AI & SOCIETY (2024) 39:1653–1665
https://doi.org/10.1007/s00146-023-01631-2

MAIN PAPER

Artificial understanding: a step toward robust AI

Erez Firt[1]

Received: 20 June 2022 / Accepted: 16 January 2023 / Published online: 15 March 2023

Abstract

In recent years, state-of-the-art artificial intelligence systems have started to show signs of what might be seen as human level intelligence. More specifically, large language models such as OpenAI's GPT-3, and more recently Google's PaLM and DeepMind's GATO, are performing amazing feats involving the generation of texts. However, it is acknowledged by many researchers that contemporary language models, and more generally, learning systems, still lack important capabilities, such as understanding, reasoning and the ability to employ knowledge of the world and common sense in order to reach or at least advance toward general intelligence. Some believe that scaling will eventually bring about these capabilities; others think that a different architecture is needed. In this paper, we focus on the latter, with the purpose of integrating a theoretical–philosophical conception of understanding as knowledge of dependence relations, with the high-level requirements and engineering design of a robust AI system, which integrates machine learning and symbolic components.

Keywords Artificial intelligence · Understanding · Robust AI · Dependence relations

1 Introduction

When OpenAI's (then) new language model, GPT (Generative Pre-Trained Transformer)-3, went public, there was an ongoing debate regarding the level of its (general) intelligence. Part of it can be attributed to conversations like the following, which highlights "common-sense", perhaps the holy grail of general intelligence:

Q: What is your favorite animal?

A: My favorite animal is a dog.

Q: Why?

A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?

A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: Are there any animals with three legs?

A: No, there are no animals with three legs.

Q: Why do not animals have three legs?

A: Animals do not have three legs because they would fall over.[1]

OpenAI was founded with the ambitious vision of the long sought for artificial general intelligence (AGI), intelligent systems that possess all the depth, variety, and flexibility of the human mind. Could GPT-3, its predecessors and alike models be the first step toward creating this sort of intelligence?

Ordinarily, I would have argued at this point that the debate concerning the level of intelligence of contemporary language models should be settled, i.e., that language models such as GPT-3, BERT, XLNet and others (e.g., the BERT modifications RoBERTa and ALBERT) cannot understand, reason or in general, think, at least not in the way humans do. However, recent developments in scaling, architecture and abilities seem to keep such a statement controversial, to a certain degree; Google's latest and greatest Pathways Language Model (PaLM), has 3 times the number of GPT-3's parameters, a new AI architecture and the ability to explain jokes and inferences, among other things.[2] Thus, the rest of this paper is dedicated to examining the debate and pointing out what might be still missing.

✉ Erez Firt
erezfirt@gmail.com

[1] Center of Humanities and AI, Haifa U. and The Technion, Haifa, Israel

[1] See https://lacker.io/ai/2020/07/06/giving-gpt-3-a-turing-test.html, accessed on May, 2022.

[2] See Dean (2021) and Chowdhery (2022).

GPT-3 is a large-scale language model: it has been trained to learn a probability distribution over tokens (roughly, pieces of information considered as discrete elements, and the building blocks of natural language) on the internet. Its training consists of being required to predict the next tokens in a sequence; the model's parameters are then iteratively modified depending on the accuracies of its predictions so that its future predictions on similar data become more accurate. Having learned the probability distribution, GPT-3 is then able—given a context (that might include a few examples of the desired form of response) and a prompt—to convincingly predict which paragraphs should come next.

PaLM is different in this sense. Instead of digesting one modality of information at a time—text in the case of GPT-3—PaLM's architecture could enable multimodal models that encompass vision, auditory, and language understanding simultaneously.

These language models are trained on TBs of text from multiple sources,[3] thus creating a model that can, given a context, correctly make medical diagnoses, draft legal text based on plain English, generate poetry, prose, philosophical reflections and art criticism and accurately translate foreign languages into English, among other things. On top of all this, current state-of-the-art language models have another impressive ability of being few-shot learner models. This means that instead of requiring additional domain-specific training to perform specialized tasks, as older models required, few-shot learners are able to accomplish that, given just a small number of examples or demonstrations.

This is an amazing engineering achievement. I have no desire to belittle its magnitude in any way whatsoever. It is also an important step on our way toward general intelligence. Nevertheless, assuming we aim to construct an autonomous machine with something similar or approaching general intelligence, it is still not enough. This is due to the fact that current language models *can only* generate sequences of tokens (in many cases only after a preliminary step of prompting or demonstration), based on a probability distribution learned from the text they were trained on. In some cases, increasing the quantity may lead to better predictions; for example, in generating responses in arithmetic cases of addition of four digit numbers (in comparison with additions of two or three digit numbers), where supposedly the reason is that calculations with many-digits numbers are encountered less frequently in the training data; or in scaling the number of parameters and constructing smart architectures, which would enable the system to overcome some of the known shortcomings of current language models, as in the case of PaLM.

However, not all cognitive abilities can depend on or be transformed into text prediction problems; where logical reasoning or common sense are needed, *most* current language models perform less successfully. I emphasize the quantifier most, because the recent release of PaLM brings to the fore a very significant point we shall elaborate on in following sections, i.e., the debate between proponents of scaling and quantity and proponents of qualitative capabilities. In other words, between those who believe that increasing the amount of training data, processing power and parameters will lead to better language models and eventually to understanding, and those who believe that we currently lack essential capabilities that cannot, in principle, emerge only by scaling. Sutton (2019), for example, argues that the history of AI research has taught us a lesson: AI systems based on data and computation are more effective than those based on human knowledge and rules, "by a large margin." Paraphrasing on a Geoff Hinton tweet[4] and the Hitchhiker's Guide to the Galaxy's answer to the question of life, the universe and everything, we can say that the quantitative approach suggests that the answer to all these questions is 42 (trillion parameters).

Recently, Firt (2020) has pointed out a few necessary conditions for general intelligence, fore and foremost among which are learning, understanding and reasoning. Here, we focus on understanding. We follow the conceptualization of understanding as knowledge of causes or knowledge of relations and dependencies between parts in a system.[5] Another important aspect of understanding is the ability to manipulate these relations in a manner which "allows the understander to anticipate what would happen if things were relevantly different…[and] make correct inferences about a world in which the relevant differences obtain." (Firt 2020: §3.1) We understand something—be it physical phenomena or the behavior of other people—when we have knowledge of it that can be grounded in causes, or more precisely, in dependence relations, of which causal dependence is but one instance.[6]

We believe that understanding is a necessary component on our path toward implementing the kind of artificial intelligence that Marcus (2020) calls 'robust AI':

"[An] intelligence that, while not necessarily superhuman or self-improving, can be counted on to apply what it knows to a wide range of problems in a systematic and reliable

[3] As per its creators, GPT-3 has been trained on over 175 billion parameters and 45 TB of text gathered from all over the web. See for example https://www.springboard.com/blog/data-science/machine-learning-gpt-3-open-ai/, accessed on 27th of April, 2022. PaLM is a 540 billion parameter language model trained on "780 billion tokens of high-quality text" (Chowdhery 2022: 3).

[4] See https://twitter.com/geoffreyhinton/status/1270814602931187715, accessed on 1st of May, 2022.

[5] For these views, see Grimm (2011, 2014) and Thorisson and Kremelberg (2017).

[6] See Grimm (2014) and Kim (2010).

way, synthesizing knowledge from a variety of sources such that it can reason flexibly and dynamically about the world, transferring what it learns in one context to another, in the way that we would expect of an ordinary adult." (ibid: 3).

The emphasis being on building an AI we can *trust,* and which must be endowed with what Marcus (ibid: 5) calls deep understanding. As Marcus stresses,

"If our AI systems do not represent and reason over detailed, structured, internal models of the external world, drawing on substantial knowledge about the world and its dynamics, they will forever resemble GPT-2: they will get some things right, drawing on vast correlative databases, but they won't understand what's going on, and we won't be able to count on them, particularly when real world circumstances deviate from training data, as they so often do." (ibid: 9).

This need for 'deep understanding' is also featured in other, closely related inquiry fields. For example, in Machine Ethics there is an ongoing discussion revolving the construction of different types of Artificial Moral Agents (AMAs). When referring to the construction of a certain type of AMA, i.e., functional AMA, Allen and Wallach (2011) suggest what they call the Hybrid approach; an approach to achieving artificial morality which integrates top-down and bottom-up approaches. In brief, top-down approaches refer mainly to rule-based AI, where human-made rules are inserted into the system, and bottom-up approaches refer to methods which begin with raw data, or some random/baseline version of the system, and learn or develop from that point on (e.g., using methods of machine learning or genetic algorithms). Again, the goal is for our future moral machines to understand the words we use (semantics), i.e., to be able to link the tokens they generate to objects in the world, and to relations these objects have with other objects (the pragmatics, the context), so it will not make disastrous mistakes when making predictions.

But what does it take to implement artificial understanding? We begin with the notion of understanding as the knowledge of relations and dependencies. The main purpose of this paper is to integrate this theoretical–philosophical conception of understanding with the high-level requirements and engineering design of a robust AI system. How do we achieve that? First, we discuss a debate between two approaches; let us term them the Deep Learning approach and the Integrative approach. Both approaches acknowledge the capabilities needed for AI systems to reach the next step. The former aims to achieve this advancement by implementing these capabilities within a framework of (deep) learning, while the latter aims to do the same thing by integrating deep learning with symbolic AI; we discuss these issues in more details below. But the main point does not rest on the details of the actual implementation. It concerns the necessary components that compose the general solution (that may have more than one way of being implemented): a data preserving structure that holds the dependence relations between entities discovered by the system. That is, a data representation of the causal and other dependence relations, which are discoverable from already-learned data correlations and the cognitive models[7] constructed by the system; these cognitive models describe the state of affairs in terms of dependencies between the entities and their location in space–time; A knowledge base containing background and common sense knowledge of the world; learning mechanisms for efficient unsupervised autonomous learning; and lastly, a reasoning system that can leverage all of the above, make inferences, predict and provide decent human readable explanations for its decisions.

Thus, the structure of this paper is as follows: in Sect. 2, we discuss the philosophical analysis of understanding as knowledge of dependence relations. In Sects. 3 and 4, we examine two debates concerning the path to understanding: the 'quantity vs. quality' debate and the Deep Learning approach vs. Integrative approach debate and propose a general high-level architecture based on the ideas presented thus far. Section 5 is comprised of our concluding remarks.

2 Understanding as knowledge of dependence relations

In this section, we focus on types of understanding usually referred to as understanding-why and objectual understanding; the former is implied in sentences that take the form "I understand why X" (for example, "I understand why this and that happened"), whereas the latter is implied in sentences that take the form "I understand X", where X can be thought of as a body of information or a subject matter.

There are several epistemological views of understanding; in this paper, we focus on what Grimm (2014) terms the traditional view, i.e., the idea that understanding derives from knowledge of causes.[8] Grimm (2014) provides a survey of other views of understanding, discusses the major objections to the traditional view, as they appear in the literature, and provides satisfactory replies to all of them. I saw no point in rehearsing them here.

In what follows, we want to depict a view, which is to some extent wider or more encompassing than understanding as knowledge of causes: Understanding arises from knowledge of relations and dependencies, of which causal

[7] We shall use cognitive models and world models interchangeably; both refer to a cognitive ability to construct inner models of the world or the immediate surroundings and examine and manipulate them.

[8] For further contemporary support for this view, see Grimm (2014), fn. #1.

dependencies or causal relations are but one instance.[9] Henceforth, when we use the notion of dependence relations or dependencies (following Kim (2010 [1994]: 183), we mean a relation between states, events, facts, properties, regularities or entities; two prominent examples of which are causal and mereological dependencies.

To understand why X, i.e., to understand why this and that happened, or to understand X, where X can be thought of as a body of information or a subject matter, one has to know what led to X (as a concrete phenomenon), or how the parts and elements of X, as a system or a structure, depend upon one another in various ways, in cases of objectual understanding.[10] Moreover, understanding, as we portray it, is also about the ability to manipulate the relations, dependencies and in general the structure that one, i.e., the understander, perceives, be it the structure of an object or the complex structure of relations between different entities populating his surrounding. Wilkenfeld (2013) suggests that one understands when one possesses a representation of that which is understood that is sufficiently robust to be manipulable for inferential and practical purposes. In other words, understanding occurs when we have a robust mental representation of the thing to be understood. This robustness is expressed by the ability of the understander to manipulate and tweak this representation to examine inferences and take actions. In the same spirit, Grimm (2011) suggests that manipulating the "system" allows the understander to "see" the way in which "the manipulation influences (or fails to influence) other parts of the system" (Grimm 2011: 11). Thus, according to Grimm, understanding the relationships between relevant parts of a subject matter amounts to manipulating the system by changing parts of it and observing the impact on the overall system. He refers to such ability as *grasping*, and suggests that it also allows the understander to anticipate what would happen if things were relevantly different. It allows the agent to make correct inferences about a world in which the relevant differences obtain.

Let us reiterate the details of 'knowledge of dependence relations' and 'understanding as manipulation of relations' and see not only how they are integrated, but also how such integration gives rise to a more complete picture of understanding. Grimm (2014) defends the traditional view of understanding as knowledge of causes by updating the way proponents of such a view should understand 'knowledge of causes'. To begin with, he defuses objections often made by opponents of the traditional view, who take knowledge of causes to mean the knowledge of the truth of a certain causal proposition, of the following structure:

(a) S has knowledge of the cause of p just in case.
(b) S knows that p because of q.

Such an account, continues Grimm (2014: §II), should be abandoned or at least supplemented. Understanding is not about knowing that a certain proposition is true or even necessarily true. To understand, one should "grasp" or "see" the "modal relationships that obtain between the properties (objects, entities) at issue. In the case of knowledge of causes in particular, what would be seen or grasped would be how changes in the value of one of the terms of the causal relata would lead (or fail to lead) to a change in the other." (ibid: §IV) In other words, understanding as knowledge of causes amounts to having the relevant causal information, i.e., how things are, and also how things could have been, or how things would turn out to be in case they were manipulated/changed/different.[11] To further refine this last statement, we can say, following Grimm (2014) that one understands when one grasps "the modal relationship that obtains between the terms of the explanation." (ibid: §VI).[12]

To conclude this modified version of knowledge of dependence relations, we stress again its two most important points: understanding is the knowledge of dependencies and relations between phenomena and entities relevant to the thing to be understood, and the ability for modal representation of these relations in a way that allows the understander to manipulate them and examine other related possibilities; or in other words, the ability for counterfactual reasoning with regards to these dependencies.

Not to get ahead of ourselves too much, but just to mention that two indispensable components of our robust AI architecture already feature prominently in this suggested

[9] The idea of understanding as knowledge of dependence relations (i.e., the idea of expanding the notion of causation to dependence) is supported by several prominent philosophers: Woodward (2003: 6) claims that "any explanation that proceeds by showing how an outcome depends… on other variables counts as causal"; Greco (2010: 9) likewise argues that "understanding involves 'grasping,' 'appreciating,' or knowing causal relations taken in the broad sense: i.e., the sort of relations that ground explanation."; and Kim (2010 [1994]: 183) argues that "dependence relations of various kinds serve as objective correlates of explanations.".

[10] See Grimm (2011). Riggs (2003: 20) emphasizes the importance of the relations among parts and between the parts and the whole, when trying to understand a subject matter. Other philosophers that support the idea that understanding is largely about grasping or coming to know the relations between entities, include Zagzebski (2001: 242, 2009: 142), who asserts that understanding "involves grasping relations of parts to other parts and perhaps the relations of part to wholes"; As Grimm (2014: §VI) puts it, "the common thought here seems to be that the primary objects of understanding are the relationships (or structures) that hold among the various elements of reality.".

[11] This description of Grimm has some interesting similarities to Pearl and Halpern (2001) Structural-Model approach. See especially §3.1.

[12] Of course, this "grasping" is fallible and the understating can be a matter of degree. See Grimm (2014: fn #21 and §V-VI, respectively).

view of understanding: cognitive models of which relations and dependencies between entities are a necessary part, and some kind of a knowledge base containing common sense or background knowledge of the world. The capability of a system to take effective action in the world (at least) partly depends on its ability to manipulate relations, and examine the cognitive models in which those relations are modified. In other words, when one understands something—an object, a situation, etc.—one is able to manipulate the thing understood effectively, e.g., in a way that will enable her to achieve goals.

More formally, a statement of the sort "agent A understands an object O",[13] is true in a certain context, if and only if A has a cognitive model R of O that A can modify in certain ways to produce R', which can then be used to manipulate or make inferences pertaining to O.

3 How to get there

As aforesaid, current state-of-the-art AI systems do not possess the kind of understanding portrayed in the previous section. In particular, just for illustration purposes, large language models such as GPT-3 perform amazingly (almost magically), when are faced with familiar tasks, i.e., tasks within their training distribution or on the rarer cases when they are able to extrapolate beyond their training distribution, but might perform miserably when faced with tasks outside their training distribution.

Since its launch, in 2020, many critics have tried to expose limitations, failures and faults in the way GPT-3 operates,[14] part of which were answered in a way that exposed faults in the objections themselves (e.g., misuse or incorrect use of prompting).[15] Marcus and Davis (2020b) were also part of this wave of criticism. Their critique is based on a few examples, where the system got it wrong. Here are a few reasoning-related illustrations (their input in bold, GPT-3's completion in italics, their explanation in brackets):

Psychological reasoning

Janet and Penny went to the store to get presents for Jack. Janet said, "I will buy Jack a top." "Don't get Jack a top," says Penny. "He has a top. He will *get a top." "I will get Jack a top," said Janet.*

[Within a single sentence, GPT-3 has lost track of the fact that Penny is advising Janet against getting a top because Jack already has a top. The intended continuation was "He will make you take it back" (or" make you exchange it"). This example was drawn directly from Eugene Charniak's 1972 PhD thesis; nearly 50 years later, it remains outside the scope of AI natural-language technology.]

Biological reasoning

You poured yourself a glass of cranberry juice, but then you absentmindedly poured about a teaspoon of grape juice into it. It looks okay. You try sniffing it, but you have a bad cold, so you can't smell anything. You are very thirsty. So *you drink it. You are now dead.*

[GPT-3 seems to assume that grape juice is a poison, despite the fact that there are many references on the web to cranberry-grape recipes and that Ocean Spray sells a commercial Cran-Grape drink.]

Overall, there is no real doubt about whether GPTs can fail and produce nonsensical text. Marcus and Davis's point is that we cannot know *when* it fails.

"The trouble is that you have no way of knowing in advance which formulations will or won't give you the right answer… The optimist will argue (as many have) that because there is some formulation in which GPT-3 gets the right answer, GPT-3 has the necessary knowledge and reasoning capacity—it's just getting confused by the language. But the problem is not with GPT-3's syntax (which is perfectly fluent) but with its semantics: it can produce words in perfect English, but it has only the dimmest sense of what those words mean, and no sense whatsoever about how those words relate to the world." (ibid.)

It is generally agreed that current state-of-the-art AI (deep learning) systems share the problem of misunderstanding or misrepresenting the world in which they reside, or in which we want them to operate. Large language models serve us here as a relatively convenient illustration of this issue—they learn how tokens relate (syntax), but not how they are connected to the physical world (semantics), under different circumstances (pragmatics).

Clearly enough, to avoid syntax-based nonsensical errors and to be able to trust these systems we need to elevate them to the next level and provide them with the ability to connect to the world. As Marcus and Davis (2019) put it:

"Start by developing systems that can represent the core frameworks of human knowledge: time, space, causality, basic knowledge of physical objects and their interactions, basic knowledge of humans and their interactions. Embed these in an architecture that can be freely extended to every kind of knowledge, keeping always in mind the central tenets of abstraction, compositionality, and tracking of individuals. Develop powerful reasoning techniques that can deal with knowledge that is complex, uncertain, and incomplete and that can freely work both top-down and bottom-up. Connect

[13] Object O is any object of understanding and it can include theories in physics, a certain proof in mathematics or logic, a person (as in, "I understand my friend"), a story or an event, an action, or a phrase in a language, to give some examples.

[14] For a structured review of GPT-3's scope and limitations, see for example Floridi and Chiriatti (2020).

[15] See the comprehensive technical blog of Gwern Branwen—https://www.gwern.net/GPT-3, Accessed 12-May-2022.

these to perception, manipulation, and language. Use these to build rich cognitive models of the world. Then finally the keystone: construct a kind of human-inspired learning system that uses all the knowledge and cognitive abilities that the AI has; that incorporates what it learns into its prior knowledge… Put all that together, and that's how you get to *deep understanding*." (ibid.)

This is already a suggestion that incorporates architectural changes that should be made to contemporary systems. Before we move to discuss these architectural changes, there are still two unresolved debates to consider, the first of which starts prior to accepting Marcus and Davis's suggestion and the second revolves around the issue of the right way to implement some or all of the features that appear in their suggestion.

The first issue concerns the controversy 'quantity vs. quality'. To be sure, we need to trust our systems and we cannot do that if, once in a while, they make a mistake that a 3-year-old would not make. This controversy revolves around the best way to achieve or improve the level of trustworthiness of these systems. One approach is to rely on scale—the idea, or hope, is that AI systems can perform better if we gather more data, make sure that the quality of the data is high, add more parameters and apply deep learning at increasingly large scales.[16] On the other hand, Marcus (2022) represents the other approach, the qualitative approach. According to him, "Scaling the measures Kaplan and his OpenAI colleagues looked at—about predicting words in a sentence—is not tantamount to the kind of deep comprehension true AI would require." (ibid.) He goes further to conclude that "research from DeepMind and elsewhere on models even larger than GPT-3 have shown that scaling starts to falter on some measures, such as toxicity, truthfulness, reasoning, and common sense…[and] making GPT-3-like models bigger makes them more fluent, but no more trustworthy." (ibid.)[17] The qualitative approach's general idea is that we need more than just scaling in the areas of data and model's parameters. We need capabilities that current state-of-the-art AI systems lack.

3.1 A deeper look into the quantitative approach—the case of PaLM

As mentioned above, since the launch of GPT-3 improvements have been made in several aspects related to large language models. In the context of the quantitative approach, which supports the general idea that scaling can lead to systems that understand the world, we should take a look at the latest and greatest in this domain, i.e., Google's Pathways Language Model (PaLM). Chowdhery et al. (2022) detail the scaling, improvements and achievements (in benchmarks, compared to other language models) of PaLM; in what follows, we take a brief look at these aspects to see whether we can draw conclusions regarding the reasonability of the quantitative/scaling approach.

Chowdhery et al. (2022) specify four main points, related to scale and architecture, which led PaLM to achieve "breakthrough performance" on a number of tasks, and more specifically, "state-of-the-art few-shot results across hundreds of natural language, code, and mathematical reasoning tasks." (ibid: 3) The following points are highlighted:

(1) Model depth and width: Training of a 540B parameter language model on 6144 Tensor Processing Units (TPU) v4 chips.
(2) No. of tokens trained: in the case of PaLM, 780 billion tokens of data.
(3) Training corpus quality: cleaner datasets from diverse sources.
(4) Architecture: The use of Pathways, a new AI architecture, which (allegedly) enables its users to train a single model to handle multiple tasks, to receive input from multiple modalities (e.g., vision, auditory, and language understanding) and make models sparse and efficient. In other words, "Pathways will enable a single AI system to generalize across thousands or millions of tasks, to understand different types of data, and to do so with remarkable efficiency." (Dean 2021).

Having said that, let us examine what can we know at this point in time and what is still uncertain. First, we know that at least up to this point on the scaling axes, scaling still achieves better performance in a number of tasks.[18] Although this cannot be argued with certainty in the case of PaLM (as there are various possible reasons for the increased performance, as we shall presently discuss), other models (e.g., GATO[19]) support this claim. Second, at least in the PaLM case, we cannot attribute the increasing success in various tasks to scaling per se, as the influence of other factors, e.g., the quality of the training data or the change in architecture, were not carefully filtered out, as is also admitted by Chowdhery et al. (2022, §13). Third, there is the issue of diminishing returns; the effect of scaling on performance may fade together with the increase in size. As the authors stress, this has not yet happened: "the results presented here suggest that the improvements from scale for

[16] See for example Kaplan et al. (2020) and Sam Altman's, Open AI's CEO, blog post, which celebrates "Moore's Law for Everything." (https://moores.samaltman.com/, Accessed on the 13-May-2022).

[17] See Rae et al. (2022) and Thoppilan et al. (2022) for support for this claim.

[18] See Chowdhery et al. (2022) and Reed (2022).

[19] See Reed (2022).

few-shot language understanding have not yet plateaued." (ibid: §14) At this point, we have no empirical evidence as to whether future scaling will keep on increasing performance, or whether the process of diminishing returns will start to manifest itself. And also, maybe more importantly, following our goal to reach robust AI that truly understands the world, the question remains, even given that scaling continues to bear fruits, in terms of increased performance, can learning from data alone provide us with understanding as knowledge of dependence relations?

In this paper, I assume that it cannot, for the following reasons: In general, machine learning uses statistical reasoning and methods to find patterns in data. After digesting a certain amount of data (i.e., the training/distribution data), the system can then make statistical predictions when given a new input. When the input is close enough to the training distribution, the prediction is more accurate. Many of the champions of machine learning agree that statistical methods are not enough for robust artificial intelligence: "Despite its success, statistical learning provides a rather superficial description of reality that only holds when the experimental conditions are fixed." (Scholkopf et al. 2021: 613) What is needed, according to these authors,[20] is to integrate "causality, with its focus on representing structural knowledge about the data generating process that allows interventions and changes," as this can take us a step toward robustness, and "can contribute toward understanding and resolving some limitations of current machine learning methods." (ibid.)

These are the fundamental intuition and reasoning that lead us to explore the qualitative approach, which basically states that current machine learning methods need some enhancement in the form of new capabilities, in order to step up. The remainder of this paper is dedicated to this approach and the two prominent alternatives to implement it, around which the next debate revolves.

4 Robust AI doth not live by data only

Ever since the beginning of AI research, two general approaches to building AI systems have been explored, implemented and tested, i.e., the rule-based approach, also referred to as symbolic AI, and the machine learning approach. The former manipulates symbols in the same way most software programs do, i.e., by having sets of symbols (codes) to represent information,[21] and by processing those symbols using mathematical and logical operations; the latter uses linear regression algorithms or structures called neural networks in order to produce statistical predictions regarding similar cases. Without going into the history of AI research over the last seven decades, it is enough for the purposes of this paper to acknowledge that current state-of-the-art systems built according to any one of these approaches have problems and certain type of tasks they perform less successfully: symbolic systems tend to be more complicated and do not perform well on tasks such as image and speech recognition and other language related tasks; Deep learning methods perform extremely well on these tasks after being trained on large data sets. On the other hand, deep learning systems are brittle (i.e., tend to be inaccurate, or even err miserably, when it comes to out-of-distribution cases), their results are difficult to explain (and therefore hard to trust) and they are associated with several ethical and socially related harms. For example, large language models can cause discrimination, exclusion and toxicity related harms (e.g., promoting stereotypes, causing unfair discrimination, inciting violence or causing offense), information hazards (e.g., providing information about how to easily perform unethical or illegal actions), human–computer interaction harms (e.g., "conversational agents" conversing with humans can exploit human psychological vulnerabilities), to name just a number of possible types of harms.[22] Thus, it seems that an obvious solution, or at least one alternative that should be explored, is the integration of the rule-based and machine learning approaches.

This integration of approaches lies at the heart of the debate between the Deep Learning approach and the Integrative approach; both parties agree that current state-of-the-art deep learning systems have the problems outlined above; both parties agree that capabilities such as reasoning, common sense and knowledge of causal relations, cognitive models and real understanding are needed. However, they defer on the way to achieve such capabilities. As Yoshua Bengio puts it, in his 2020 AI debate with Gary Marcus, "We would like to build in some of the functional advantages of classical AI rule based symbolic manipulation in neural nets, but in an implicit way."[23] His rival in this debate, Gary Marcus, is a known champion of the Integrative approach and has for long advocated for what has been recently known as the Symbolic-Deep-Learning or Neuro-Symbolic AI approach.[24]

[20] See also related work of Bengio et al. (2020).

[21] For example, in the universally used ASCII code, the binary numbers 01000001 and 01000010 stand for (are symbols for) the letters A and B, respectively.

[22] See Weidinger (2021).

[23] See min. 46:30, https://www.youtube.com/watch?v=EeqwFjqFvJA, Accessed 22-May, 2022.

[24] Representative studies of this approach include Mao (2019), Raedt et al. (2020), Oltramari (2020), Chitnis et al. (2021), and a relevant interesting research summary from IBM: https://research.ibm.com/blog/ai-neurosymbolic-common-sense, Accessed 25-May-2022.

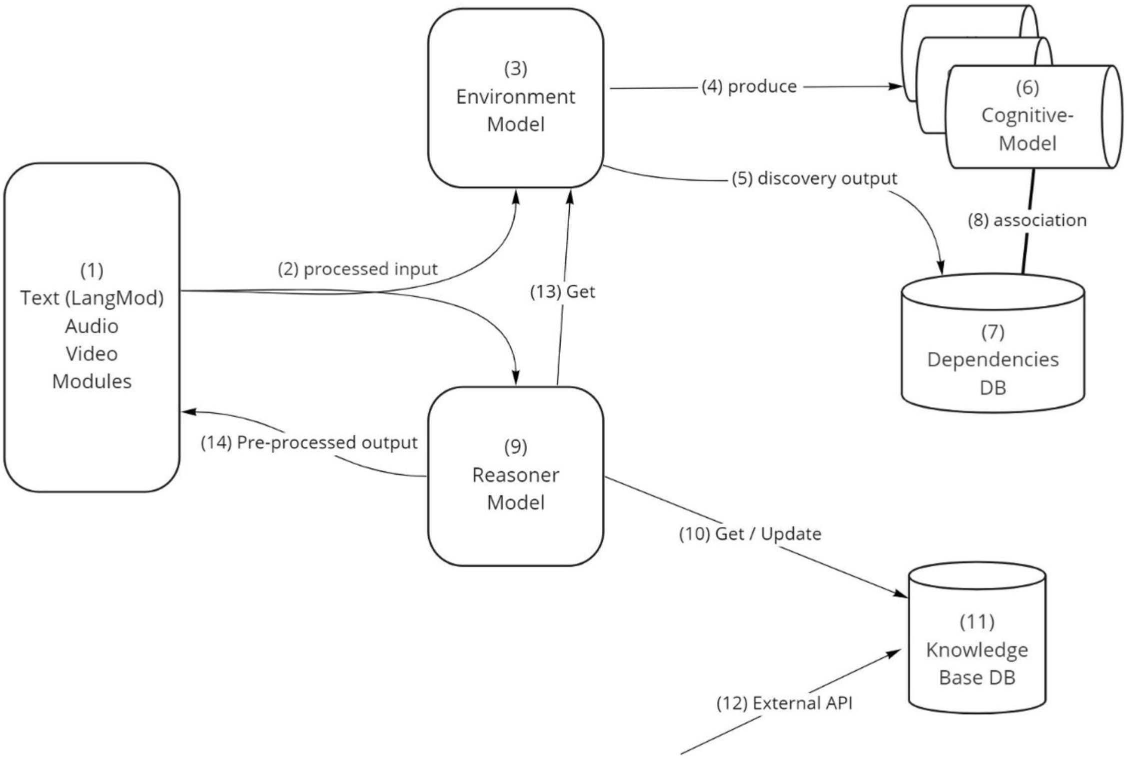
As regards the Integrative approach, Marcus predicts that "within a few years… many people will wonder why deep learning for so long tried to do so largely without the otherwise spectacularly valuable tools of symbol manipulation; virtually all great engineering accomplishments of humankind have rested on some sort of symbolic reasoning," (Marcus 2020: 16–17) and goes on to specify what he believes to be the key components of such robust systems: a reasoning component, "that can leverage large-scale background knowledge efficiently, even when available information is incomplete," (ibid: 40) cognitive models, without which "systems like these are lost. Sometimes they get lucky from statistics, but lacking cognitive models they have no reliable foundation with which to reason over." (ibid: 42) In addition, understanding, which according to Marcus is closely related to the ability to infer a model of the thing we wish to understand, "and ultimately to be able to make inferences about how it operates, and what might happen next." (ibid: 41).

Interestingly enough, his rival in this debate, Yoshua Bengio, agrees with the gist of these statements. In his 2020 talk[25] in 'AI in 2020 & Beyond' he specifies the following problems for (what was then) current deep learning systems: the number of samples a system should be trained on, so it may operate (more or less) accurately, the dependence on human-provided labels, the kind of errors deep learning systems make, which implies their lack of understanding. To overcome these problems and reach the goal of robustness and trustworthiness we require and want, he then stresses several points to be achieved: generalization beyond the training distribution, in order to perform well under unknown circumstances; the development of the ability to create and manipulate inner world models, which can give these systems a capability, analogous (at least, to a certain extent) to human imagination; the ability to discover and manipulate causal structures, to be able to better cope with changes in their surroundings; and the development of the ability to employ common sense, so these systems may get better at understanding the world around them and in this way acquire a better sense of the meaning of the concepts we use.

Bottom line, when examining what both researchers suggest, one can come to an understanding that both sides agree on many important high-level details; the rest may seem like implementation details. Thus, in the remainder of this section, I sketch a suggestion for a high-level software architecture for an artificial understander: the main entities, their role, and their inter-relations. In this suggestion, I follow the general Integrative approach, and offer a solution wherein the initial proposals made by learning sub-systems can be verified and validated against the system's knowledge of the world and common sense on the one hand, and its perception of its surroundings, the entities populating it and their relations, on the other hand.[26] The following is a high-level design of such architecture, succeeded by descriptions of the different components and their relations:

[25] See https://www.youtube.com/watch?v=GibjI5FoZsE, Accessed 24-May-2022.

[26] See for example Sychev (2021: 731–2) for a similar approach: "we can develop a human-like intelligence as a system where neural networks generate new ideas and strategies given the context and random noise … symbolic reasoning assesses their applicability and the level of risk using available knowledge before trying them in the environment, then the ideas that passed logical verification are implemented under conscious control.".



(1) One or more modules of learning systems,[27] providing the system with audio, visual and text input information. This information can take several forms: visual information (e.g., post-processed images or videos containing object information), audio and text containing queries, descriptions or conversational input, etc. In other words, these modules provide sensory input to the system. They receive or actively collect information from and about its surroundings, e.g., text input that can be received as textual requests, visual input from cameras, auditory input that can be transformed into text input and can be analyzed to extract content, sentiment and other pragmatic aspects.[28] Based on this analysis, a coordinator sub-module can then construct a request for additional information, based on the received processed input, as part of a call to other services in the system, as detailed below.

(2) Following (1), processed input may contain information regarding objects and their relations (the result of processed images, videos, audio and text), suggested output in the form of generated text, requests for additional supplementing data (in the form of requests for data to be retrieved from other services in the system), or requests for the execution of certain tasks (e.g., a request from the Reasoner Model service to perform a logical analysis of a certain proposition or claim).

(3) The Environment model is a sort of a gateway/ Façade,[29] whose responsibility is to allow multiple types of operations[30] on data related to the system's

[27] As stated in Dean (2021), "Pathways could enable multimodal models that encompass vision, auditory, and language understanding simultaneously."

[28] For example, when conversing with a human, the system can analyze the auditory input and extract the semantic value as well as the sentiment, and integrate this analysis with a visual analysis of the posture and gestures, for example, to gain a more complete understanding of the pragmatics of the conversation.

[29] The Façade design pattern is a software structural design pattern whose responsibility is to hide the complexities of a system and provide an interface using which any user of the system can access it.

[30] These types of operations, commonly implemented to support persistent storage applications, are usually referred to by the acronym CRUD: Create, Read, Update, Delete.

surroundings: cognitive models [see (4, 6)], relations and dependencies between objects populating the environment [see (5, 7)] and the relation between these models and the dependencies [see (8)].

(4) Following (3), and by employing the processed input from (2), the environment model is responsible for producing or updating cognitive/world models.[31] World models are internal models of the surroundings or environment in which the agent or system resides or wish to take action in. The accuracy of the model, in terms of the objects contained in it, their properties, relations and location, determine the system's ability to manipulate and act within the environment this world model describes, making the constant updating of the model necessary. This internal model can refer to any type of environment (e.g., physical/realistic, fictional, and a game world).

(5) Use of discovery algorithms to extract relations and dependencies—mostly causal dependencies—from the processed input in (2).[32]

(6) Cognitive/world model storage, composed of static world models (in case of models of static scenarios such as images, or descriptions of a certain static fictional-realistic state of affairs), or a timeline series of world models, in which frame-like world models are connected together by a timeline, creating a description of the surroundings over time. World models, following (4), contain information and references to entities populating the relevant environment, their properties (e.g., their spatial properties, but also relevant facts about them or references to knowledge base entries, or dependencies-db entries containing relevant additional information), and their relations.

(7) Dependencies database (DB): contains all of the dependencies between entities populating the different world models. Dependencies represent the different relations between entities (e.g., some sort of spatial relations, causal relations, part-whole relations and more, depending on the domain), and are the result of discovery processes (5). Each entry in the dependencies DB contains information about the entities (e.g., a complex entity-id representing the entity and the world model it is contained in), the dependency type, and additional information according to the dependency type (e.g., a reference to a previous/subsequent chain in a causal chain).

(8) World model dependencies associations: stands for the relation between a world model and a dependency. Each world model has zero or more (0-*) dependencies that are related to it. Each dependency has at least one (i.e., the one it was discovered in, more if it lasts over time in different frames in the same time series of a world) world model it is related to.

(9) The Reasoner model draws its inspiration from systems like CYC,[33] and neuro-symbolic architectures that apply reasoning, e.g., for question answering.[34] The general idea is to transform the problem, the concept or the question, given in a natural-language form, into an abstract form that captures its conceptual meaning; the system can then reason about this form using its knowledge and common sense bases, as well as environmental data, and produce a meaningful and explainable result. The Reasoner can apply deductive, inductive and abductive reasoning, which means that it can reason effectively (i.e., reach an hypothesis or a theory that best explains the available data,[35] much like humans do) with partial and uncertain information.

(10) Retrieve or update knowledge [see (11)] related to reasoning processes (retrieval) or text, images, video, audio data (update). In addition, one of the ways to develop and increase the knowledge contained in the system's knowledge base (11) is to add the products of reasoning processes to it; this is one of the ways by which the system can learn from experience and reuse the past conclusions of reasoning processes.

(11) Knowledge base: contains general concepts, rules and domain-specific extensions of both. For example, concepts can relate to time (hour, day, night), space (geography, spatial relations), emotions, culture, history and much more. Rules involve these concepts, e.g., that most people sleep at night, for several hours at a time, lying down; that no two objects can occupy the same space at the same time; that causes precede or start at same time as their effects. Another important aspect of knowledge bases is the contextual validity of their assertions,[36] i.e., the fact that assertions

[31] Henceforth, we will use cognitive and world models interchangeably.

[32] See for example Scholkopf et al. (2021) and Bengio et al. (2020).

[33] See CYC technology overview, https://www.cyc.com/wp-content/uploads/2019/09/Cyc-Technology-Overview.pdf, Accessed 29-May-2022.

[34] See for example the research summary from IBM: https://research.ibm.com/blog/ai-neurosymbolic-common-sense, Accessed 25-May-2022.

[35] See Vogel (1998).

[36] See CYC technology overview (fn. #28), Sect. 5.6.

can be true at one time, but not at another; in these circumstances but not in those; in one culture but not in another. This is crucial for handling contradictions and the non-universality of most convictions.

(12) Application Programming Interface (API) for externally inserting new concepts or rules to the knowledge base.

(13) Retrieval of environmental data, i.e., data related to the surroundings, the entities populating it and their mutual dependencies, which can be of help during reasoning processes. For example, to infer the best explanation, the Reasoner Model must employ all available data, however partial and uncertain, which are stored either in the knowledge base or as part of the Environmental data.

(14) The output of the Reasoner model can take several forms, depending on the context. For example, it can be employed as a filtering mechanism to filter out certain instances of generated text suggested by the system, based on common sense and knowledge reasoning; it can provide a step-by-step understandable logical argument in answer to a query; it can provide an answer to a question that takes into account the physical surroundings, and more. Similar (but in an opposite way) to the process described in (1), where a Coordinator sub-module takes the output of the learning modules and transforms it into a form that can be passed to the other sub-modules of the system, here, the Coordinator sub-module receives the output of the Reasoner and transforms it into a human understandable form that can be presented to the end-users of the system.

5 Concluding remarks

As previously mentioned, the main purpose of this paper is to implement artificial understanding by integrating our chosen theoretical–philosophical conception of understanding with the high-level requirements and engineering design of a robust AI system. In Sect. 2, we outline a notion of understanding as the knowledge of relations and dependencies. Then, we follow the Integrative approach, which aims to achieve the capabilities needed for AI systems to reach the next step by integrating deep learning with symbolic AI. In this final section, we examine how the high-level architecture presented in the previous section puts into practice the following theoretical–philosophical conception of understanding: knowledge of dependencies and relations between phenomena and entities relevant to the thing to be understood, and the ability for modal representation of these relations in a way that allows the understander to manipulate them and examine other related possibilities; or in other words, counterfactually reason about them.

How does the architecture outlined in the previous section corresponds to this conception of understanding? First, it allows the understander (i.e., the system) to keep track of the dependencies in its surroundings. This is accomplished by providing a data preserving structure that holds the dependence relations between entities discovered by the system. The language, audio and visual modules are in constant interaction with the environment; they provide a constant flow of information processed into a pre-defined abstract representational form; this input is then used by the Environment Model to construct world model frames that describe the surroundings at a certain instant of time on the one hand, and to run discovery algorithms to extract dependencies between entities on the other hand. Discovery algorithms can also make use of the Reasoner Model to retrieve and employ knowledge and common sense to identify and catalogue correlations in the input as (causal) dependencies. The world models containing the entities and the dependencies are stored and associations between them are created, as described in the previous section. Second, the Reasoner Model can make use of the Environment Model to construct additional world models (based on the ones discovered thus far) and associate different dependencies with them, so it can then reason counterfactually about them. In other words, it allows the Reasoner to tweak duplicates of existing world models. This enables the Reasoner to manipulate inner world models and their associated dependencies, to examine counterfactual courses of action and their results, and thus, reach better informed decisions.

As can be seen from the above description, the outlined architecture integrates learning models with symbolic models. It captures the essence of understanding as 'knowledge of dependencies' by discovering and storing dependencies between identified entities in its surroundings, and by enabling the different sub-modules of the system to interact in a way that allows the creation and manipulation of world models and their associated inter-relations similar to the one identified by the system; thus, enabling the system to 'imagine' what could have been if similar circumstances took place. This kind of architecture meets the requirements of the theoretical–philosophical conception of understanding described in Sect. 2, by enabling a robust representation of the dependence relations, the creation of world models which allow manipulation and tweaking of those relations, so it is able to observe their impact and anticipate what would happen if things were relevantly different. To follow our final statement of Sect. 2, in the proposed architecture, our agent A can have a cognitive model R of the object O; it can create a cognitive model R' of O, manipulate it and make inferences pertaining to O, based on these manipulations; thus, agent A understands O.

It is not my intention to argue that this is the only alternative for the implementation of artificial understanding, or even the best of all alternatives. The debate portrayed in previous sections is not settled and it may be the case that similar capabilities may be implemented by adopting other approaches.[37] However, at this point in time, the Integrative approach seems more promising and following it to implement understanding in artificial systems will surly yield more robust AI systems.

Data availability Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

Declarations

Conflict of interest On behalf of all the authors, the corresponding author states that there is no conflict of interest.

References

Allen C, Wallach W (2011) Moral machines: contradiction in terms, or abdication of human responsibility? In: Lin P, Abney K, Bekey G (eds) Robot ethics: the ethical and social implications of robotics. MIT Press, Cambridge, pp 55–68

Bengio Y et al. (2020) "A meta-transfer objective for learning to disentangle causal mechanisms." ArXiv abs/1901.10912

Chitnis R et al. (2021). "Learning neuro-symbolic relational transition models for bilevel planning." ArXiv abs/2105.14074

Chowdhery et al. (2022). "PaLM: Scaling Language Modeling with Pathways". https://doi.org/10.48550/arXiv.2204.02311

Dean J (2021) "Introducing pathways: a next-generation AI architecture". Google's Blog, https://blog.google/technology/ai/introducing-pathways-next-generation-ai-architecture/, Accessed 17 May 2022

Firt E (2020) The missing G. Ai&society 35:995–1007

Floridi L, Chiriatti M (2020) GPT-3: its nature, scope, limits, and consequences. Mind Mach 30:681–694. https://doi.org/10.1007/s11023-020-09548-1

Greco J (2010) Achieving knowledge. Cambridge University Press, New York

Grimm S (2011) Understanding. In: Berneker S, Pritchard D (eds) The Routledge companion to epistemology. Routledge, New York

Grimm SR (2014) Understanding as knowledge of causes. In: Fairweather A (ed) Virtue epistemology naturalized. Springer International Publishing, pp 329–345

Kaplan J et al. (2020) "Scaling laws for neural language models". arXiv 2001.08361

Kim J (2010) Explanatory knowledge and metaphysical dependence. Essays in the metaphysics of mind. Oxford University Press, New York

Mao J et al. (2019) "The neuro-symbolic concept learner: interpreting scenes words and sentences from natural supervision." ArXiv abs/1904.12584

Marcus G (2020) "The next decade in AI: four steps towards robust artificial intelligence". https://arxiv.org/abs/2002.06177

Marcus G (2022) "Deep learning is hitting a wall". https://nautil.us/deep-learning-is-hitting-a-wall-14467/. Accessed on 13 May 2022

Marcus, G. and Davis, E. (2019). Rebooting AI: Building Artificial Intelligence We Can Trust. Vintage Books.

Marcus G, Davis E (2020b) "GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about". MIT technology review. https://www.technologyreview.com/2020b/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/, Accessed 11-May-2022

Oltramari A et al. (2020) "Neuro-symbolic architectures for context understanding." Knowledge graphs for explainable artificial intelligence. Neuro-symbolic Architectures for Context Understanding. https://doi.org/10.48550/arXiv.2003.04707

Pearl J, Halpern JY (2001) "Causes and explanations: a structural-model approach—part ii: explanations". In Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI), San Francisco, CA: Morgan Kaufmann

Rae JW et al. (2022) "Scaling language models: Methods, analysis and insights from training Gopher". arXiv 2112.11446

Raedt De L et al. (2020) "From statistical relational to neuro-symbolic artificial intelligence." ArXiv abs/2003.08316

Reed S et al. (2022) "A Generalist Agent." ArXiv abs/2205.06175

Riggs W (2003) Understanding 'virtue' and the virtue of understanding. In: DePaul M, Zagzebski L (eds) Intellectual virtue: perspectives from ethics and epistemology. Oxford University Press, New York, pp 203–226

Scholkopf B et al (2021) Toward causal representation learning. Proc IEEE 109:612–634

Sutton R (2019) "The Bitter Lesson". http://incompleteideas.net/IncIdeas/BitterLesson.html. Accessed 07 May 2022

Sychev O (2021) Combining neural networks and symbolic inference in a hybrid cognitive architecture. Procedia Comput Sci 190:728–734

Thoppilan R et al. (2022) "LaMDA: Language models for dialog applications". arXiv 2201.08239

Thorisson KR, Kremelberg D (2017) Do machines understand? A short review of understanding & common sense in Artificial Intelligence. (AGI 2017 conference). http://alumni.media.mit.edu/~kris/ftp/AGI17-UUW-DoMachinesUnderstand.pdf. Accessed 07 May 2022

Vogel J (1998) "Inference to the best explanation". In The Routledge Encyclopedia of Philosophy. Taylor and Francis. Retrieved 4 Dec 2022, from https://www.rep.routledge.com/articles/thematic/inference-to-the-best-explanation/v-1. doi: https://doi.org/10.4324/9780415249126-P025-1

Weidinger, L., et al. (2021). "Ethical and social risks of harm from Language Models." ArXiv abs/2112.04359.

Wilkenfeld D (2013) Understanding as representation manipulability. Synthese 190(6):997–1016

Woodward J (2003) Making things happen: a theory of causal explanation. Oxford University Press, New York

[37] An additional prominent approach is championed by Yoshua Bengio, who is considered one of the founding fathers of our current-days deep learning systems. To reiterate what was already mentioned at the beginning of Section IV, both the approaches mentioned here agree on the capabilities that current systems lack. However, each approach suggests a different path to implementation. Bengio believes that we should keep the current framework of deep learning, but "build in some of the functional advantages of classical AI rule-based symbolic manipulation in neural nets, but in implicit way." (~46:30, from the AI Debate between Gary Marcus and Yoshua Bengio, https://www.youtube.com/watch?v=EeqwFjqFvJA, Accessed 04-Dec-2022). For Bengio's opinion, see also his keynote lecture "Deep Learning Cognition" in AI in 2020 and Beyond.

Zagzebski L (2001) Recovering understanding. In: Steup M (ed) Knowledge, truth, and duty: essays on epistemic justification, responsibility, and virtue. Oxford University Press, New York

Zagzebski L (2009) On epistemology. Wadsworth, Belmont

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.