989), which is interesting from the perspectives of knowledge representation and computational deductive reasoning:

- From the point of view of knowledge representation, it characterises an expressive language, capable of expressing logical reasoning and representing complex domains using first-order predicates and functions. Non-monotonicity is conveyed by the non-classical negation-as-failure, through which the inclusion of new facts in a logical theory can retract previous theorems from it, thus providing increased expressiveness in comparison with classical monotonic logics.

- From the point of view of computational deductive reasoning, it relies on a sound and complete algorithmic deductive machinery. Stratification ensures that for each logical theory, there is a single semantic model.

The expressiveness of this logic is suitable to represent the knowledge required to assist the appraisal of the moral worth of actions related to the development and adoption of AI products. This logic, already adjusted for the computable framework described in this article, contains the following alphabet, given a *maximum cardinality* $N < \infty$:

- A finite set of *constants*:

$$C = \{c_1, c_2, ..., c_{nc}\}, nc \leq N.$$

- Given a *maximum arity* $K < \infty$, for each $k \in \{1, 2, ..., K\}$:

  - a finite set of k-ary function symbols:

$$\mathcal{F}^k = \{f_1^k, f_2^k, ..., f_{nfk}^k\}, nfk \leq N.$$

  - a finite set of k-ary predicate symbols:

$$\mathcal{P}^k = \{p_1^k, p_2^k, ..., p_{npk}^k\}, npk \leq N.$$

- A countable set of *variables*:

$$X = \{x_1, x_2, ...\}.$$

- *Connectives*: $\neg$, $\wedge$, $\leftarrow$.
- *Quantifiers*: $\forall$, $\exists$.

Using this alphabet, the following expressions are defined:

- *Terms t* are constants c, variables x or, inductively, function applications on terms $f^k(t_1,...,t_k)$.
- *Atoms* P^a are predicate applications on terms $p^k(t_1,...,t_k)$.
- *Literals L* are atoms P^a (positive literals) or negated atoms $\neg P^a$ (negative literals).
- *Free* and *bound* variables are defined as usual in logical theories.
- *Positive clauses* are positive literals in which all free variables are universally quantified:

$$\forall(p^k(t_1, ..., t_k)).$$

A positive clause is *stratified* if it does not have any free variable.

Query clauses are negative literals in which all free variables are existentially quantified:

$$\forall(\neg p^k(t_1, ..., t_k)) \equiv \neg\exists(p^k(t_1, ..., t_k)).$$

Normal clauses are universally quantified expressions in "rule form":

$$\forall(P^a \leftarrow P_1^a \wedge ... \wedge P_m^a \wedge \neg P_{m+1}^a \wedge ... \wedge \neg P_n^a),\ m, n \geq 0.$$

In a normal clause, the literal P^a is called the *conclusion* and the conjunction $(P_1^a \wedge ... \wedge P_m^a \wedge \neg P_{m+1}^a \wedge ... \wedge \neg P_n^a)$ is called the *premise*.

- *Stratified normal clauses* are normal clauses in which free variables occur only in positive literals inside the premise.
- *Normal theories* are finite collections of clauses in which all clauses are positive or normal.
- *Stratified normal theories* are normal theories in which all clauses are stratified.

A normal theory is expected to characterise a complete description of the relevant solutions of a specified problem. When a query clause is added to a normal theory, consistency is checked, and, in case the inclusion of the query generates an inconsistency, the opposite of the query is proven by contradiction.

Intuitively, stratification adds to normal theories full specification across negations, as universally quantified variables are allowed only in positive literals inside the premise. For this reason, this language is tailor-made for the specification of guidelines and recommendations in systems development and decision making in problem solving. As already pointed out, verification by contradiction in stratified normal theories is also convenient from a computational point of view, as deductive reasoning can be implemented for these theories. In fact, computational reasoning for these theories is the heart of the programming language *PROLOG*, which was once a prominent language to develop AI systems (Wielemaker, 2012).

A step-by-step description of the proposed framework expressed as a stratified normal theory is as follows. In this description, we employ self-descriptive names for the predicates and terms. We also adopt the *PROLOG* syntax notation and represent variables using names starting with capital letters, and predicates and other terms using names starting with small letters.

The theory is planned to be used to build *product justification documents*, which can be queried to verify whether a specification of actions related to the development and adoption of AI products is morally praiseworthy, in the sense that it is justified and motivated by the defense of Virtue, as expressed by cardinal virtues. In other words, Virtue should be the goal of actions related to product development and adoption.

Actions are performed by agents and related to specific products. Actions are also *justified* by their expected effects and *motivated* by a chain of concepts that leads to cardinal virtues. Justifications and

motivations are characterised using implication relations and can be distinguished according to the conclusions of clauses in which they occur. Justifications should also be motivated by Virtue; hence all implicative chains should point to Virtue. The verification of a specification of actions must therefore indicate that Virtue is the final goal of any action.

Operationally, this verification can be performed using a query clause of the form

$$\neg virtue(namedPr, namedAg).$$

The goal should be to count on a stratified normal theory of specifications of actions such that, together with this query, the theory would become inconsistent. The terms *namedPr* and *namedAg* indicate, respectively, a specific product and a specific agent whose moral appraisal is being verified. Intuitively, this query clause states that Virtue cannot be present given the product *namedPr* and the agent *namedAg*.

Observable virtuous behaviour should be motivated by Virtue. Hence, virtues such as the cardinal virtues ought to be motivated by Virtue, which can be expressed using the following clause:

$$\forall(virtue(Pr, Ag) \leftarrow cardinal(Pr, Ag, Crd)).$$

Intuitively, the combination of the query clause and this normal clause entail that there cannot be a cardinal virtue *namedCrd* that can be observed given *namedPr* and *namedAg*.

An operational rendition of how to exert cardinal virtues is a little more elaborate. The general principle to be considered is that such behaviour should be the effect of a conjunction of events, none of which being considered harmful. This can be formalised as follows:

- The only function explicitly employed in this formalisation is the *list constructor*, recursively defined to build the corresponding data structure usual in computer programs. We employ the standard *PROLOG* notation $[H|T]$ to indicate the first element H and the rest T of the list.
- Given that a cardinal virtue should be the effect of a conjunction of events, a ground list of event names should be the premise for any specified virtue to be observed.

Harms caused by specific events should be checked in two steps:

1. *Potential harm* should be explicitly identified as a set of pairs $\langle Events, Harms \rangle$, in which *Events* would indicate a conjunction of events and *Harms* would indicate a set of potential harms entailed by this conjunction of events.
2. In order for any specific *Harm* present in the set *Harms* to occur:
    - A conjunction of *causing events* must occur, and

- A conjunction of *preventive events* must *NOT* occur.

For a behaviour to be considered virtuous, it should be grounded on the expected conjunction of events, none of which being provably harmful. This characterisation can be captured using the following clause:

- $\forall(cardinal(Pr, Ag, namedCrd) \leftarrow all(Pr, Ag, namedCrd, namedCrdEvents) \wedge potHarms(Events, Harms) \wedge subset(Events, namedCrdEvents) \wedge \neg exists(Pr, Ag, namedCrd, Harms))$.

Intuitively, given a specified cardinal virtue *namedCrd* and a specified conjunction of events *namedCrdEvents* expressed as a list of event names, the conjunction of events is a prerequisite for the cardinal virtue, if there is no occurring harm belonging to a list *Harms* linked to a list *Events* which is a subset of *namedCrdEvents*.

Every event must be caused either by another event or by an action motivated by virtuous behaviour:

- $\forall(event(Pr, Ag, Crd, namedEvent) \leftarrow event(Pr, Ag, Crd, namedEvent'))$.
- $\forall(event(Pr, Ag, Crd, namedEvent) \leftarrow action(Pr, Ag, Crd, namedAction))$.
- $\forall(all(Pr, Ag, Crd, [namedEvent|Events]) \leftarrow event(Pr, Ag, Crd, namedEvent) \wedge all(Pr, Ag, Crd, Events))$.
- $\forall(all(Pr, Ag, Crd, []))$.

Occurrence of a particular $Harm \in Harms$ can be verified using the following clauses:

- $\forall(exists(Pr, Ag, Crd, [Harm|Harms]) \leftarrow actualHarm(Pr, Ag, Crd, Harm))$.
- $\forall(exists(Pr, Ag, Crd, [Harm|Harms]) \leftarrow \neg actualHarm(Pr, Ag, Crd, Harm) \wedge exists(Pr,Ag,Crd,Harms))$.
- $\forall(actualHarm(Pr, Ag, Crd, namedHarm) \leftarrow all(Pr, Ag, Crd, namedCrdEvents) \wedge \neg all(Pr, Ag, Crd, namedCrdEvents'))$.

This stratified normal theory can be complemented with specific named agents, products, cardinal virtues, harms, and events, as well as statements of potential harms and actions. The query clause can be used to verify whether Virtue is the motivation for actions, and a minimality verification with respect to the semantic model that supports the complemented theory can be used to ensure that no actions involving the agents under consideration and related to the products under consideration are performed for motivations other than the pursuit of Virtue.

In Section 4 a concrete example of the use of this tool is presented, illustrating the tool in action. Given that this article is focused on ethical requirements for AI, before the presentation of this example, we present an overview of the field of AI.

3. An Overview of AI

AI can be considered as a scientific endeavour, a specialisation in engineering, and a niche for business development. The term *"Artificial Intelligence"* was used for the first time to name a research field at the workshop held at Dartmouth College (USA) in 1956, which aimed at the validation of the conjecture that *"(...) every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it"* (McCarthy et al., 1955). It focused on *intelligent behaviour*, studied using incrementally refined models of intelligence and adopting biological intelligent behaviour as a reference.

As an important methodological consideration given this characterisation of AI, it should be observed that concepts such as *General Intelligence*, referring to models that could be considered fully equivalent to biological intelligence, or *Super Intelligence*, referring to models that could surpass biological intelligence, cannot be valid targets of AI as a scientific initiative, considering that both concepts would indicate results of experiments which have failed to model the intelligent behaviour that guided their design.

Intelligent behaviour is multifaceted, and models have rendered specific aspects of intelligence, such as natural language processing, image recognition, task planning, and deductive reasoning, resulting in fragmentation of AI. Computational tools to validate conjectures about intelligent behaviour have been reused in general problem solving, thus attracting the interest of fields as diverse as engineering, product design, and business development. Tools can be organised in four groups:

1. **Computational deductive reasoning**: declarative programming languages – most remarkably logic programming languages such as *PROLOG* (Apt, 1991; Apt et al., 1997); tools for automated reasoning in non-classical logics applied to knowledge representation and inference (Fagin et al., 2004; van Harmelen et al., 2008); heuristic methods for efficient solution of the boolean satisfiability problem (Een & Sorensson, 2003; Moskewicz et al., 2001); ontologies and applications in knowledge representation and problem solving (Fensel, 2001; Uschold & Gruninger, 1996).
2. **Natural Language Processing (NLP)**, following two approaches: (a) *Foundational*, based on computational reconstructions of linguistic phenomena to understand human language and generate useful tools for text translation, document summarisation, information extraction and question answering (Wilks, 2005); and (b) *Statistical*, working on observable self-referential co-occurrences of language patterns in discourse (Johri et al., 2021).

3. **Adaptive and heuristic optimisation**: simulation of collective intelligence and how behaviour can be locally and dynamically adjusted to optimise global performance, characterising the field of Evolutionary Computing (Eiben & Smith, 2015).
4. **Stepwise function approximations,** *aka* **Machine Learning (ML)**, which has received unprecedented attention given recent impressive empirical results, following two approaches: (a) *Foundational*, considering learning as a phenomenon, hence interacting with Cognitive Science (Perconti & Plebe, 2020), and generating computational learning models that are *(i)* sufficiently expressive to approximate function classes efficiently (Augustine, 2024; Gavranovic, 2024; Jia et al., 2024) and *(ii)* probabilistically reliable given reasonable assumptions about domains of interest and available samples (de la Vega et al., 2023; Levine et al., 2018; Valiant, 1984; Valiant, 1999); and (b) *Performance-oriented*, focusing on the design and use of benchmarks to highlight the features of engineered systems (Kopalidis et al., 2024; Prata et al., 2024; Rubachev et al., 2024; Wan et al., 2024).

For any tool to be trusted in mission critical, risk-prone, or costly scenarios, quality and reliability assessment must be rigorous, transparent, and generate detailed guidelines and requirements for appropriate use, covering data quality requirements and statistical assumptions about input data and corresponding expected outcomes. Ethical concerns can arise from inadequate use of tools, potentially due to purposefully biased information about their capabilities. Furthermore, AI tools have been used to stimulate expectations based on the foreseen consequences of their use, generate collective responses in advance of such expectations, and build revenues based on these responses (Bender et al., 2021; Gebru & Torres, 2024; Inie et al., 2024; Lemke & Monett, 2024; Monett & Grigorescu, 2024; Paullada et al., 2021; Vallor, 2024). Scrutable guardrails established as guidelines for responsible AI and managed by trusted institutions can be useful in preventing misinformation about AI products.

Agents to be considered in the appraisal of the moral worth of actions related to AI products can be *designers and developers*, *vendors*, and *users*. Designers and developers are related to the science and technology of AI, while vendors and users are related to the business of AI. Designers, developers, and vendors provide users with products, and users decide about their adoption.

4. Moral Worth and AI

As described in Section 2, the appraisal of the moral worth of actions can be aligned with Virtue Ethics, and this alignment can be formalised as stratified normal theories, thus enabling executable specifications of how actions related the development and adoption of products can be verified with respect to the pursuit of virtuous behaviour.

The pursuit of Virtue is clearly advantageous to society. Given the present tacit organisation of business interactions in use globally, tangible advantages must be offered to business corporations in order for any guidance for actions to be accepted. We argue that an indirect advantage of Virtue-driven actions can be institutional longevity by observing that most very long-standing businesses share the following organisational attributes:

- Perennial management aligned with well-defined and stable values, in many cases operationalised as ownership by a family or some other stable and cohesive social institution.
- Provision of services and products which are valued as socially beneficial across generations, such as hospitality, catering, healthcare, education, and culture.
- High regard to the importance of labour, and corresponding prioritisation of maintenance of workers' quality of life.

Other attributes are observed in very long-standing businesses (Ahn & Park, 2018; Tapies & Fernandez Moya, 2012). We highlight these ones because they can be easily aligned with Virtue-driven actions. As an illustration of how the presented framework could be used in practice, we consider the following specific values for variables in the theory presented in Section 2:

- **Agent**: the vendor of an AI product.
- **Product**: a chatbot designed to generate informational products of a particular type, e.g. how to design marketing campaigns for innovative consumer products.
- **Cardinal virtue**: *Justice*.
- **Harms**: (1) Hype about product capabilities, which can inflate expectations about what it can generate, and (2) Misuse of product, mistakenly assuming that it can design anything at all.
- **Actions**: (1) Distribute the product across appropriate channels to reach potential users, (2) Adopt fair pricing practices, (3) Adopt carelessly aggressive marketing, (4) Provide users with detailed product documentation and (5) Focus on corporate goals only and disregard the interests of the users.

The stratified normal theory can be instantiated using these values as follows:

- ∀(*cardinal*(*Pr, Ag,* **justice** ← *all*(*Pr, Ag,* **justice**, [**availability**, **fairPrice**])
  ∧ *potHarms*(*Events, Harms*)
  ∧ *subset*(*Events,* [**availability**, **fairPrice**]) ∧ ¬*exists*(*Pr, Ag,* **justice**, *Harms*)).
- ∀(*event*(*Pr, Ag, Crd,* **availability**) ← *action*(*Pr, Ag, Crd,* **distribution**)).
- ∀(*event*(*Pr, Ag, Crd,* **fairPrice**) ← *action*(*Pr, Ag, Crd,* **fairPricing**)).
- ∀(*event*(*Pr, Ag, Crd,* **productPush**) ← *action*(*Pr, Ag, Crd,* **aggrMrkng**)).

- ∀(*event*(*Pr, Ag, Crd,* **properInfo**) ← *action*(*Pr, Ag, Crd,* **documentation**)).
- ∀(*event*(*Pr, Ag, Crd,* **distortedGoals**) ← *action*(*Pr, Ag, Crd,* **selfCentricStr**)).
- ∀(*all*(*Pr, Ag, Crd,* [*Event*|*Events*]) ← *event*(*Pr, Ag, Crd, Event*) ∧ *all*(*Pr, Ag, Crd, Events*)).
- ∀(*all*(*Pr, Ag, Crd,* [])).
- ∀(*exists*(*Pr, Ag, Crd,* [*Harm*|*Harms*]) ← *actualHarm*(*Pr, Ag, Crd, Harm*)).
- ∀(*exists*(*Pr, Ag, Crd,* [*Harm*|*Harms*]) ← ¬*actualHarm*(*Pr, Ag, Crd,Harm*) ∧ *exists*(*Pr, Ag, Crd, Harms*)).
- ∀(*actualHarm*(*Pr, Ag, Crd,* **hype**) ← *all*(*Pr, Ag, Crd,* **productPush**) ∧ ¬*all*(*Pr, Ag, Crd,* **properInfo**)).
- ∀(*actualHarm*(*Pr, Ag, Crd,* **misuse**) ← *all*(*Pr, Ag, Crd,* [**productPush**, **distortedGoals**]) ∧ ¬*all*(*Pr, Ag, Crd, properInfo*)).

The theory can be complemented with statements of potential harms such as the following one:

- *potHarms*(**availability**, [**hype**,**misuse**]).

Finally, the theory can be complemented with statements of actions such as the following ones:

- *action*(**chatbot**, **vendor**, **justice**, **distribution**).
- *action*(**chatbot**, **vendor**, **justice**, **fairPricing**).
- *action*(**chatbot**, **vendor**, **justice**, **aggrMrkng**).
- *action*(**chatbot**, **vendor**, **justice**, **documentation**).
- *action*(**chatbot**, **vendor**, **justice**, **selfCentricStr**).

Figure 1. *PROLOG* source code for the *vendor* example.

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
virtue(Pr, Ag) :-
  cardinal(Pr, Ag, _).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
cardinal(Pr, Ag, justice) :-
  all(Pr, Ag, justice, [availability, fairPrice]),
  potHarms(Events, Harms),
  subset(Events, [availability, fairPrice]),
  \+ exists(Pr, Ag, justice, Harms).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
potHarms([availability], [hype, misuse]).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
all(P, Ag, V, [S|LS]) :-
  event(P, Ag, V, S),
  all(P, Ag, V, LS).
all(_, _, _, []).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
exists(P, Ag, V, [H|_]) :- actualHarm(P, Ag, V, H).
exists(P, Ag, V, [H|LH]) :- \+ actualHarm(P, Ag, V, H),
  exists(P, Ag, V, LH).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

```
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
actualHarm(P, Ag, V, hype) :-
  all(P, Ag, V, [productPush]),
  \+ all(P, Ag, V, [properInfo]).
actualHarm(P, Ag, V, misuse) :-
  all(P, Ag, V, [productPush, distedGoals]),
  \+ all(P, Ag, V, [properInfo]).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
event(P, Ag, V, availability) :- action(P, Ag, V, distribution).
event(P, Ag, V, fairPrice) :- action(P, Ag, V, fairPricing).
event(P, Ag, V, productPush) :- action(P, Ag, V, aggrMrkng).
event(P, Ag, V, properInfo) :- action(P, Ag, V, documentation).
event(P, Ag, V, distedGoals) :- action(P, Ag, V, selfCentricStr).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
action(P_ground, Ag_ground, V_ground, Ac_ground) :-
  lactionExists(P_ground, Ag_ground, V_ground, LAc_ground),
  member(Ac_ground, LAc_ground).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
lactionExists(chatbot, vendor, justice, [distribution, fairPricing,
      aggrMrkng, documentation, selfCentricStr]).
%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%%
```

This stratified normal theory is, evidently, a small illustration of what a realistic specification may look like. Nevertheless, it contains the structural elements that are required for appraisal of the moral worth of actions. This theory can be transcribed to a programming language such as *PROLOG*. The corresponding *PROLOG* source code is presented in Figure 1.

When triggered with the query

$$\neg \textit{virtue}(\textbf{chatbot}, \textbf{vendor}).$$

it will reply *YES* if actions are Virtue-driven, or *NO* otherwise (in the particular case of the theory presented in this example, it would reply *YES*). Upon verification whether actions can lead to "dead ends", i.e. events that do not support Virtue directly, the theory would also permit verification about "subreptitious" actions that would come together with Virtue-driven ones but could, in fact, be at service of other goals.

5. Conclusion

In the present work, we have adopted Virtue (as characterised by Stoicism) as measure of the moral worth of actions, and introduced a framework for the appraisal of the moral worth related to the design, development, distribution and adoption of AI products, which should be useful to support existing guidelines for the responsible use of AI.

Given recent developments in AI, particularly in Deep Learning and Generative Machine Learning (Ekundayo & Ezugwu, 2025), and considering the suggested use of the tool presented in Section 4, which can be adopted by designers, developers, vendors and users of AI products to verify and clarify their intentions when acting within their respective scopes of responsibility, it comes as a surprise that so many AI products are being distributed without a full clarification of their capabilities and limitations. Effort has been made by accomplished researchers to describe in detail the algorithms that fuel these products (Barrow, 2024; de la Vega et al., 2023; Gavranovic, 2024; Jia et al., 2024; Levine et al., 2018), but research that focuses on the extent to which recent technological accomplishments relate to AI is rare (Mollo & Milliere, 2023). Our hope is that this article can contribute to the development of actionable and verifiable guidelines for the proper development *and* utilisation of AI.

Acknowledgement

This work was partially supported by FAPESP – Project 2020/09850-0.

References

Ahn, S. Y., & Park, D. J. (2018). Corporate social responsibility and corporate longevity: The mediating role of social capital and moral legitimacy in Korea. *Journal of Business Ethics*, *150*(1), 117-134. https://doi.org/10.1007/s10551-016-3161-3

Apt, K. R. (1990). Logic Programming, Handbook of Theoretical Computer Science. *Van Leeuween (Manag. Ed) Vol*, *2*, 493-574.

Apt, K. R., Blair, H. A., & Walker, A. (1988). Towards a theory of declarative knowledge. In *Foundations of deductive databases and logic programming* (pp. 89-148). Morgan Kaufmann.

Apt, K. R. (1997). *From logic programming to Prolog* (Vol. 362). London: Prentice Hall.

Augustine, M. T. (2024). A survey on universal approximation theorems. *arXiv:2407.12895*. https://doi.org/10.48550/arXiv.2407.12895

Barrow, N. (2024). Anthropomorphism and AI hype. *AI and Ethics, 4*(3), 707-711. https://doi.org/10.1007/s43681-024-00454-1

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021, March). On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency* (pp. 610-623).

Brennan, T. (2015). The stoic theory of virtue. In *The Routledge companion to virtue ethics* (pp. 31-50). Routledge.

Cath, C., Wachter, S., Mittelstadt, B., Taddeo, M., & Floridi, L. (2018). Artificial intelligence and the 'good society': the US, EU, and UK approach. *Science and engineering ethics, 24*(2), 505-528. https://doi.org/10.1007/s11948-017-9901-7

De La Vega, N., Razin, N., & Cohen, N. (2023). What Makes Data Suitable for a Locally Connected Neural Network? A Necessary and Sufficient Condition Based on Quantum Entanglement. *arXiv:2303.11249*. https://doi.org/10.48550/arXiv.2303.11249

Eén, N., & Sörensson, N. (2003, May). *An extensible SAT-solver. In International conference on theory and applications of satisfiability testing* (pp. 502-518). Berlin, Heidelberg: Springer Berlin Heidelberg.

Eiben, A. E., & Smith, J. E. (2015). *Introduction to evolutionary computing*. springer.

Ekundayo, O. S., & Ezugwu, A. E. (2025). Deep learning: Historical overview from inception to actualization, models, applications and future trends. *Applied Soft Computing*, *181*, 113378. https://doi.org/10.1016/j.asoc.2025.113378

Fagin, R., Halpern, J. Y., Moses, Y., & Vardi, M. (2004). *Reasoning about knowledge*. MIT press.

Fensel, D. (2001). Ontologies. In *Ontologies: A silver bullet for knowledge management and electronic commerce* (pp. 11-18). Berlin, Heidelberg: Springer Berlin Heidelberg.

Fitting, M., & Ben-Jacob, M. (1990). Stratified, weak stratified, and three-valued semantics. *Fundamenta Informaticae*, *13*(1), 19-33. https://doi.org/10.3233/FI-1990-13104

Floridi, L. (2024). Why the AI hype is another tech bubble. *Philosophy & Technology*, *37*(4), 128. https://doi.org/10.1007/s13347-024-00817-w

Gambelin, O. (2024). *Responsible AI: Implement an Ethical Approach in Your Organization*. Kogan Page Publishers.

Gavranović, B. (2024). Fundamental Components of Deep Learning: A category-theoretic approach. *arXiv:2403.13001*. https://doi.org/10.48550/arXiv.2403.13001

Gebru, T., & Torres, É. P. (2024). The TESCREAL bundle: Eugenics and the promise of utopia through artificial general intelligence. *First Monday*, 29(4). https://doi.org/10.5210/fm.v29i4.13636

Inie, N., Druga, S., Zukerman, P., & Bender, E. M. (2024, June). From" AI" to Probabilistic Automation: How Does Anthropomorphization of Technical Systems Descriptions Influence Trust?. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2322-2347).

Jedan, C. (2009). *Stoic Virtues: Chrysippus and the Religious Character of Stoic Ethics*. A&C Black.

Jia, Y., Peng, G., Yang, Z., & Chen, T. (2024). Category-theoretical and topos-theoretical frameworks in machine learning: A survey. *arXiv:2408.14014*. https://doi.org/10.48550/arXiv.2408.14014

Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021, March). Natural language processing: History, evolution, application, and future work. In *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020* (pp. 365-375). Singapore: Springer Singapore.

Kolaitis, P. G. (1991). The expressive power of stratified logic programs. *Information and Computation*, *90*(1), 50-66. https://doi.org/10.1016/0890-5401(91)90059-B

Kopalidis, T., Solachidis, V., Vretos, N., & Daras, P. (2024). Advances in facial expression recognition: A survey of methods, benchmarks, models, and datasets. *Information, 15*(3), 135. https://doi.org/10.3390/info15030135

Kunen, K. (1989). Signed data dependencies in logic programs. *The Journal of Logic Programming*, *7*(3), 231-245. https://doi.org/10.1016/0743-1066(89)90022-8

LaGrandeur, K. (2024). The consequences of AI hype. *AI and Ethics*, *4*(3), 653-656. https://doi.org/10.1007/s43681-023-00352-y

Lemke, C., & Monett, D. (2024). AI-based service systems: digital-ethical issues and their impact on value co-creation. In *Handbook of Services and Artificial Intelligence* (pp. 228-249). Edward Elgar Publishing.

Levine, Y., Yakira, D., Cohen, N., & Shashua, A. (2018, February). Deep Learning and Quantum Entanglement: Fundamental Connections with Implications to Network Design. In *International Conference on Learning Representations*.

Markelius, A., Wright, C., Kuiper, J., Delille, N., & Kuo, Y. T. (2024). The mechanisms of AI hype and its planetary and social costs. *AI and Ethics*, *4*(3), 727-742. https://doi.org/10.1007/s43681-024-00461-2

Markovits, J. (2010). Acting for the right reasons. *Philosophical Review*, *119*(2), 201-242. https://doi.org/10.1215/00318108-2009-037

Markovits, J. (2012). Saints, heroes, sages, and villains. *Philosophical Studies*, *158*(2), 289-311. https://doi.org/10.1007/s11098-012-9883-x

McCarthy, J., Minsky, M. L., Rochester, N., & Shannon, C. E. (1955). A PROPOSAL FOR THE DARTMOUTH SUMMER RESEARCH PROJECT ON ARTIFICIAL INTELLIGENCE.

McLarney, E., Gawdiak, Y., Oza, N., Mattmann, C., Garcia, M., Maskey, M., ... & Little, C. (2021). *Nasa framework for the ethical use of Artificial Intelligence (AI)*. NTRS - NASA Technical Reports Server.

Mollo, D. C., & Millière, R. (2023). The vector grounding problem. *arXiv:2304.01481*. https://doi.org/10.48550/arXiv.2304.01481

Monett, D., & Grigorescu, B. (2024, October). Deconstructing the AI myth: Fallacies and harms of algorithmification. In *European Conference on e-Learning* (Vol. 23, pp. 242-248). Academic Conferences International Limited.

Moskewicz, M. W., Madigan, C. F., Zhao, Y., Zhang, L., & Malik, S. (2001, June). Chaff: Engineering an efficient SAT solver. In *Proceedings of the 38th annual Design Automation Conference* (pp. 530-535).

O'Regan, G. (2024). *Ethical and Legal Aspects of Computing: A Professional Perspective from Software Engineering*. Springer Nature.

Paullada, A., Raji, I. D., Bender, E. M., Denton, E., & Hanna, A. (2021). Data and its (dis) contents: A survey of dataset development and use in machine learning research. *Patterns*, *2*(11). https://doi.org/10.1016/j.patter.2021.100336

Perconti, P., & Plebe, A. (2020). Deep learning and cognitive science. *Cognition*, *203*, 104365. https://doi.org/10.1016/j.cognition.2020.104365

Placani, A. (2024). Anthropomorphism in AI: hype and fallacy. *AI and Ethics*, *4*(3), 691-698. https://doi.org/10.1007/s43681-024-00419-4

Prata, M., Masi, G., Berti, L., Arrigoni, V., Coletta, A., Cannistraci, I., ... & Bartolini, N. (2024). LOB-based deep learning models for stock price trend prediction: a benchmark study. *Artificial Intelligence Review*, *57*(5), 116. https://doi.org/10.1007/s10462-024-10715-4

Rubachev, I., Kartashev, N., Gorishniy, Y., & Babenko, A. (2024). Tabred: Analyzing pitfalls and filling the gaps in tabular deep learning benchmarks. *arXiv:2406.19380*. https://doi.org/10.48550/arXiv.2406.19380

Shahriari, K., & Shahriari, M. (2017, July). IEEE standard review - Ethically aligned design: A vision for prioritizing human wellbeing with artificial intelligence and autonomous systems. In *2017 IEEE Canada international humanitarian technology conference (IHTC)* (pp. 197-201). IEEE.

Sliwa, P. (2016). Moral worth and moral knowledge. *Philosophy and Phenomenological Research*, *93*(2), 393-418. https://doi.org/10.1111/phpr.12195

Snow, N. E. (Ed.). (2017). *The Oxford handbook of virtue*. Oxford University Press.

Tàpies, J., & Fernández Moya, M. (2012). Values and longevity in family business: evidence from a cross-cultural analysis. *Journal of Family Business Management*, *2*(2), 130-146. https://doi.org/10.1108/20436231211261871

UNESCO. (2022). *Recommendation on the ethics of artificial intelligence*. United Nations Educational, Scientific and Cultural Organization.

Uschold, M., & Gruninger, M. (1996). Ontologies: Principles, methods and applications. *The knowledge engineering review*, *11*(2), 93-136. https://doi.org/10.1017/S0269888900007797

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, *27*(11), 1134-1142. https://doi.org/10.1145/1968.1972

Valiant, L. G. (1999, May). Robust logics. In *Proceedings of the thirty-first annual ACM symposium on Theory of Computing* (pp. 642-651).

Vallor, S. (2016). *Technology and the virtues: A philosophical guide to a future worth wanting*. Oxford University Press.

Vallor, S. (2024). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press.

Vallor, S., & Rewark, W. J. (2018). An introduction to data ethics (*Course module). Santa Clara, CA: Markkula Center for Applied Ethics*.

Van Harmelen, F., Lifschitz, V., & Porter, B. (Eds.). (2008). *Handbook of knowledge representation* (Vol. 1). Elsevier.

Wan, Y., Bi, Z., He, Y., Zhang, J., Zhang, H., Sui, Y., ... & Yu, P. (2024). Deep learning for code intelligence: Survey, benchmark and toolkit. *ACM Computing Surveys*, *56*(12), 1-41. https://doi.org/10.1145/3664597

Widder, D. G., & Hicks, M. (2024). Watching the generative AI hype bubble deflate. *arXiv:2408.08778*. https://doi.org/10.48550/arXiv.2408.08778

Wielemaker, J., Schrijvers, T., Triska, M., & Lager, T. (2012). SWI-PROLOG. *Theory and Practice of Logic Programming*, *12*(1-2), 67-96. https://doi.org/10.1017/S1471068411000494

Wilks, Y. (2005). *The history of natural language processing and machine translation*. Encyclopedia of language and linguistics.