Symbolic and Statistical Theories of Cognition: Towards Integrated Artificial Intelligence[⋆]

Yoshihiro Maruyama

Research School of Computer Science
College of Engineering and Computer Science
The Australian National University
yoshihiro.maruyama@anu.edu.au

Abstract. There are two types of approaches to Artificial Intelligence, namely Symbolic AI and Statistical AI. The symbolic and statistical paradigms of cognition may be considered to be in conflict with each other; the recent debate between Chomsky and Norvig exemplifies a fundamental tension between the two paradigms, which is arguably in parallel with a conflict on interpretations of quantum theory as seen between Bohr and Einstein, one side arguing for the probabilist or empiricist view and the other for the universalist or rationalist view. In the present paper we explicate and articulate the fundamental discrepancy between them, and explore how a unifying theory could be developed to integrate them, and what sort of cognitive rôles Integrated AI could play in comparison with present-day AI. We give, inter alia, a classification of Integrated AI, and argue that Integrated AI serves the purpose of humanising AI in terms of making AI more verifiable, more explainable, more causally accountable, more ethical, and thus closer to general intelligence. We especially emphasise the ethical advantage of Integrated AI. We also briefly touch upon the Turing Test for Ethical AI, and the pluralistic nature of Turing-type Tests for Integrated AI. Overall, we believe that the integrated approach to cognition gives the key to the next generation paradigm for AI and Cognitive Science in general.

Keywords: Symbolic AI; Statistical AI; Integrated AI; Categorical AI; Classification of Integrated AI; Five Features of Integrated AI; Ethical Turing Test; Integrated Turing Test; Chomsky vs. Norvig Debate; Bohr vs. Einstein Debate

1 Introduction: MIT's AI Lab, Now and Then

Neil Thompson at MIT and his collaborators recently published an intriguing article entitled "The Computational Limits of Deep Learning" [38], arguing in the following way:

[⋆] This is still a preliminary pre-proceedings version, and not the final version; the Springer LNCS proceedings version published after the conference shall be slightly different from, and in particular slightly longer than, the current version.

> [P]rogress along current lines is rapidly becoming [...] unsustainable. Thus, continued progress [...] will require dramatically more computationally-efficient methods, which will either have to come from changes to deep learning or from moving to other machine learning methods.

There are many other problems in statistical machine learning, such as explainability and ethical issues as we shall discuss below. How could we overcome them? What sort of changes would be necessary for the next generation of Artificial Intelligence (and Cognitive Science in general)? A possible approach to overcome the limitations of statistical machine learning would be the integration of Symbolic and Statistical AI; at least some part of what Statistical AI is bad at is what Symbolic AI is good at. Deductive reasoning and inductive learning would be arguably the two fundamental wheels of the human mind (even though there may possibly be yet another wheel of human cognition).

One of the earliest ideas of Integrated AI comes from Marvin Minsky, the 1969 Turing Award winner and co-founder of MIT's AI Lab, who proposes the integration of Symbolic and Connectionist AI in particular (aka. Logical and Analogical AI, or Neat and Scruffy AI) as a form of Integrated AI in his 1991 article [32]:

> Our purely numerical connectionist networks are inherently deficient in abilities to reason well; our purely symbolic logical systems are inherently deficient in abilities to represent the all-important "heuristic connections" between things — the uncertain, approximate, and analogical linkages that we need for making new hypotheses. The versatility that we need can be found only in larger-scale architectures that can exploit and manage the advantages of several types of representations at the same time. Then, each can be used to overcome the deficiencies of the others.

In light of this, the Minsky's conceptions of Symbolic (Logical) and Connectionist (Analogical) AI may be compared with the Reichenbach's well-known (yet debated) conceptions of the context of justification ("reason well") and the context of discovery ("making new hypotheses"). From the philosophy of science point of view, the cognitive capacities of discovery and justification are arguably the conditions of possibility of science as a human intellectual enterprise, which would make it compelling to combine the two paradigms of AI. Yet at the same time, there is a fundamental tension between the two paradigms, as exemplified by the Chomsky versus Norvig debate (Peter Norvig is Google's research director; Chomsky is one of the founders of Cognitive Science as well as the father of modern linguistics), which is arguably in parallel with the Bohr versus Einstein debate on the nature of quantum reality, as we shall see below.

In the following, we revisit the Chomsky-Norvig (and Bohr-Einstein) debate(s) to elucidate the discrepancy between the two paradigms, and place it in a broader context of science and philosophy (Section 2). And we discuss how the two paradigms could be integrated and why that matters at all, especially from an ethical point of view; besides these, we discuss Turing-type tests for

Ethical AI and Integrated AI (Section 3). We finally conclude with outlooks for the future of Artificial Intelligence and Cognitive Science (Section 4).

2 The Fundamental Tension between Symbolic and Statistical Paradigms of Cognition

In everyday life, we use both logical reasoning and statistical inference to make various judgments; deduction and induction are indispensable part of everyday life as well as scientific investigation. At the same time, we cannot precisely tell which part of human cognition is essentially symbolic, and which part of it is essentially statistical. It could, for example, happen that all functions of the human mind can be simulated by logical means or by statistical means alone; if this sort of reduction is possible, the apparent dualism of logic and statistics may collapse. For instance, automated theorem proving has been advanced within the symbolic paradigm of AI, but there is now some evidence that it can be done more efficiently within the statistical paradigm, especially with the help of deep learning (see, e.g., [2, 24, 35]). AI research today seems to make it compelling to reconsider the relationships between deductive reasoning and inductive learning. In this section, we focus upon a fundamental tension between the two paradigms, which manifests in the present landscape of AI as well as the history of science and philosophy as we shall discuss in the following. Let us begin with the Maxwell's interesting view of Nature.

2.1 Maxwell's Dualistic View of Nature and Humans as its Observers

In a 1850 letter to Lewis Campbell, Maxwell [30] asserts as follows:

> [T]he true Logic for this world is the Calculus of Probabilities.

Maxwell is known for his great contributions to electromagnetism and statistical mechanics. Philosophically, he was seemingly influenced by the British tradition of empiricism, which puts a strong emphasis on the contingent nature of reality. His empiricist tendency may be observed in the following passage as well [30]:

> [A]s human knowledge comes by the senses in such a way that the existence of things external is only inferred from the harmonious (not similar) testimony of the different senses, understanding, acting by the laws of right reason, will assign to different truths (or facts, or testimonies, or what shall I call them) different degrees of probability.

Yet Maxwell was not a naïve empiricist. His statistical mechanics is in harmony with the empiricist thought; however, his theory of electromagnetism, which is arguably his greatest contribution to science, is rather closer to the rationalist thought in the continental tradition that sticks to the absolute, universal nature of truth, or the mechanistic view of Nature as shared by Newton and Laplace. Indeed, Maxwell [30] argues as follows:

> [O]ur experiments can never give us anything more than statistical information [...] But when we pass from the contemplation of our experiments to that of the molecules themselves, we leave a world of chance and change, and enter a region where everything is certain and immutable.

The "molecules themselves" in Maxwell's thought may be compared with the Kant's idea of things themselves. So, whilst asserting that the true logic of the world is the calculus of probabilities, he maintains the universal conception of truth as being immune to chance and change. Experiments only allow us to access statistical information, but he thought there is something beyond that, namely some Platonistic realm of absolute truths. The probabilist view somehow coexisted with the universalist view in his thought. The Maxwell equations in his theory of electromagnetism embody the latter, whilst his statistical mechanics the former.

2.2 The Chomsky versus Norvig Debate on the Nature of Science and Cognition

There was already some tension between the probabilist and universalist views at the time of Maxwell; it culminates in the contemporary debate between Noam Chomsky, who defends the universalist position, and Peter Norvig, who defends the probabilist position. Gold [20] recapitulates their debate in the following manner:

> Recently, Peter Norvig, Google's Director of Research and co-author of the most popular artificial intelligence textbook in the world, wrote a webpage extensively criticizing Noam Chomsky, arguably the most influential linguist in the world. Their disagreement points to a revolution in artificial intelligence that, like many revolutions, threatens to destroy as much as it improves. Chomsky, one of the old guard, wishes for an elegant theory of intelligence and language that looks past human fallibility to try to see simple structure underneath. Norvig, meanwhile, represents the new philosophy: truth by statistics, and simplicity be damned.

Norvig basically takes the empiricist position, emphasising the "world of chance and change" (in terms of Maxwell) and thus the necessity of statistical analysis; Chomsky, by contrast, takes the rationalist position, emphasising the universal nature of linguistic structure and of scientific laws in general. To clarify Chomsky's view, Katz [23] interviewed Chomsky at MIT; Chomsky criticises Statistical AI, or Statistical Cognitive Science in general, in the following manner:

> [I]f you get more and more data, and better and better statistics, you can get a better and better approximation to some immense corpus of text [...] but you learn nothing about the language.

Chomsky does not deny the success of statistical methods in prediction and other tasks; rather, he is concerned with the nature of scientific understanding, his point being akin to the recent issue of explainability in machine learning, especially deep learning. Chomsky even argues, in the same interview, that statistical analysis allows us to "eliminate the physics department" in his extreme Gedanken experiment:

> [I]t's very different from what's done in the sciences. So for example, take an extreme case, suppose that somebody says he wants to eliminate the physics department and do it the right way. The "right" way is to take endless numbers of videotapes of what's happening outside the video, and feed them into the biggest and fastest computer, gigabytes of data, and do complex statistical analysis [...] you'll get some kind of prediction about what's gonna happen outside the window next. In fact, you get a much better prediction than the physics department will ever give.

Chomsky is not just concerned with cognition and language, but also with the nature of science in general. Chomsky argues that there are two conceptions of science, the universalist one aiming at the scientific understanding of Nature (and Cognition) and the probabilist one aiming at the engineering approximation of data; he says as follows in the same interview [23].

> These are just two different concepts of science. The second one is what science has been since Galileo, that's modern science. The approximating unanalyzed data kind is sort of a new approach, not totally, there's things like it in the past. It's basically a new approach that has been accelerated by the existence of massive memories, very rapid processing, which enables you to do things like this that you couldn't have done by hand. But I think, myself, that it is leading subjects like computational cognitive science into a direction of maybe some practical applicability.

To Chomsky, it is a wrong direction to go, especially from a scientific, rather than engineering, point of view. Chomsky argues that statistical analysis is just "butterfly collecting"; Norvig [34] himself succinctly recapitulates Chomsky's points as follows:

> Statistical language models have had engineering success, but that is irrelevant to science [...] Accurately modeling linguistic facts is just butterfly collecting; what matters in science (and specifically linguistics) is the underlying principles [...] Statistical models are incomprehensible; they provide no insight.

To Chomsky, data science is engineering rather than science; science must confer understanding. As the above passage clearly shows, Norvig actually understood Chomsky's points very well, and still strongly disagreed. Norvig [34] argues for the necessity of statistical analysis in the science of language on the grounds of the contingent nature of language per se:

> [L]anguages are complex, random, contingent biological processes that are subject to the whims of evolution and cultural change. What constitutes a language is not an eternal ideal form, represented by the settings

> of a small number of parameters, but rather is the contingent outcome of complex processes. Since they are contingent, it seems they can only be analyzed with probabilistic models.

The Chomsky versus Norvig debate may be compared with the Bohr versus Einstein debate on the ultimate nature of quantum reality, especially the issue of the EPR (Einstein-Podolsky-Rosen) paradox and non-local correlations [7]. The Chomskyan linguistics aims at explicating the eternal ideal form of language, and Chomsky is very much like Einstein, who believed that probabilities arise in quantum mechanics because the formulation of quantum mechanics is still incomplete, i.e., there are some hidden variables ("small number of parameters") to make it a deterministic theory like classical mechanics. The universalist's strongest possible presupposition is that there are always universal (or deterministic) principles underlying apparently complex (or probabilistic) phenomena. From the universalist perspective, statistics is more like a compromise than an ultimate solution to the problem of understanding Nature. On the other hand, Norvig argues that the irreducible complexity of natural language and its evolution makes it compelling to use probabilistic models, just as Bohr argued for the necessity of probabilities in quantum mechanics and so for the completeness of it. To the probabilist, there is nothing lurking behind statistics; it can simply happen that certain phenomena in Nature are inherently probabilistic. That is to say, there is just the surface without any depths underlying it (incidentally, such an idea has been discussed in twentieth century continental philosophy as well); certain probabilistic theories are already complete.

It is a common view that Bohr won the debate with Einstein, whose understanding of quantum theory was proven to be misconceived by the celebrated Bell theorem [4], even though there are some non-local deterministic formulations of quantum theory, such as Bohmian mechanics, to which the assumptions of the Bell theorem do not apply and which thus partially realise Einstein's dream [6]. Bell-type theorems in physics are called No-Go theorems because they mathematically refute certain forms of classical realism, which, therefore, is a wrong direction to go. If there are similar theorems in AI, Norvig could mathematically refute Chomsky; however, there is no such theorem known at the moment. And in the case of AI in particular, there is some hope for reconciling the two camps as we shall discuss in the next section, before which we briefly touch upon the tension between the two paradigms in the context of natural language semantics in particular.

2.3 A Manifestation of the Fundamental Tension in Natural Language Semantics

The success of Natural Language Processing in the statistical paradigm is mostly due to the so-called Vector Space Model (VSM) of Meaning. Turney and Pantel [39] indeed argue as follows:

> The success of the VSM for information retrieval has inspired researchers to extend the VSM to other semantic tasks in natural language process-

> ing, with impressive results. For instance, Rapp (2003) used a vector-based representation of word meaning to achieve a score of 92.5% on multiple-choice synonym questions from the Test of English as a Foreign Language (TOEFL), whereas the average human score was 64.5%.

The Vector Space Model of Meaning is statistical semantics of natural language, and based upon what is called the Distributional Hypothesis [39]: "words in similar contexts have similar meanings." This is some sort of semantic contextualism, and semantic contextualism is a form of holism about meaning, since the meaning of a word is determined with reference to a larger whole, namely contexts, without which meaning cannot be determined. In the Vector Space Model of Meaning, for instance, meaning vectors are derived on the basis of a large amount of linguistic contexts, without which meaning vectors cannot be determined.

In contrast to this statistical semantics, which builds upon contextualism, there is another paradigm of natural language semantics, namely symbolic semantics, which builds upon compositionalism, the view that the meaning of a whole is determined with reference to the meaning of its parts. In contextualism, the meaning of a part is only determined with reference to a larger whole, and thus compositionalism is in sharp contrast with contextualism. The tension between Chomsky and Norvig in the narrow context of linguistic analysis may be understood as rooted in this conflict between compositional and contextual semantics. The compositionality camp includes Montague as well; he expresses an opinion sympathetic with Chomsky as follows [33]:

> There is in my opinion no important theoretical difference between natural languages and the artificial languages of logicians; indeed, I consider it possible to comprehend the syntax and semantics of both kinds of languages within a single natural and mathematically precise theory. On this point I differ from a number of philosophers, but agree, I believe, with Chomsky and his associates.

Both the principle of compositionality and the principle of contextuality have their origins in Frege's philosophy of language. It is puzzling why Frege endorsed both of them, especially in light of the above view that there is a fundamental conflict between compositionality and contextuality. Michael Dummett, a well known commentator on Frege, was clearly aware of this, pointing out a "difficulty which faces most readers of Frege" [16]:

> It was meant to epitomize the way I hoped to reconcile that principle, taken as one relating to sense, with the thesis that the sense of a sentence is built up out of the senses of the words. This is a difficulty which faces most readers of Frege [...] The thesis that a thought is compounded out of parts comes into apparent conflict [...] with the context principle [...]

According to more recent commentators on Frege, it is actually not so obvious whether Frege really endorsed any of them. Pelletier [36], for example, concludes that Frege endorsed neither of them; Janssen [22] argues that Frege only endorsed the principle of contextuality.

Compositionality is essential in the so-called productivity of language; thanks to the compositional character of language, we can compose and comprehend entirely new sentences. Frege [17] was already aware of this connection between compositionality and productivity:

> It is astonishing what language can do. With a few syllables it can express an incalculable number of thoughts, so that even a thought grasped by a terrestrial being for the very first time can be put into a form of words which will be understood by someone to whom the thought is entirely new.

Compositionality allows us to understand some other striking characteristics of natural language. Davidson [13], for example, points out that compositionality is essential in the learnability of language:

> It is conceded by most philosophers of language, and recently by some linguists, that a satisfactory theory of meaning must give an account of how the meanings of sentences depend on the meanings of words. Unless such an account could be supplied for a particular language, it is argued, there would be no explaining the fact that we can learn the language: no explaining the fact that, on mastering a finite vocabulary and a finitely stated set of rules, we are prepared to produce and to understand any of a potential infinitude of sentences. I do not dispute these vague claims, in which I sense more than a kernel of truth.

Yet these do not necessarily imply that the principle of contextuality or statistical semantics based on it cannot account for those properties of natural language (see, e.g., [26], which also explains the tension between compositionality and contextuality in more detail; for contextuality across physics and cognitive science, see [27, 28]).

Statistical semantics in Natural Language Processing has been highly successful in various domains of application and actually implemented in a variety of real-world systems. It is however known to suffer from lack of structure; it mostly ignores the inherent structure of language such as grammar. Turney-Pantel [39], for example, argue in the following manner:

> Most of the criticism stems from the fact that term-document and word-context matrices typically ignore word order. In LSA, for instance, a phrase is commonly represented by the sum of the vectors for the individual words in the phrase; hence the phrases house boat and boat house will be represented by the same vector, although they have different meanings.

The same criticism applies to what is generally called the bag-of-words model in information retrieval [39]. It is truly amazing that statistical semantics has been so successful whilst ignoring the structure of language mostly. The intrinsic structure of language is what Chomsky has investigated for a long time. So Gold [20] says that modern AI technologies would make "Chomskyan linguists cry":

> Norvig is now arguing for an extreme pendulum swing in the other direction, one which is in some ways simpler, and in others, ridiculously more complex. Current speech recognition, machine translation, and other modern AI technologies typically use a model of language that would make Chomskyan linguists cry: for any sequence of words, there is some probability that it will occur in the English language, which we can measure by counting how often its parts appear on the internet. Forget nouns and verbs, rules of conjugation, and so on: deep parsing and logic are the failed techs of yesteryear.

Yet at the same time, substantial improvement in computational efficiency and other respects has been achieved recently with the integration of symbolic and statistical methods in Natural Language Processing (see, e.g., [21]). And there is now some movement to integrate the two paradigms in linguistic and other contexts (for integrations in Natural Language Processing, see, e.g., [3, 12, 21]). The integration of Symbolic and Statistical AI works beyond Natural Language Processing, allowing for different advantages, and this is what we are going to address in the following section (for more detailed discussions on linguistic issues concerning Symbolic and Statistical AI, we refer to [26]).

3 Towards Integrated Artificial Intelligence and Integrated Cognitive Science

Symbolic AI is good at principled judgements, such as logical reasoning and rule-based diagnoses, whereas Statistical AI is good at intuitive judgements, such as pattern recognition and object classification. The former would amount to what is called the faculty of reason and understanding, and the latter to the faculty of sensibility in terms of the Kantian epistemology or philosophy of mind. McLear [31] explains these fundamental faculties of human cognition in the following manner:

> Kant distinguishes the three fundamental mental faculties from one another in two ways. First, he construes sensibility as the specific manner in which human beings, as well as other animals, are receptive. This is in contrast with the faculties of understanding and reason, which are forms of human, or all rational beings, spontaneity.

So, from this Kantian point of view, animal cognition, as well as human cognition, is equipped with the faculty of sensibility to recognise the world, and yet the faculty of reason and understanding is a striking characteristic of human cognition only. If this view is correct, Statistical AI may not be sufficient for realising human-level (or super-human) machine intelligence. The Kantian philosophy of mind suggests that both Symbolic and Statistical AI are indispensable for human-level artificial intelligence. If so, it would be essential for the next generation of AI to overcome the symbolic-statistical divide and integrate the two paradigms of cognition.

3.1 The Integrated Paradigm: A Classification of Integrated AI

Let us discuss Integrated AI in more detail in the following. We give a classification of three levels of Integrated AI, and propose Turing-type tests for it. In addition we argue, inter alia, that Integrated AI is a promising approach to Ethical AI or Just AI.

Let us begin with some history of AI. Broadly speaking, historical developments of Artificial Intelligence may be summarised as follows [10]:

- First-generation AI: Search-based Primitive AI.
- Second-generation AI: Deductive Rule-based Symbolic AI (aka. GOFAI, i.e., Good Old-Fashioned Artificial Intelligence).
  - Examples of Symbolic AI: expert systems based on production rules; automated reasoning and planning; theorem provers and verification; and so fourth.
- Third-generation AI (present-day AI): Inductive Learning-based Statistical AI (with successful applications in industry today).
  - Examples of Statistical AI: neural networks and deep learning; support vector machines and kernel methods; Bayesian networks and their variants such as Markov networks; and so fourth.

The next generation AI, then, might be:

- Fourth-generation AI (in the coming future): Towards Integrated AI, namely the integration of Symbolic and Statistical AI.

[10] presents a similar perspective on future developments of AI.

In order to explicate different ways to conceive Integrated AI, let us now give a conceptual classification of Integrated AI in terms of three levels of integration (i.e., task-oriented integration, modular mechanism integration, and seamless mechanism integration) in the following manner:

1. Task-oriented integration: integration at the level of each concrete problem solving, namely integration made for (or dependent upon) a given particular task, which is thus applicable to the specific type of problems only.
   - Examples of task-oriented integration: Statistical Theorem Proving, which generates candidate proofs via statistical methods, and then verify their correctness via symbolic methods [2, 24, 35]; Safe Learning, which combines deductive reachability analysis with statistical machine learning in order to determine safe regions for safety critical systems to operate [1, 40].
2. Modular mechanism integration: integration at the level of modular mechanisms, namely integration with symbolic and statistical components modularly separated to each other.

  - Examples of modular mechanism integration: the compositional distributional model of natural language processing, which derived word vectors via statistical methods based upon what is called the distributional hypothesis ("words in similar contexts have similar meanings"), and compose sentence vectors from word vectors via symbolic methods based upon the logical theory of formal grammar such as Lambek's pregroups [12, 21].
3. Seamless mechanism integration: integration at the level of integrated mechanism, namely integration as a single mechanism unifying symbolic and statistical approaches to cognition.
   - Examples of seamless mechanism integration: Markov logic network, which is a general framework to combine first-order logic and Markov networks, "attaching weights to first-order formulas and viewing them as templates for features of Markov networks" [14] (see also [15]); neural-symbolic computing [5, 19], which is one of the oldest approaches to the integrated paradigm of Artificial Intelligence, and aims to integrate statistical connectionism and symbolic representationism within a general framework for learning and reasoning.

Task-independent methods are desirable in order to develop Integrated AI for different purposes in a systematic manner. Task-oriented integrations only work for specific tasks, but both modular and seamless mechanism integrations can work for more general purposes, just as the compositional distributional model of natural language processing mentioned above works for a broad variety of linguistic tasks. Note that there is no implication like seamless integrations are generally better than modular integrations. Modular integrations could be more useful than seamless ones, for example, on the ground that results in each paradigm can be transferred and applied directly. Note also that machine learning frameworks are usually task-independent, even though each problem solving algorithm is made in a task-dependent manner; this means that the mathematical essence of learning is independent of the nature of each concrete task, and that is why machine learning is a theory of learning.

3.2 Five Features of Integrated AI: Making AI More Verifiable, Explainable, Accountable, Ethical, and thus More Human

Integrated AI is not just for improvement of computational performance; it is expected to resolve difficulties in Statistical AI via the methods of Symbolic AI. Desirable characteristics of Integrated AI would be as follows (cf. [29]):

1. Verifiability: Integrated AI should allow us to verify the results of its problem solving such as prediction and classification (which Statistical AI is good at, whereas Symbolic AI is good at verification).
2. Explainability: Integrated AI should allow us to explain the results of its problem solving, e.g., the reason why they have obtained rather than others; explainability is seemingly one of the strongest concerns in recent AI research.

3. Causality: Integrated AI should allow us to account for causal relationships as well as correlational ones; this is particularly important in data science, which must be able to account for causal laws if it aims to qualify as proper science on its own.
4. Unbiasedness: Integrated AI should allow us to make unbiased judgements or to correct their biases learned from biased real-world data; this 'debiasing' function shall be discussed below in more detail.
5. Generality: Integrated AI would allow for AGI, namely Artificial General Intelligence; this may be too strong a requirement, yet developing general intelligence would be one of the ultimate purposes of AI research.

Caliskan et al. show in their recent *Science* article [9] that:

> Semantics derived automatically from language corpora contain human-like biases.

Machine learning, or data-driven science enabled with it, is descriptive in the sense that it basically learns anything in data, regardless of whether it is good or bad. It is like a very obedient child, who may mimic some bad behaviour of parents or teachers without considering whether it is good or bad. By contrast, Integrated AI can be normative as well as descriptive; it can, for example, be equipped with top-down rules or norms to prevent bias learning and to make AI more ethical. This may count as a striking feature of Integrated AI, especially from the perspective of AI ethics.

There might be no means for purely statistical AI to prevent itself from learning biases from biased data; the better it approximates the given biased data, the better it learns those biases contained in it. This ironically suggests that those AI systems that are inferior in learning performance can actually be superior, in terms of unbiasedness, to those that are superior in learning performance (something analogous could happen in the human society as well). Put another way, there are things one should not learn from experience (i.e., empirical data) as well as things one should learn from it. And rational agents must be able to distinguish between them on the ground of some norms or rules, which can be incorporated via Symbolic AI. Integrated AI would thus be a right framework for Unbiased Ethical AI (aka. Just AI; see also [29]); this would be crucial in developments of AI for the Social Good, which has been sought after in the present, more and more AI-laden society.

Social implementation of AI systems would require them to be ethical; ethics, or ethical behaviour, may also be considered to be constituents of intelligence. But how could we judge whether AI is morally good or not? There could be something like the Turing Test to do that. For example, the Ethical Turing Test could be formulated in the following manner:

- The Ethical Turing Test (aka. Misleading Turing Test): we try to deceive AI with biased data or reasoning; still AI must be able to make correct judgments whilst being not deceived by us humans.

AI must be able to follow (or simulate) correct behaviour in the original Turing Test; in the above Ethical Turing Test, AI must be able to correct its behaviour, and so it may also be called the Dual Turing Test. The Dual Turing Test can be more difficult to pass than the original Turing Test, because correcting wrong answers is often more complex than giving correct answers (the author is familiar with this phenomenon in his experience of teaching logic to hundreds of students and correcting their mistakes every week during the term). The Ethical / Dual Turing Test requires AI to be resilient with respect to different biases, which do exist in real-world situations. Although Statistical AI has been shown to learn different biases from real-world data in an inductive, bottom-up manner, nevertheless, Integrated AI could pass the Ethical Turing Test with the help of Symbolic AI, which gives top-down rules and principles to make it immune to potential biases.

To test Integrated AI, we could rely upon other types of Turing-type tests as well, such as the Verification Turing Test, in which AI must be able to give both answers to questions and the verification of them. In general, the plurality of Turing-type tests would be essential; there may be no single experimental scheme to test different aspects of intelligence at once. If so, multiple tests are required to test different facets of intelligence. Human intelligence is so versatile that no single experiment allows for an adequate assessment of different aspects of it. So the plurality of Turing-type tests would be essential for conceiving the Turing Test for Integrated AI (for related Turing-type Tests, see also [29]).

There could be Chinese-Room-style counterarguments against these Turing-type Tests. Highly non-ethical AI could pass the Ethical Turing Test above just by simulating ethical behaviour in a superficial manner. Superintelligent AI, e.g., could not be deceived by us, bur rather could deceive us in many ways, whilst hiding its unethical nature from us. This means it could easily pass the Ethical Turing Test. Yet the same thing may happen in the human case as well. Just as there is no effective method to test the ethical nature of human beings, there would be no ultimate Turing Test for Ethical AI, neither (for related issues, see also [25, 29]).

4 Concluding Remarks: The Integrated Paradigm as a Transdisciplinary Trading Zone

We have discussed the fundamental tension between the symbolic and statistical paradigms of AI, and the possibility of integrating and unifying them, together with various advantages to do so, including the ethical one in particular. What is particularly interesting in the present landscape of AI is, in our opinion, that the debate between the symbolic and statistical camps look very much like the classic debate between the universalist and the probabilist views of Nature (including Cognition and Intelligence as part of it), and that the debate is directly relevant to urgent issues in AI, such as verifiability, explainability, causal accountability, and algorithmic biases, as we have discussed above.

AI and Machine Learning, therefore, would allow us not only to revive the classic debate between the universalist and the probabilist in the past, but also to place it in different novel contexts relevant to the present society. The central tenet of the present paper is that Integrated AI, if it could be developed in the right way, would serve the purpose of solving those urgent issues in AI. Yet at the same time, we would also contend that philosophical debates in the past (and present) could be useful inputs to the design and development of Integrated AI, since they are closely linked with the urgent issues in AI as we have discussed above. In light of these, Integrated AI could be a transdisciplinary place (or trading zone in Peter Galison's terms [18]) where different sorts of intellectual cultures are allowed to meet each other, as well as a theoretical foundation for the next generation AI technology (note that Galison also contrasts Image and Logic, which Statistical and Symbolic AI are about).

We have mainly focused upon AI rather than Cognitive Science in general; even so, our arguments would mostly apply to Cognitive Science as well as AI. There are, as a matter of course, symbolic and statistical paradigms in Cognitive Science just as well, and integrating them would be beneficial in many ways. From the AI point of view, the principal merits of the integrated paradigm would be developments of solutions to problems such as explainability and algorithmic biases. Yet from the Cognitive Science point of view, the principal advantages of the integrated paradigm would rather be the integrated understanding of fundamental faculties of the human mind, especially the faculty of reason and understanding on one hand, and the faculty of sensibility on the other. Integrated Cognitive Science could, hopefully, lead to something like a cognitive theory of everything (or a theory of every-cognition). A physical theory of everything is concerned with a unified understanding of general relativity and quantum theory, and a cognitive theory of everything with a unified understanding of the faculty of reason and understanding and the faculty of sensibility.

It is a highly non-trivial issue how to actually develop Integrated AI or what kind of mathematical framework allows us to lay down a theoretical foundation for the integrated paradigm of AI and Cognitive Science in the first place. We have touched upon different approaches to Integrated AI throughout the paper, some of which can be expressed in terms of category theory. We generally believe that category theory could give a principal methodology to integrate the two paradigms, as it has indeed played such unificatory rôles in the sciences, and succeeded in integrating different paradigms even across different sciences. In light of this, many approaches to Integrated AI could even be understood and unified under one umbrella, namely Categorical AI, hopefully.

Acknowledgements. The author would like to thank his colleagues, both past and present, for various inputs enabling this work; most recently, the Humanising Machine Intelligence project across computer science and philosophy at ANU, of which the author is an affiliated member, has affected some aspects and wordings of the present work. The author hereby acknowledges that this work was supported by JST PRESTO (grant code: JPMJPR17G9)

References

1. A. K. Akametalu, S. Kaynama, J. F. Fisac, M. N. Zeilinger, J. H. Gillula, C. J. Tomlin, Reachability-based safe learning with Gaussian processes, *Proceedings of CDC*, pp. 1424-1431, 2014.
2. K. Bansal, S. M. Loos, M. N. Rabe, C. Szegedy, and S. Wilcox, HOList: An Environment for Machine Learning of Higher Order Logic Theorem Proving, *Proceedings of ICML*, pp. 454-463, 2019.
3. M. Baroni et al, Frege in space: A program of compositional distributional semantics, *Linguistic Issues in Language Technology*, vol. 9, pp. 5-110, 2014.
4. J. S. Bell, *Speakable and Unspeakable in Quantum Mechanics: Collected Papers on Quantum Philosophy*, Cambridge University Press, 2004.
5. T. R. Besold et al., Neural-Symbolic Learning and Reasoning: A Survey and Interpretation, arXiv:1711.03902, 2017.
6. D. Bohm and B. Hiley, *The Undivided Universe: An Ontological Interpretation of Quantum Theory*, Routledge Chapman & Hall, 1993.
7. M. Born, *The Born Einstein Letters*, Walker and Company, 1971.
8. Z. Bouraoui et al., From Shallow to Deep Interactions Between Knowledge Representation, Reasoning and Machine Learning, arXiv:1912.06612.
9. Caliskan et al., Semantics derived automatically from language corpora contain human-like biases, *Science*, vol. 356, pp. 183-186, 2017.
10. CDRS, Research and Development on the Fourth Generation of AI, Strategic Proposal, CRDS-FY2019-SP-08, 2019. ,
11. N. Chomsky, Keynote Panel: The Golden Age – A Look at the Original Roots of Artificial Intelligence, Cognitive Science, and Neuroscience, *MIT Symposium on Brains, Minds, and Machines*, http://languagelog.ldc.upenn.edu/myl/PinkerChomskyMIT.html (retrieved on 23 June 2019).
12. B. Coecke et al., Mathematical foundations for a compositional distributional model of meaning, *Linguistic Analysis*, vol. 36, pp. 345-384, 2010.
13. D. Davidson, Truth and Meaning, *Synthese*, vol. 17. pp. 304-323, 1967.
14. P. Domingos, S. Kok, D. Lowd, H. Poon, M. Richardson, and P. Singla, Markov Logic, *Springer LNCS*, vol. 4911, pp. 92-117, 2008.
15. P. Domingos and D. Lowd, Unifying logical and statistical AI with Markov logic, *Communications of the ACM*, vol. 62, pp. 74-83, 2019.
16. M. Dummett, *The Interpretation of Frege's Philosophy*, London: Duckworth, 1981.
17. G. Frege, Compound thoughts, *Mind*, vol. 72, pp. 1-17, 1963 (originally 1923).
18. P. Galison, *Image & logic: A material culture of microphysics*, The University of Chicago Press, 1997.
19. A. Garcez, M. Gori, L. Lamb, L. Serafini, M. Spranger, and S. Tran, Neural-Symbolic Computing: An Effective Methodology for Principled Integration of Machine Learning and Reasoning, arXiv:1905.06088.
20. K. Gold, Norvig vs. Chomsky and the Fight for the Future of AI, *TOR.COM*, 21 June 2011.
21. E. Grefenstette et al., Experimental support for a categorical compositional distributional model of meaning, Proceedings of EMNLP'11, pp. 1394-1404, 2011.
22. T. Janssen, Frege, contextuality and compositionality, *Journal of Logic, Language, and Information*, vol. 10, pp. 87-114, 2001.
23. Y. Katz, Noam Chomsky on Where Artificial Intelligence Went Wrong, *The Atlantic*, 1 November 2012.

24. G. Lederman, M. N. Rabe, E. A. Lee, and S. A. Seshia, Learning Heuristics for Automated Reasoning through Deep Reinforcement Learning, arXiv:1807.08058.
25. Y. Maruyama, AI, Quantum Information, and External Semantic Realism: Searle's Observer-Relativity and Chinese Room, Revisited, *Fundamental Issues of Artificial Intelligence*, Synthese Library, vol. 376, pp. 115-127, 2016.
26. Y. Maruyama, Compositionality and Contextuality: The Symbolic and Statistical Theories of Meaning, *Springer LNCS*, vol. 11939, pp. 161-174, 2019.
27. Y. Maruyama, Contextuality across the Sciences: Bell-type Theorems in Physics and Cognitive Science, *Springer LNCS*, vol. 11939, pp. 147-160, 2019.
28. Y. Maruyama, Rationality, Cognitive Bias, and Artificial Intelligence: A Structural Perspective on Quantum Cognitive Science, *Springer LNCS*, vol. 12187, pp. 172-188, 2020.
29. Y. Maruyama, The Conditions of Artificial General Intelligence, *Springer LNCS*, vol. 12177, pp. 242-251, 2020.
30. J. C. Maxwell, *The Scientific Letters and Papers of James Clerk Maxwell: 1846-1862*, Cambridge: Cambridge University Press, 1990.
31. C. McLear, Kant: Philosophy of Mind, *Internet Encyclopedia of Philosophy*, retrieved on 2 February 2020.
32. M. L. Minsky, Logical Versus Analogical or Symbolic Versus Connectionist or Neat Versus Scruffy, *AI Magazine*, vol. 12, pp. 34-51, 1991.
33. R. Montague, Universal grammar, *Theoria*, vol. 36, pp. 373-98, 1970.
34. P. Norvig, On Chomsky and the Two Cultures of Statistical Learning, in: W. Pietsch et al. (eds.), *Berechenbarkeit der Welt?*, Wiesbaden: Springer, 2017.
35. A. Paliwal, S. M. Loos, M. N. Rabe, K. Bansal, and C. Szegedy, Graph Representations for Higher-Order Logic and Theorem Proving, *Proceedings of AAAI*, pp. 2967-2974, 2020.
36. F. J. Pelletier, Did Frege Believe Frege's Principle?, *Journal of Logic, Language, and Information*, vol. 10, pp. 87-114, 2001.
37. Z. G. Szabó, Compositionality, *Stanford Encyclopedia of Philosophy*, 2017.
38. N. C. Thompson et al., The Computational Limits of Deep Learning, arXiv:2007.05558, 2020.
39. P. Turney and P. Pantel, From Frequency to Meaning: Vector Space Models of Semantics, *Journal of Artificial Intelligence Research*, vol. 37, pp. 141-188, 2010.
40. W. Zhou and W. Li, Safety-Aware Apprenticeship Learning *Proceedings of CAV*, pp. 662-680, 2018.