**arXiv manuscript No.**
(will be inserted by the editor)

# Making AI Meaningful Again

**Jobst Landgrebe · Barry Smith**

January 9th 2019



**Abstract** Artificial intelligence (AI) research enjoyed an initial period of enthusiasm in the 1970s and 80s. But this enthusiasm was tempered by a long interlude of frustration when genuinely useful AI applications failed to be forthcoming. Today, we are experiencing once again a period of enthusiasm, fired above all by the successes of the technology of deep neural networks or deep machine learning. In this paper we draw attention to what we take to be serious problems underlying current views of artificial intelligence encouraged by these successes, especially in the domain of language processing. We then show an alternative approach to language-centric AI, in which we identify a role for philosophy.

**Keywords** Artificial intelligence · deep neural networks · semantics · logic · Basic Formal Ontology (BFO)

## 1 The current paradigm of AI: Agnostic Deep Neural Networks (dNNs)

An AI application is a computer program that can create an output in response to externally derived input data in a way that is similar to the ways humans react to corresponding environmental stimuli. In what follows we will focus on AI applications that work with natural language, where the currently dominant paradigm is provided by what is called agnostic deep machine learning. The latter is a subfield of applied mathematics in which input-output-tuples of data are used to create stochastic models, in a process often (somewhat simplistically) referred to as training. The inputs are connected to outputs probabilistically, which means that there is a certain (a priori unknown but measurable) likelihood that a given input will be associated with a given output. The models are referred to as stochastic because they work by utilizing the fact that the data on which they draw is probabilistic in this sense. The models are, in addition, agnostic – which means

J. Landgrebe
Cognotekt GmbH, Bonner Str. 209, D-50996 Köln, Germany

B. Smith
University at Buffalo, Buffalo, NY, USA E-mail: phismith@buffalo.edu

that they do not rely on any prior knowledge about the task or about the types of situations in which the task is performed, and they are often "end to end," which means that they are meant to model an entire process such as answering a letter or driving a car. The models are, finally, deep in the sense that their architecture involves multiple layers of networks of computational units (thus not, for example, because of any depth in their semantics). For agnostic deep learning to be useable in creating an AI application, a number of conditions must be satisfied:

1. A sufficient body of training data must be available in the form of tuples of input and output data. These are digital representations of, respectively, a situation in response to which an action is required, and an action of the corresponding sort [16]. The classical AI-application in this sense is the spam filter, whose initial output data were created using annotations, in this case adding the label "spam" to email inputs.
2. Computers must be able to represent the training material they receive in digital form, so that it can be processed using the computing resources available today [7].
3. The annotated training tuples must be noise-poor – that is, similar inputs should lead to similar outputs. This is because machine learning requires repetitive patterns – patterns that have arisen in a recurring, rather than erratic, process. The behaviour of human email users when identifying spam forms a repetitive process of the needed sort. This is because users of email have a motive to become experts in successful identification of spam, since they are aware of the high costs of failure. The movement of the oil price over time, in contrast, is an example of an erratic process.
4. The data input must be abundant, since a machine-learning algorithm is a stochastic model that needs to represent the entire variance which characterises the situation in which the model is to be used. Because in language applications the overall complexity of the relationship between input and output is typically very high, the models will need many parameters. For mathematical reasons these parameters can only be computed (through the type of optimisation process otherwise called "training") on the basis of huge data sets. If the training sets are too small, there is a high chance that novel input data will not have the properties of the data sampled in the training distribution. The model will then not be able to produce an adequate output under real production conditions.

Most of the AI applications in current use, for example in product recommendation or advertisement placement, draw on machine learning approaches of this type. To establish the training set for the first spam filters, developers needed to collect millions of input-output data tuples, where inputs are emails received by humans and outputs are the classifications of these emails by their respective recipients either as spam or as valid email. They then train a machine-learning model using these data tuples and apply the result to new emails. The goal is that the model should replicate the human reaction it has been trained with, which means: identify spam in a way that matches the behaviour of a typical human. In applications such as this, it is only a very simple type of knowledge – knowledge that is captured by simple input-output-tuples – that is given to the machine by its mathematician or AI-engineer trainers. However, application developers may wish to improve the model that is generated by the algorithm from the data by

selecting for training purposes only those tuples that have certain desired properties (as when, in building training models for autonomous cars, they select driving behaviour of mature females rather than that of teenage males). The quality of the performance of the machine can on this basis even surpass that of the average human because the trainers of the model select only the most desired sorts of responses from what may be a much more considerable variance exhibited in the behaviour of humans. Thus they may select data that has been somehow validated by experts for correctness, creating what is called a "gold standard" set of annotations. Because the engineer uses prior knowledge about data quality when making such selections, this is equivalent to an – albeit minimalistic – usage of prior knowledge in machine learning. Machine learning with neural networks can out-perform even the strongest human performance, but only in three types of cases:

- where the behaviour that is modelled consists of truly repetitive processes with narrow scope and with data that can be easily represented adequately in digital form (for example spam filters, shopping recommendations) – this achieves an efficiency higher than is obtainable by humans
- in hypothesis-based pattern identification (for example in the recent identification by a dNN of a correlation between retinal patterns and cardiovascular risk factors [31] – this achieves an effectiveness higher than with humans
- in reinforcement learning, a method used in certain narrowly defined situations of the sort that arise in games (for example in GO [37] or First Person Shooters [18]) and in contexts that can be framed like games [39]– this achieves both efficiency and effectiveness higher than with humans.

Examples of usage are: (i) driving a car on a highway, (ii) scientific pattern-search applications, e.g. in biology or astronomy, (iii) robotics, e.g. industrial plant cleaning. But unfortunately, each of these types of situations is highly restrictive and context-specific, and none is available where we are dealing with natural language input.

## 2 Applying Agnostic Deep Neural Networks in the field of language understanding

To understand how modern agnostic deep neural network AI works in the language domain, consider the most prominent production example, which is that of machine translation as illustrated by Google translate[1]. A recent publication authored by Google Brain[2] and Google Research, with the title "Attention is all you need" [42], provides a representative example. The stochastic models described in this paper were trained for the translation of English to German and of English to French. To train Transformer– which is the best-performing "big" model described in the paper– the authors encoded the language material at their disposal using byte-pair encoding, which encodes each single-sentence input into an encoding vector of 1024

[1] https://translate.google.com/

[2] This is the official name of Googles AI department. While Googles machine-learning engineers are certainly among the leading representatives of their craft, the name nonetheless reveals a certain hubris.

real numbers (rounded to a certain number of decimal places). This is a complexity-reducing encoding, which means (very roughly) that it treats each sentence simply as a series of signs. This allows the encoding process to retain certain important features of the input sentences because relevant sentence patterns are repeated in many sentences in a similar way, and these sentences are shown to the algorithm[3].

But at the same time, it necessarily leads to the discarding of many subtleties of these sentences. This is because the embedding of the sentence loses relations not only between words within the sentence but also between sentences. A further feature of the experiments reported is that the models used are trained with quite small amounts of training data: 36 million sentence pairs for English-French, and only 4.5 million for English-German. The models are completely agnostic: they have no knowledge of linguistics, for example, because they have no knowledge of anything at all. Rather, they just try to mimic the human translations (or rather the corresponding sets of vectorially simplified input-output pairs) they learned from. The principal problem of this approach, however, is that embedding into 1024 encoding real numbers leads to the discarding of all information pertaining to the contexts of the input sentences. That this has adverse consequences becomes clear when we reflect that, in all language interpretation processes, even for single sentence inputs, humans use prior knowledge to contextualise the sentences they receive. As an example, consider how a typical reader of this text would contextualise the single sentence: "In the beginning was the word."

### 2.1 Results thus far

How well, then, do these models do? Transformer, specifically, creates a model that achieves a sentence-level score of 28.4 for English-German and 41.8 for English-French using the BLEU metric, which measures on a scale from 0 to 100 the degree of matching of the machine-translation with a human gold-standard translation [30]. A score of 100 can never be achieved, because there are always several valid translations for any given sentence and not all of them can be in the gold-standard set. But 75-85 could be achieved in theory. Such a score would be excellent, and it would correspond to the translation abilities of an average bilingual speaker. The scores achieved by Transformer, in contrast, which are reported as the state-of-the art in machine translation, are low. To illustrate the limitations of the approach, Hofstadter used input sentences with a high degree of cross-contextualisation[4].

[3] For example, the algorithm learns to translate the German word Mehl into flour because this pair is repeated many times in training sentences. But it will fail to translate "Wir haben Mehl Befehl gegeben zu laufen" into the adequate "We ordered Mehl to run". It rather gives out the nonsensical "We have ordered flour to run" (result produced on Jan. 7, 2019). The translation fails because there are not enough training examples to learn the martial usage of surnames without title.

[4] Douglas Hofstadter provides the following illustration of the lack of semantics in translate.google.com in "The Shallowness of Google Translate", *The Atlantic*, January 30, 2018, Text by Hofstadter: In their house, everything comes in pairs. Theres his car and her car, his towels and her towels, and his library and hers. Google Translate: Dans leur maison, tout vient en paires. Il y a sa voiture et sa voiture, ses serviettes et ses serviettes, sa bibliothque et les siennes. Translated back into English by Google: In their house everything comes in pairs. There is his car and his car, their napkins and their napkins, his library and theirs.

### 2.2 General limitations of machine learning

Major limitations of current deep learning paradigms have been identified already (for example in [24]). They include first of all a set of quite general problems affecting stochastic models of any sort– not only deep neural nets but also traditional regression and classification approaches [16], including graph-based stochastic models (Bayesian Networks).

The first of these limitations turns on the huge data need of stochastic models, which may employ millions of parameters. Transformer, for example, has 213 million parameters and needs at a minimum billions of data tuples to become useful even for the sorts of rough translation produced by google translate. This limitation is already of considerable importance given that, leaving aside the resources of internet giants such as Google, there are few real-world examples of data available in the sorts of quantities needed to deal with complex outcomes using any sort of stochastic approach. Second, all stochastic models require a stable environment. The quality of their output depends on how well they reflect the real-world input-output relationship they are aiming to represent. Where this relationship is erratic, there can be no good model (consider the oil price example above). But even where the relationship is stable, the model will quickly become invalid if the input-output relationship changes on either side even in some minor way. This is because the model does not generalise. Once fed with data as input that do not correspond to the distribution it was trained with, the model will fail without alerting the user that it is failing[5]. This explains why stochastic spam filters and similar applications are so vulnerable to changing situations, and why they so often need re-training. And the more complex the application, the more demanding will be the re-training of end-to-end neural networks that is required upon change of input constellations (for example when new types of sensors are introduced in driverless cars). The costs for such re-training will vary, of course, with the complexity of the input and the accuracy requirements of the network. But there is a third group of limitations, turning on the fact that the output of all stochastic models is, by definition, approximative. Models of this sort can yield only the most probable output for any given input and model, and this output often falls below even the average human output. For many imaginable useful purposes, however, the output should ideally be at least as reliable as the behaviour not of the average but of a qualified subset of human reference samples; this is very hard to achieve in language-focused applications using dNNs only. We can better understand the limitations of stochastic models when we reflect on how humans interpret reality. Unlike machines, humans are able spontaneously and immediately to attribute meaning to the world they experience. This is because the human species has evolved with a complex set of dispositions to react immediately in highly specific ways to specific sorts of external stimuli. Human beings are, along many dimensions, tuned to the environments in which they live. The entities that we experience are spontaneously assigned meanings that reflect their relevance to our survival, meanings that are assigned using machinery that has been hard wired into our brains. The belief that stochastic models can learn to make decisions without benefit of prior hardwiring of this sort is as naive as the old tabula rasa theories that were once the staple of empiricist philosophers and of

[5] Deterministic AI models do not generalize either, but they report their failures.

their empirical psychologist followers. Such views were criticized by J. J. Gibson in his ecological theory of perception [6][13], and they were experimentally refuted in the work on infant cognition of Carey [4], Gopnik [14], Keil [20], [21], Kim and Spelke [22], who demonstrated that infants (and primates, [32]) possess a large body of categorical and structural knowledge about the world of solid objects long before they even start acquiring the grammar of their mother tongue [23]. Indeed, it seems that language acquisition presupposes the working of a common set of ontological distinctions on the side of language learners, including the distinction between objects and processes, between individuals and categories, between natural and accidental properties of objects, and so forth. Even Bayesian models for concept learning based on similarity acknowledge (i) the need for a prior genus-individual distinction to explain the mechanics behind generalization and (ii) the existence of a prior meta-heuristic linking membership in a class to property instantiation [40, 41]. As Rehder formulates the matter, categorization relies on inferences about the causal role of putative essences in producing observable features [34]. The latter, in other words, are merely secondary, derivative; and all the naive knowledge brought to bear by the infant follows from the natural and universal supposition that things belong to classes sharing similar properties [26, 29]. Even children as young as 3 years old believe that the insides of objects are relevant in determining class membership [10, 12] and [20]. According to Carey and Xu [4] (p. 207) experiments on object recognition suggest that there is an object tracking system in the infant– a system that tracks three-dimensional, bounded, and coherent physical entities, and fails to track perceptually specified figures that have a history of non-cohesion. And what holds of infant cognition in general holds also of infant language leaning and language competence in particular, where the capability of object tracking grounds the use of nouns and pronouns. Indeed part of the background source of this empirical work on infant ontology was formed by Chomskys ideas on innate universal grammar [6]. Gelman and Byrnes [11] make explicit reference to these ideas when they assert that they are able to "determine how languages and conceptual systems are constrained by examining the forms and meanings that children construct, and which errors they fail to make" [11], compare [27], p. 47. For our purposes here, it is crucial that the AI applications running on todays computers can simulate at best only small fragments of the hard-wired human capabilities revealed in such research. This means that they can simulate only small fragments of the semantics underlying human language use. As we shall see, neural networks have in this respect even more severe limitations than traditional logic-based AI approaches to the modeling of human cognition. This is because the formal ontologies used in the latter involve direct representations of the sorts of objects, processes and attributes (and associated nouns, verbs and predicates) used by human beings in perceiving, acting and speaking. Neural networks attempt to build relation-rich content of this sort out of gigantically large numbers of features represented using numerical input vectors or matrices by estimating what amounts to a very large polynomial (this is what a neural network is) with the help of an optimization procedure. This seems to be infeasible even for simple ontologies of the RDF-sort made up of structures of the type: entity A– relates to– entity B [15].

[6] Indeed they were criticized, 200 years earlier, by Immanuel Kant in 1781 in his *Critique of Pure Reason*.

### 2.3 Limitations applying specifically to deep neural networks (dNNs)

As humans process sensory input data, they assign meanings to the objects and events which the data represent (and from which the sensory content originates), experiencing these objects and events as belonging to a specific sort of categorical structure. But dNNs do not use any of the target-derived properties of the input data that humans spontaneously use when they assign meaning to the data which they receive through experience. The result is a tremendous brittleness of dNN capabilities. Moosavi et al. [28] describe how high-performance neural networks developed for image classification can be perturbed to completely misclassify images when the input material is mixed with a perturbation image. For example, what is at first correctly classified by the system as a flagpole is classified as a labrador after the system is very slightly perturbed. Perturbations of an analogous sort do not cause problems for humans at all. Jo and Bengio [19] more recently showed that dNNs work merely by learning certain surface-statistical regularities from images: the green grass that forms the typical background of a cow, for example, is contrasted with the grey of asphalt that forms the typical background of a car. They can be perturbed so easily precisely because they do not learn what the images are about and the sort of world to which the imaged objects and events belong.

The same holds also of the dNNs constructed for language processing purposes. A recent paper by Chen et al. [5] proves what, given what was said above, we should in any case expect, namely that dNNs lack core computational features of traditional approaches to syntactic language analysis pioneered by Chomsky using probabilistic context-free grammars [6]. As the authors show, while it is required of every valid stochastic model that it compute a valid probabilistic distribution, this condition is not in general satisfied by dNNs working from language input. But without this ability, there can be no computational representation of semantics, and the paper by Feng et al. [9] shows, the language constituents used by dNNs to make predictions in question-answering or textual entailment tasks make no sense to humans at all in most cases[7]. This in turn means that dNNs, whatever it is that they are doing , cannot be modeling the semantics that need to be captured in order to extract information from texts, a crucial task in natural language processing for automation purposes.

Zheng et al. [43] provide a poignant example of the low quality currently being achieved for tasks of this sort, with a low net information extraction (IE) accuracy rate [8]. This reveals just how low the expectations in the field have become. The inability to compute natural language semantics is also illustrated by the recent

[7] One example described in [9] rests on the input: "In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments". The described system correctly answers the question: "What did Tesla spend Astors money on?" with a confidence of 0.78 (where 1 is the maximum). The problem is that it provides exactly the same answer with a similar degree of confidence as its response to the nonsensical question: "did?"

[8] The reported $F_1$-score of 0.52 seems quite high but most of the training material is synthetic and the reported outcome only concerns information triples, which cannot be used for applied IE. The results is poignant because the paper in question won the 2017 Prize for Information Extraction of the Association for Computational Linguistics, globally the most important meeting in the language AI field.

misclassification of the United States Declaration of Independence as hate speech by the Facebook filter algorithm[9].

dNNs are also unable to perform the sorts of inferences that are required for contextual sentence interpretation. The problem is exemplified by the following simple example:

"The cat caught the mouse because it was slow" vs.
"The cat caught the mouse because it was quick."

What is the "it" in each of these sentences? To resolve anaphora requires inference using world knowledge– about persistence of object identity, catching, speed, roles of predator and prey, and so forth. Thus far, however, little effort has been invested into discovering how one might engineer such prior knowledge into dNNs (if indeed this is possible at all). The result is that, for all current applications, dNN models are still very weak, as they can only learn from the extremely narrow correlations available in just that set of annotated training material on the basis of which they were created– with the exception of game-like situations in which training material can be generated synthetically, esp. in reinforcement learning.

And worse: because the dNNs rely exclusively on just those correlations, they are also unable to distinguish correlation from causation, as they can model only input-output-relationships in ways that are agnostic to questions of evidence and causality. They can detect, for example, that there is some sort of relationship between smoking and lung cancer. But they cannot determine the type of relation that is involved unless references to this relation and to relevant types of relata themselves form part of the annotated corpus. Unfortunately, to create the needed annotated gold standard corpora– one for each domain of interest– is hugely expensive in terms of both time and expertise. Thus to make dNNs work effectively in language applications not only are enormous collections of data required. For many applications at least– for example those involving the tracing of causality– the investment of considerable amounts of human expertise is needed also.

One final problem is that, in part because they do not incorporate prior knowledge, dNNs lack transparency– the models work as black boxes, so that their engineers cannot tell how the network worked to yield any given output. This poses a major challenge in areas where we need to reproduce the behaviour of the network, for example in case of disputes over liability. Taken together, these problems rule out entirely the use of machine learning to drive mission-critical AI systems– for example systems capable of driving cars or of managing nuclear power stations. They are too brittle and unstable against variations in the input, can easily be fooled, lack quality and precision, and fail completely for many types of language understanding or where issues of liability can arise. Even at their very best, they remain approximative, and so any success they achieve is still, in the end, based on luck rather than on modus ponens.

## 3 Making AI meaningful again

To overcome these problems, ways need to be found to incorporate prior knowledge into the AI algorithms. One attempt to do this is to enhance Bayesian Networks

[9] https://www.theguardian.com/world/2018/jul/05/facebook-declaration-of-independence-hate-speech

with an explicit relationship semantics [33], which allows the model designer to build in knowledge describing entity relationships before using data to train the weights of these relationships. This reduces the learning effort by providing a rudimentary form of prior knowledge. But unfortunately, the expressivity of the resulting models is too low to represent the sorts of complex contexts relevant to human language understanding. Furthermore, they are not exact, secure, or robust against minor perturbations. They are also not transparent, and thus they are not meaningful in the sense that humans cannot reliably understand how they work to achieve given results. The goal of meeting this requirement is now dubbed "explainable AI", and we believe that the most promising strategy for achieving this goal lies in building applications that work in accordance with the ways humans themselves assign meaning to the reality that surrounds them. To this end, a semantics-based representation is highly desirable that is able to deal with language as it is actually used by human beings. The representation should be able to incorporate prior knowledge based on low to medium amounts of input material of the sorts found in typical real-world situations. For humans do not find meaning in data. Rather, they find meaning in the objects and events that surround them, and in the affordances that these objects and events support. This entails a different sort of AI application, in the building of which not only mathematics and computer science play a role, but also philosophy. Part of what is needed here is to be found already in early attempts to create AI-systems under the heading of what is sometimes called strong logic-based AI. Already in the 1960s, the use of (first-order) logic for AI modeling purposes was regarded as attractive because it was seen as enabling exact inference.[10] The most interesting example of this strong AI for our purposes here is in the work of Patrick Hayes, a philosopher who first made his name with a paper co-authored with John McCarthy, commonly accredited with having founded the discipline of AI research. The paper is titled "Some Philosophical Problems from the Standpoint of Artificial Intelligence" and it lays forth for the first time the idea behind the calculus of situations [25]. In the subsequent years Hayes set forth the idea of what he called nave physics, by which he meant a theory, consisting of various modules called ontologies, that would capture the common-sense knowledge (sets of common-sense beliefs) which give humans (or robots) the capacity to reason and plan and navigate through the world [17]. The theory is axiomatized using first-order logic (FOL) and Hayes proposed that something of the order of 10,000 predicates would need to be encoded if the resulting theory was to have the power to simulate human reasoning about physical objects of the sorts that are encountered by humans in their everyday lives[11]. The problem with Hayes approach, as with strong (FOL-based) AI in general is that to mimic even simple human reasoning in real time would require a reasoning engine that is decidable, and this implies a severe restriction on the expressiveness of the logic that can be used. Standardly, one ends up with a very weak fragment of FOL such as that encapsulated nowadays in the s0-called Web Ontology Language (OWL, see below). OWL is restricted for example in that it can capture at most relational information involving two-place relations,

[10] An excellent summary can be found in [36].

[11] Hayes' conception of an ontology as the formalization of our knowledge of reality continues today in the work of Tom Gruber, whose Siri application, implemented by Apple in the iPhone, is built around a set of continuously evolving ontologies representing simple domains of reality such as restaurants, movies, and so forth.

**Table 1** Minimal desiderata for a real-world AI language-processing system

| Property | System | Example |
|---|---|---|
| Exactness | needs to be able to be exact where necessary and not always restricted to the merely approximative | in the insurance domain: automated validation and payment of a claim |
| Security | needs to avoid insecurities of the sort which arise, for example, when even slight perturbations lead to drastically erroneous outputs | in autonomous driving: avoid harmful consequences of adversarially manipulated traffic signs |
| Robustness | needs to be able to work reliably in a consistent way even given radical changes of situation and input, or to detect critical changes and report on its own inability to cope | language that is not understood by the system is sent for inspection by a human |
| Data parsimony | needs to be trainable with thousands to millions of data points (rather than billions to trillions– magnitudes which rarely occur in reality) | in the domain of business correspondence: automation of letter-answering on the basis of just a few thousand examples per class of letter |
| Semantic fidelity | needs to be able to incorporate contextual interpretations of input situations | in the domain of sentiment analytics: the Declaration of Independence should not be classified as hate speech |
| Inference | needs to be able to compute the consequences of given inputs, to distinguish correlation from causality (thus requiring the ability to reason with time and causation) | determination of required actions on the basis of text input, for example in automated processing of medical discharge summaries |
| Prior knowledge use | needs to be able to use prior knowledge to interpret situations | understanding that issuing a declaration of inability to pay implies earlier receipt of a payment request |

and it has a similarly diminished quantifier-syntax. For this and many other reasons, logic-based systems have never reached the point where they were able to drive AI-applications. They did however spawn the development of a huge body of mechanical theorem proving tools [35] and they contributed to the development of modern computational ontologies, which helped to transform biology into an information-driven discipline [2]. Both of these developments are, as we shall see, essential for the sort of logic-based AI that is being developed today.

### 3.1 How philosophy is being reinserted into AI

We will show in what follows how augmenting deep neural networks and other stochastic models with certain philosophically driven elements of logic-based AI is now allowing us to create AI applications that are already solving real-world problems. We present an example of how the sort of philosophy-driven ontology machinery proposed here can be made to work. To understand its functionality,

we start by giving details about the minimal requirements which we believe a real-world AI system must satisfy (listed in Table 1). These requirements cannot be satisfied by agnostic machine-learning systems alone, as they presuppose the ability to deal with the semantics of human (natural) language. They can be satisfied, we believe, only by using systems which combine components from dNNs with methods associated with traditional, logic-based AI in such a way as to allow incorporation of prior knowledge. What is the role of philosophy here? First, it was philosophers, including the mathematician-philosopher Gottlob Frege, who developed the methods which enable the expression in exact logical form of knowledge otherwise expressed in natural language. FOL itself was invented by Frege in 1879, and since then the FOL framework has been refined and extended in order to reach the stage where it is possible to represent natural language in a formal, computable manner[12]. Second, philosophers have from the very beginning attempted to understand how human language works, how language relates to the world, and how philosophers themselves use language in their own thinking. Think of Aristotles Organon and Book VII of his Metaphysics. In the 20th century, an entire branch of the discipline– called analytical philosophy– has grown up around this topic [8]. We shall see, too, that the discipline of formal ontology, too, has in recent years achieved considerable maturity in part as a result of the influence of philosophical ideas. The last four properties listed in Table 1 can only be achieved on the basis of a formal, computable representation of prior and world knowledge using a computable representation of the natural language semantics of the systems inputs and possibly also of its outputs. This computational representation needs two major elements: (a) a set of logical formalisms that can store and manipulate language in Turing-machines, and (b) a framework which enables one to define the meanings of the elements of the language, constituted by what are nowadays called formal ontologies. We start with the logical formalisms. Natural language is very hard to express in a logical framework, and for this a combination of different logical formalisms is needed to represent the input language (no matter whether this is used in training dNN models, in using these models in practice, or for knowledge insertion into algorithms). These logical formalisms must enable computation, for which not necessarily decidability, but robustness and completeness are needed.[13] Several such logical dialects are needed to cover different aspects of the input language (such as propositions, quantified entities and their relationships, temporal relationships and modalities). The requirement is that they are able to achieve, when used together, a good approximation of natural language semantics while also enabling computability and domain-specificity.[14] The syntactical representation of FOL was standardised in 2007 as Common Logic (CL) (ISO/IEC 24707:2007, now revised as ISO/IEC 24707:2018). Combined with a set of robust and complete propositional modal logics, CL enables a rich representation and manipulation of

[12] An overview is given in [3].

[13] Decidability is not required because with robustness and completeness logical inference is possible but may not terminate. In practice, algorithms are stopped after a maximum computation period defined by fiat, when the case is handed to a human for decision.

[14] Note that the originality of this approach compared to initiatives like, for example, CyC, stems from the restriction to small domains, the automated translation of natural language into logic and the usage of different logical dialects to represent different aspects of natural language.

linguistic content as well as automated machine inference (modulo the need for a time-out when specific computations take too long).

#### *3.1.1 Incorporating Ontologies*

Ontologies can be divided into two types. On the one hand are domain ontologies, which are formal representations of the kinds of entities constituting a given domain of inquiry together with the relations between such entities [38]. On the other hand are top-level ontologies, which represent the categories that are shared across a maximally broad range of domains– categories such as object, property, process and so forth. Each ontology is built around a simple a taxonomic hierarchy in which the types of entities are related to each other by the relation of greater and lesser generality (an analogue of the subset relation that holds between the instances of such types). Domain ontologies have enjoyed considerable success in the formalisation of the descriptive content of scientific theories above all in many areas of biology (see especially the Gene Ontology, [2], where they served initially as controlled, structured vocabularies for describing the many new types of entities discovered in the wake of the Human Genome Project. As more and more such ontologies came to be developed and applied to the annotation of more and more different types of data, the need arose to standardise these ontologies using a common top-level ontology framework, and it was in this context that there arose Basic Formal Ontology (BFO, [1]) which is used as shared top-level ontology in some 300 ontology initiatives (currently under development as an ISO standard under ISO/IEC: 21838-1 (Top-Level Ontologies: Requirements) and ISO/IEC: 21838-2 (BFO)). The use of a common top level allows multiple ontologies to facilitate standardised exchange between parties communicating data about entities in different domains. Through the incorporation of formal definitions they also allow the application of basic inference mechanisms when interpreting data exploiting taxonomic and other relations built into the ontology. For logic-based AI applications, more powerful ontologies are needed which reflect the full spectrum of language constituents and of their logical counterparts. They must enable the expression not only of traditional taxonomical and mereological relations but also for example of synonymy relations and reasoning at both the lexeme (single word) and phrase level. The terms in such ontologies will be defined using formulae of FOL (for example as standardized in CL). These formulae relate entities to each other via a transitive network of formulae as illustrated in the following simplified example:

*Natural language input:* Boats can float on water. To float, they enclose air, which gives them buoyancy.

*Formal representation:* $\text{boat}(x) \wedge \text{water}(y) \wedge R_1(x,y) \wedge \text{air}(z) \wedge (R_2(x,z) \wedge \text{buoyancy}(w) \wedge R_3(x,z,w)) \rightarrow R_1(x,y)$

Here lower-case letters represent entities, the $R_i$ represent relations, $R_1 = \text{floats\_on}$, $R_2 = \text{encloses}$ $R_3 = \text{gives}$ (quantifiers not shown for the sake of readability).

### 3.2 Putting the philosophy-driven machinery to work

To represent in logical form the full meaning of a given complex natural language expression E in a given domain and for a given purpose, we will need algorithms which, given E, can generate a corresponding logical formula using those terms in the relevant ontology which are counterparts of the constituent simple expressions in E. These algorithms, together with a consistent set of supporting ontologies, can thereby allow the representation in machine-readable form not merely of single expressions but of entire texts, even of entire bodies of literature, in which domain-specific knowledge is communicated in natural language form. To see how philosophy is already enabling applied science along these lines, let us look at a real-world example of an AI automaton used to automatically generate expert technical appraisals for insurance claims. Today, such claims are validated by mid-level clerks, whose job is to compare the content of each claim– for example the line items in a car repair or cardiologist bill– with the standards legally and technically valid for the context at issue. Deviations from the standard are detected and corresponding amounts are subtracted from the indemnity amount with a written justification for the reduction. Digitalization has advanced sufficiently far in the insurance world that claims data can be made available in structured digital form (the bills lines are stored as separate attributes of a table in a relational database). On the other hand, however the relevant texts specifying standards have until recently been represented only as free text strings. Now, however, by using technology along the lines described above it is possible to automate using AI both the digital representation of these standards and the results of the corresponding comparisons between standards and claims data. To achieve this, we developed an application that has the capability to do all of the following:

1. Recognise the exact type of bill and understand the context in which it was generated
2. Understand the type and contents of the bill (both the textual and the structured, quantitative content)
3. Transform the bills contents into a logico-mathematical representation
4. Identify the pertinent standards by querying the corresponding insurance knowledge base
5. Determine from 3. and 4. the appropriate repair benchmark for a claim repair of the relevant type and situation and determine the correct procedure to fulfil the claim. Claim, benchmark and procedure are here all expressed in mathematical logic.
6. Compare the bill to its benchmark by identifying departures from logical equivalence of line items in the claim from those in the benchmark.
7. Subtract the items on the bill that do not match the reference you are really subtracting money not items, surely
8. Output the justification for the subtractions

As to 1., human beings are able to make such assignments of context spontaneously– both for entire artefacts such as bills and for the single lines which are the constituent strands within such artefacts. Human beings live in a world which is meaningful in precisely this respect. The ability to make such assignments of context is, as we saw, difficult to replicate on the part of machines. As to 2., for the textual content in a bill to be interpretable we need ontologies covering both

the objects to which reference is made and the contexts and information artefacts associated therewith. We need also formal definitions of the relevant characteristics of these objects, of the terms used in the relevant insurance rules, and so forth. Together, these constitute the ontology mentioned below. The ontologies are built by hand, but involve a minimal amount of effort for those with expertise in the relevant domain (here: the contents of repair bills). These definitions are entailed by the bill and benchmark texts, which are automatically processed into logical representations without human interference. Viewed logically, the steps are as follows:

$$\text{text} \rightsquigarrow \Gamma \tag{1}$$

where $\rightsquigarrow$ means automated translation, and $\Gamma$ is the set of k-order intensional logic formulae[15] generated by the translation.

$$\Gamma \curvearrowright \Delta \tag{2}$$

where $\Delta$ is the set of (first-order or propositional modal) logic formulae automatically generated ($\curvearrowright$) from $\Gamma$.

$$\Delta \vdash \phi_i \in \Omega, \forall i = 1 \dots n \tag{3}$$

where $\vdash$ means: entailment using mechanical theorem proving, and $\phi_i$ is one of $n$ human-authored domain-specific formula entailed by $\Delta$.

$\Omega$ is an ontology comprising human-authored domain formulae $\phi_i$. Note that $\Omega$ is specific to a type of text (for instance repair bills) and to a pertinent context (for instance the regulation under which the repair occurs). $\Delta \cap \Omega \neq \varnothing$ only holds if the input text matches the type and context of the ontology.

In total, the process looks like this:

$$\text{text} \rightsquigarrow \Gamma \curvearrowright \Delta \vdash \phi_i \in \Omega, \forall i = 1 \dots n$$

Where the only manual input– creation of $\Omega$– has to be performed only once, at system design time.

As the case of BFO shows, philosophy is part of what is required to create in consistent fashion the successive layers of ontologies required to realize a system of the sort described. Knowledge of logic and adequate representation of reality is used to select appropriate logical dialects for different contexts and situations: temporal or deontic phenomena require a different logic than more basic phenomena which can be rendered using predicate logic. The mechanism described in equation (1) is used for steps 1, 2 and in part for steps 3 and 4 of the above list. Of course, these steps are realised not in a form as philosophical but rather as software of a sort which comprises both stochastic models and mechanical theorem provers. But their configuration, which conveys what is called the business process, is created using standard methods (including methods from linguistics) to formulate the logic of this process. For step 4, a set of context-dependent permissible benchmarks must be created. Its elements consist of a structured set of formulae as well as quantitative (typed variable) information. When a new bill arrives, the system

[15] 'k-order' means that predicates of the logic can predicate over other predicates arbitrarily often. 'Intensional' means that the range of predication in the logic is not restricted to existing entities.

performs steps 1-3 and then obtains an adequate benchmark from the computers memory. It then instantiates its typed variables in order to identify the values of those parameters relevant to the given situation, as exemplified in the following: consider (i) a car repair reference which specifies the replacement of then door, then the instantiation will be the actual part number for the door of the given car model. Or (ii) on a cardiologist bill for a coronary angiography, age, sex and diagnostic status of the patient are variables that would be used for parametrisation to identify where the reference matches the current situation and where it fails to match. The remaining operations of the machinery are purely technical a matter of software engineering and mathematics. The philosophy-driven AI application is now used by several German insurance companies, dental clinics and a leading academic hospital. It meets all the requirements listed in Table 1:

- *Exactness* – it has an error rate below 0.3%, which is below the best human error rate of 0.5%, because it will detect if it cannot entail any formula from text.
- *Security* – it is secure, as its stochastic model never works on its own, but misreactions to perturbing input would be detected by the logical model working right after it.
- *Robustness* – it is robust as it will always realise if it cannot interpret a context properly. It requires very little data for training (data parsimony) as the semantic training space it provides separates the data points so well.
- *Fidelity* – it has semantic fidelity. It not only allows, but it is based on inference and can easily use prior and world knowledge in stochastic (as Bayesian net) and deterministic (logical) from.

However, it is a very specific system, far away from the unrealistic notion of general artificial intelligence. It is rather an exact, philosophy-driven system of artificial instincts, as Stanislaw Lem coined the term in his essay "Weapon systems of the 21st century" (1983). It demonstrates the ability of philosophy to make applied sciences work.

## 4 Conclusion

The type of philosophy-driven AI application described in the above is not a one-off example; such systems are being successfully used in a range of different domains. Moreover, the method in question is generalizable to data of many different sorts, in principle– as the breadth of the available ontologies is extended and the sophistication of the algorithms is enhanced– without limit. We believe that these facts have implications beyond the merely technical (and, associated therewith, pecuniary). For they point to a new conception of the role of philosophy in human affairs. Many, especially in the twentieth century, have proclaimed the death of philosophy. Others have seen in philosophy a merely compensatory role– whereby philosophers might offer some sort of substitute for those traditions which in former times gave human beings the ability to interpret their lives as meaningful but which have since been eroded through the spread of modern science and technology. And the ways in which humans lives are meaningful– are full of meaning– did indeed play a role in our argument. Rather, we view the question of the role of philosophy from a broader historical perspective, drawing on the ways in which,

beginning already with the Greeks, philosophers have helped to lay the groundwork for social upheavals of the sorts associated, for example, with the birth of democracy or of market institutions, of new artifacts such as Cartesian coordinates, and sometimes of entire scientific disciplines. In this light we have shown in this paper that one place to look for a role for philosophy in the present day lies in the way philosophy can be used– and is already being used– to strengthen and enable applied sciences in the digital era. More specifically, we have demonstrated that philosophy can be of value in the creation of useful and realistic artificial intelligence applications.

**Acknowledgements** We would like to thank Prodromos Kolyvakis for his valuable review of the manuscript.

## References

[1] R. Arp, B. Smith, and A. Spear. *Building Ontologies with Basic Formal Ontology.* Cambridge, MA: MIT Press, 2015.

[2] M. Ashburner. "Gene Ontology: Tool for the unification of biology". In: *Nature Genetics* 25 (2000), pp. 25–29.

[3] G. S. Boolos, J. P. Burgess, and R. C. Jeffrey. *Computability and Logic.* Cambridge: Cambridge University Press, 2007.

[4] S. Carey and F Xu. "Infants' knowledge of objects: Beyond object files and object tracking." In: *Cognition* 80 (2001), pp. 179–213.

[5] Yining Chen et al. "Recurrent Neural Networks as Weighted Language Recognizers". In: *CoRR* abs/1711.05408 (2017).

[6] N Chomsky. "Three models for the description of language." In: *IRE Transactions on Information Theory* 2 (1956), pp. 113–124.

[7] S. B. Cooper. *Computability Theory.* London: Chapman & Hall/CRC, 2004.

[8] M. Dummett. *Origins of Analytical Philosophy.* Boston, MA: Harvard University Press, 1996.

[9] Shi Feng et al. "Right Answer for the Wrong Reason: Discovery and Mitigation". In: *CoRR* abs/1804.07781 (2018).

[10] S. Gelman. *The essential child: Origins of essentialism in everyday thought.* London: Oxford Series in Cognitive Development, 2003.

[11] S. A. Gelman and J. P. Byrnes. *Perspectives on Language and Thought.* Cambridge, MA: Cambridge University Press, 1991.

[12] S. A. Gelman and H. M. Wellman. "Insides and essences: Early understandings of the non-obvious." In: *Cognition* 38(3) (1991), pp. 213–244.

[13] J. J. Gibson. *An Ecological Theory of Perception.* Boston, MA: Houghton Miflin, 1979.

[14] A. Gopnik. "Explanation as orgasm and the drive for causal understanding". In: *Cognition and explanation.* Ed. by Keil F. and Wilson R. Cambridge, MA: MIT Press, 2000.

[15] V Gutierrez-Basulto and S. Schockaert. "From Knowledge Graph Embedding to Ontology Embedding? An Analysis of the Compatibility between Vector Space Representations and Rules". In: *Principles of Knowledge Representation and Reasoning: Proceedings of the Sixteenth International Conference,*

*KR 2018, Tempe, Arizona, 30 October - 2 November 2018*. 2018, pp. 379–388. URL: https://aaai.org/ocs/index.php/KR/KR18/paper/view/18013.

[16] T Hastie, T Tishirani, and Friedman J. *The elements of statistical learning*. 2nd ed. Berlin: Springer, 2008.

[17] P. J. Hayes. "The second naive physics manifesto." In: *Formal Theories of the Common-Sense World*. Ed. by J. R. Hobbs and R.C. Moore. Norwoord, 1985.

[18] M. Jaderberg and W. M. Czarnecki. "Human-level performance in first-person multiplayer games with population-based deep reinforcement learning." In: (2018). arXiv: 1807.01281v1.

[19] Jason Jo and Yoshua Bengio. "Measuring the tendency of CNNs to Learn Surface Statistical Regularities". In: *CoRR* abs/1711.11561 (2017).

[20] F. Keil. *Concepts, Kinds and cognitive development*. Cambridge, MA: MIT Press, 1989.

[21] F. Keil. "The growth of causal understanding of natural kinds." In: *Causal Cognition*. Ed. by D. Premack and J. Premack. London: Oxford University Press, 1995.

[22] I.K. Kim and E.S. Spelke. "Perception and understanding of effects of gravity and inertia on object motion." In: *Developmental Science* 2(3) (1999), pp. 339–362.

[23] A. Leslie. *The representation of perceived causal connection in infancy*. Oxford: University of Oxford, 1979.

[24] G. Marcus. "Deep Learning: A Critical Appraisal." In: (2018). arXiv: 1801.00631.

[25] J. McCarthy and P.J. Hayes. "Some philosophical problems from the standpoint of artificial intelligence". In: *Machine Intelligence* 4 (1969), pp. 463–502.

[26] Doug Medin and Brian H. Ross. "The specific character of abstract thought: Categorization, problem solving, and induction." In: *Advances in the Psychology of Human Intelligence* 5 (Jan. 1989).

[27] R. Millikan. *On clear and confused ideas*. Cambridge, MA: Cambridge Studies in Philosophy, 2001.

[28] Seyed-Mohsen Moosavi-Dezfooli et al. "Universal adversarial perturbations". In: *CoRR* abs/1610.08401 (2016).

[29] Karen O. Solomon, Doug Medin, and Elizabeth Lynch. "Concepts do more than categorize". In: *Trends in cognitive sciences* 3 (Apr. 1999), pp. 99–105. DOI: 10.1016/S1364-6613(99)01288-7.

[30] K. Papineni et al. "BLEU: a method for automatic evaluation of machine translation." In: *ACL*. ACL, 2002, pp. 311–318.

[31] R. Poplin, A. V. Varadarajan, and K Blumer. "Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning." In: *Nature Biomedical Engineering* 2 (2018), pp. 158–164.

[32] D. J. Povinelli. *Folk physics for apes: The chimpanzee's theory of how the world works*. London: Oxford University Press, 2000.

[33] Probabilistic Graphical Models: Principles and Techniques. *Koller, D. and Friedman, N.* Cambridge, MA: MIT,

[34] B. Rehder. "A causal model theory of categorization." In: *Proceedings of the 21st annual meeting of the Cognitive Science Society*. 1999, pp. 595–600.

[35] A. Robinson and A. Voronkov. *Handbook of automated reasoning*. Cambrdige, MA: Elsevier Science, 2001.

[36] S. Russell and P. Norvig. *Artificial Intelligence: A Modern Approach*. Harlow, Essex: Pearson Education, 2014.

[37] David Silver et al. "Mastering the Game of Go with Deep Neural Networks and Tree Search". In: *Nature* 529.7587 (Jan. 2016), pp. 484–489. ISSN: 0028-0836. DOI: 10.1038/nature16961.

[38] B. Smith. "Ontology". In: *Blackwell Guide to the Philosophy of Computing and Information*. Blackwell, 2003, pp. 155–166.

[39] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. Cambridge, MA: The MIT Press, 2018.

[40] J. B. Tenenbaum. *A Bayesian framework for concept learning*. Cambridge, MA: Massachusetts Institute of Technology, 1999.

[41] J. B. Tenenbaum and T. L. Griffiths. "Generalization, similarity, and Bayesian inference." In: *Behavioral and brain sciences* 24(4) (2001), pp. 629–640.

[42] A Vaswani et al. "Attention Is All You Need". In: *CoRR* abs/1706.03762 (2017). arXiv: 1706.03762. URL: http://arxiv.org/abs/1706.03762.

[43] Suncong Zheng et al. "Joint Extraction of Entities and Relations Based on a Novel Tagging Scheme". In: *CoRR* abs/1706.05075 (2017).