Explaining Everything

David Davenport

Computer Engineering Dept., Bilkent University, Ankara 06800, Turkey.   
email: david@bilkent.edu.tr

**Abstract** Oxford physicist David Deutsch recently claimed that AI researchers had made no progress towards creating truly intelligent agents and were unlikely to do so until they began making machines that could produce creative explanations for themselves. Deutsch argued that AI must be possible because of the Universality of Computation, but that progress towards it would require nothing less than a new philosophical direction: a rejection of inductivism in favour of fallibilism. This paper sets out to review and respond to these claims. After first establishing a broad framework and terminology with which to discuss these questions, it examines the inductivist and fallibilist philosophies. It argues that Deutsch is right about fallibilism, not only because of the need for creative explanations but also because it makes it easier for agents to create and maintain models—a crucial ability for any sophisticated agent. However, his claim that AI research has made no progress is debatable, if not mistaken. The paper concludes with suggestions for ways in which agents might come up with truly creative explanations and looks briefly at the meaning of knowledge and truth in a fallibilist world.

Introduction

Is it possible to explain everything? Doing so would certainly involve explaining how we human beings not only explain the world around us, but also do all the other complicated things we do. Might it involve something more, something beyond even our abilities? These are questions that David Deutsch, an Oxford physicist renowned for his work on quantum computation, attempts to answer in his recent book (Deutsch, 2011), "The Beginning of Infinity". He claims that any agent (natural or artificial) endowed with a certain minimum set of capabilities, can begin this infinite journey towards explaining everything. He goes on to suggest, however, that Artificial Intelligence (AI) research has made no progress towards this goal. What's missing, he argues, is the ability to autonomously generate creative new explanations, and he traces this failure to the inductivist philosophy he sees as inherent in much AI work, promoting fallibilism as the way forward.

This paper examines Deutsch's claims, finding some merit in them, but also suggesting that some AI research is indeed moving in the direction he suggests. I begin by considering his claims about AI's lack of progress and the problems we still face with even basic terminology. I then take an in-depth look at the inductivist philosophy that Deutsch claims is to blame for AI's woes, and examine the fallibilist philosophy he proposes we adopt instead. I will argue that the philosophy we choose directly impacts the architecture and implementation of the agent, and that ultimately the fallibilist approach not only makes more sense, but makes it much easier to construct and maintain the internal (mental) models necessary for true intelligence. These considerations may even have implications for philosophy, requiring us to take a fundamentally different view of what constitutes knowledge and truth.

Progress in AI?

Clearly, we do not yet have the sort of intelligent machines Science Fiction writers and Hollywood films have long envisaged —and perhaps that's a good thing,— but has there really been no progress towards AI, as Deutsch claims? To some extent the answer depends on who you ask, what their expectations are and what they see as the goal of AI research. For some, especially philosophers and neuroscientists, the purpose of AI research is to understand how the human brain works its magic. While I can't vouch for the philosophers, neuroscientists are certainly beginning to develop a good understanding of the (low-level) biological basis of cognition, though they still have a very long way to go. For others, engineers in particular, the aim is usually just to build better, smarter machines. Whether such devices solve problems in the same way that humans do is usually irrelevant; they may be inspired by solutions seen in the human or animal kingdom, but they are just as likely to adopt a completely different approach if they can find one that produces results more efficiently than than evolution has thus far offered. Of course, most such devices are restricted in scope. Missile control systems, anti-lock braking systems, are just a few examples of literally hundreds of thousands of these mundane systems that incorporate algorithms and techniques developed by AI researchers, many of which perform far better and more reliably than humans could possibly manage. To date, very few machines have matched, let alone surpassed, human-level performance in what most people would consider intellectual tasks—playing chess, diagnosing human illnesses, playing Jeopardy, etc. Those high-profile systems that have, such as IBM's Deep Blue, which beat World chess champion Gary Kasparov in 1997, and its recent successor Watson, which won the TV game show Jeopardy just last year, are far from general-purpose; being focused on a specific task, most are simply unable to cope with even minor changes to that task or to the environment (though Watson is more flexible than most).

While "clever", none of these machines could be described as being really "intelligent". But, then, what is intelligence? Surprisingly, we do not even have a good definition of intelligence. Humans are the only widely-accepted example of intelligent beings, and even this status is open to debate. Legg collected and analysed around 70 definitions from the literature and from experts, synthesising his own definition: "Intelligence measures an agent's ability to achieve goals in a wide range of environments." Notice that this is an essentially behaviourist definition, as it doesn't concern itself with how the agent achieves its goals, only that it does achieve them.

Given that the very notion of intelligence is so vague, is it any wonder that progress towards it has been painfully slow? How do we know we are even on the right path, let alone getting closer? IQ tests, as applied to humans, have a long and chequered history . Legg discusses intelligence tests for humans, animals and machines—the latter including not only the infamous Turing Test, but also compression tests, linguistic complexity tests, multiple cognitive abilities, competitive games, etc.

Real intelligence requires achieving goals "in a wide range of environments", something most AI systems to date have found difficult to achieve. It was this obvious lack of generality that sparked the Artificial General Intelligence (AGI) program. Ben Goertzel, the researcher who coined the term AGI, responded to Deutsch's claims in "The real reasons we don't have AGI yet" , listing: the weakness of current computer hardware, the relatively minimal funding, and the integration bottleneck—that is, "intelligence depends on the emergence of certain high-level structures and dynamics", and "we have not discovered the one algorithm or approach capable of yielding the emergence of these structures", ... "integrating a number of different AI algorithms and structures is tricky" and "so far the integration has not been done in the correct way". All of which sound more like excuses that merely confirm Deutsch's claim about the lack of progress in AI. And yet Deutsch, like Goertzel, passionately believes that AGI must be possible.

“Despite this long record of failure, AGI must be possible. And that is because of a deep property of the laws of physics, namely the universality of computation. This entails that everything that the laws of physics require a physical object to do can, in principle, be emulated in arbitrarily fine detail by some program on a general-purpose computer, provided it is given enough time and memory. … [In the 1980s] I proved it using the quantum theory of computation.

So is Deutsch correct in his analysis? Others in the field have accused him of dabbling in topics outside his area of expertise and of being unaware of recent developments in the field. Whether Deutsch is aware of current research in AI and cognitive systems, I don't know; nowadays it is difficult for anyone to follow research in such a broad field. Even so, we shouldn't dismiss ideas just because they come from an outsider; without the burden of prior knowledge they may recognise things that those "in the trenches" are just too close to see.

Unfortunately, we continue to encounter the same difficulties when trying to define many other terms commonly used in discussing AI, including "information", "explanation", "representation", "truth", "knowledge", "computer", and "computation". This, surely, is an indication of a field that lacks solid foundations, its practitioners locked in endless wrangling over terminology. In an attempt to prevent such misunderstandings, at least within this paper, I will begin by outlining how I see some of these notions. I will not attempt to give precise definitions, but rather indicate the general ideas and relationships as I see them.

Agent Architecture, Models and Terminology

As depicted in Figure 1, agents are a part of the world—that is, we can view them as having a physical existence (a body)[[1]](#footnote-1) in a physical environment (the world)[[2]](#footnote-2). Agents have sensors (inputs), that allow them to detect only certain very limited aspects of the world, and physical actuators (outputs), that enable them to bring about (very limited) changes in the world. In between the sensory input and the actuator output stands the agent's "control" mechanism, which connects (maps) the inputs to outputs.

C:\Users\David\Documents\research2013\CreativeBlocks\Figure1.emf

**Fig. 1.** An agent has a body and a control mechanism  
that mediates between inputs & outputs.

This mapping may be a very crude and direct mechanism that produces the same response whenever a particular input is sensed—for example, the automatic light inside a refrigerator that switches on when the door is opened, a thermostatically controlled radiator, or Watt's centrifugal engine speed governor. Such devices, while useful, barely qualify as agents. Of more interest, are mechanisms whose mapping involves multiple internal states, and agents that can learn from (i.e. modify the mapping as a result of) their interactions with the environment, such that they may produce different responses in essentially similar circumstances. Agents can vary dramatically in terms of the number of inputs, outputs, and states available for the mapping, as well as the internal organisation/mechanism responsible for managing all this.

The sort of agents we are here primarily concerned with, are ones that can create and maintain internal (mental) "models" of their environment, models they can use to predict the effect of their actions or inaction, so as to enable them to select the output(s) most likely to benefit them. In other words, agents have "goals", that is, they desire to bring about certain states in the world[[3]](#footnote-3). In the case of biological agents, such goals include satisfying very basic survival needs—for food and water, for flight from predators, for reproduction, etc. If the agent exists in a social group, then more abstract, long-term goals may also be apparent—e.g. those related to job, responsibility, relationships, etc. The agent's task then, is to select appropriate action sequences in order to satisfy its goals as best it can, given its limitations and the vagaries of the world. The degree to which agents succeed in managing their world is a measure of their intelligence (c.f. Legg's definition of intelligence in the previous section). Given the huge variability (seemingly) inherent in the environment, the more sophisticated an agent's ability to "understand" its world (to model and predict it), the more successful it is likely to be.

An agent's model, then, is its "understanding". What is a model? A model is something that allows us to answer questions about another system (the system-being-modelled—the environment, for instance) without actually needing to interact with the other system. Such a substitute is particularly useful when the system itself is too big or too small to manipulate, or when doing so would be too dangerous, or when it doesn't actually exist but we would like to know how it would react were we to build it. Models are useful in everyday situations too: from predicting where a moving object will reappear after being obscured, to calculating how to move one’s arm and hand to pick up a cup of coffee. Even very conventional tasks, such as stock-control, can be formulated in terms of models. The model need only correspond to the system-being-modeled in ways relevant to answering the questions that will be asked of it. In most cases, this will require the model have states that can be mapped onto those of the system-being-modeled and that the sequence through which such states evolve from given initial conditions, correspond to those that the system-being-modelled would go through in the same circumstances. The model, then, can be used to simulate the relevant aspects of the system-being-modelled. Of course, such a model may not be entirely accurate and external influences acting on the system-being-modelled may result in it behaving in ways unforeseen by the model (e.g. in the stock-control system, parts may be mislabelled or stolen, leading to inaccuracies in the model). Long term predictions are thus likely to be increasingly error-prone. Notice that the speed of simulation is not especially relevant, what matters—what defines the simulation—is the sequence of states/events[[4]](#footnote-4). Of course, for an agent interacting with the real world, being able to perform simulations rapidly enough is vital, as it can mean the difference between life and death.

Any physical system with dynamic (causal) characteristics that can be mapped to those of the system-being-modelled, can act as a model. This notion of model is clearly material-independent; that is, models of the same system could be implemented (instantiated) in many different ways and in a variety of different materials, all of which would allow the user to run the same simulations and so answer the same set of questions[[5]](#footnote-5)[[6]](#footnote-6). Models may be constructed either by finding an existing physical system with the appropriate causal structure, by constructing such a system anew, or, more commonly nowadays, by programming a general purpose computer.

The digital stored-program (von Neumann) computer is a universal computing device that can easily be set up ("programmed") so as to model (almost) anything. A program (algorithm), then, is an abstract specification for a causal system that can be used as a model, represented in a language and form that the machine can interpret so as to set up the appropriate causal connections. A program (algorithm) defines a model and so a set of causal systems (computers) that would implement the model. A program transforms a universal machine into a specific single purpose machine. We also commonly say that a program specifies a computation, meaning either a single run of the model with given inputs or the entire set of runs of which it is capable. Whilst the term "computer" has become practically synonymous with the general purpose von Neumann machine, it should now be clear that the term is much more general, applying equally to general and specific purpose machines howsoever constructed. Deutsch sees the "universality of computation" as a "law of physics", which seems a strange categorisation given that computation is specified independently of the implementation (and that, as we will see shortly, the very notion of a "law" is at odds with what Deutsch is proposing). His claim that any physical system "can be emulated in arbitrary fine detail" seems correct, though for the wrong reason, since computation need not be restricted by time and memory as are machines of the von Neumann ilk— another instance of the same physical substance could be used to perform the computation.

Returning to Figure 1, the agent's control system (a computational model) affects the body, which in turn affects the environment; the loops—the feedback that results in subsequent input to the control system—may occur wholly within the body, and/or directly or indirectly via the environment, including other agents. This "picture" shows that the embodied, embedded and situated approaches to cognition merely emphasise different aspects of the same basic arrangement, while all remaining computational in nature (i.e. Computationalism is "Still the only game in town" (Davenport, 2012), given the understanding of computation outlined above).

One very important step on the path to explaining everything is the acquisition of language. Agents can use their physical outputs to generate signs/symbols in the form of sounds and/or the placement of materials, in such a way that these come to have meaning for other agents and for the agent itself. The inputs an agent receives can then include such audio and visual linguistic elements, as well as the "raw" input from the environment. We can now define "information" as input that "forms" the model or input that "informs" the model. That is, an agent's input contributes to the model (as experiences) and is used by the model to decide what is "out there" and hence how to react. Such linguistic elements enable the agent to bring to mind things which are not currently in their immediate environment (something especially useful once the agent has developed an inner-voice) effectively freeing them from the here-and-now and allowing the conjunction of arbitrary concepts. The outputs an agent makes now also include linguistic elements, in the form of commands, requests, questions, answers, descriptions, and explanations, to induce other agents to make changes in the environment or to asking them for information so as to update the model. Other elements, however, are intended to communicate the contents/structure of the agent's model, so as to update the mental models of other agents.

Note that we often speak of the information being contained in the model and of the model being the explanation for some phenomena. Deutsch points out that prediction is not explanation—you can predict that the woman sawn in half by the magician will be unharmed, but what you really want is an explanation of how the trick works and why your observations fail you. Thus, when Deutsch uses the term "explanation" he is actually referring to the knowledge inherent in the agent's model, rather than the spoken statement of its content. This usage is understandably common in scientific circles, since science aims to provide an objective picture of a phenomena; a scientific explanation refers, in essence, to the "knowledge" that is common to the models in the heads of all scientists. Woodward (2011) provides a good overview of theories of scientific explanation.

Given this general outline of cognitive agents as being ones that employ computational models to maximise their chances of success, the difficult technical question of just how a physical mechanism can create, maintain and use such models, arises. There has been no shortage of suggestions from philosophers and the AI community, as to how to do this, yet to-date there seems little consensus regarding the proper approach. It is here, I believe, that philosophy really matters and that Deutsch's claim that much work in AI has adopted the wrong, "inductivist", philosophy makes sense. Deutsch argues that if progress is to be made, AI must instead adopt "fallibilism" in order to be able to generate novel hypotheses. I agree, and I also believe that fallibilism will make it easier to create and use models. To understand why, we first need to examine both of these philosophies.

Inductivism vs. Fallibilism

Induction was (and in many ways still is) seen as fundamental to science and the scientific method. It was part and parcel of the romantic view of science and scientists discovering Knowledge and revealing the Truth about Nature, that has long held sway (at least in Western thought). Induction, in essence, involves taking a set of individual observations and constructing a universal rule (or law) from them[[7]](#footnote-7). The paradigmatic example involves deriving from the observations that the Sun rose in the East today, and yesterday, and the day before that, and indeed for all of recorded history, the rule that, "the Sun rises in the East every day" (hence justifying the inference that it will rise in the East tomorrow). Similarly, the observation that people get old and die, allows us to induce the law that, "all men are mortal" (so that, given "Socrates is a man", we can deduce that "Socrates is mortal"). Another classic example takes observations of lots of white swans, to derive the rule that, "all swans are white". While the former examples appear reasonable, the latter shows the weakness of this form of reasoning. That it is invalid becomes apparent the moment you encounter your first black swan (and is further eroded when you learn that there are actually swans which are both black and white). The problem, as pointed out long ago by Hume, is that it is simply impossible to examine all relevant instances, as would be necessary to guarantee a law-like relation. The fact that induction is demonstratively invalid thus presents a very serious problem for a scientific method that seemingly relies on examining the world and discovering such universal laws (Vickers, 2013).

It was Karl Popper (1959) who pointed out that no matter how many matching instances were uncovered, scientific "laws" were actually just theories that could never be confirmed (proven correct), but that could be falsified. A single failure to match (like the discovery of a black swan), could in principle invalidate a scientific theory. In practice, of course, scientists do not immediately eliminate a theory the moment they come across a single mismatch. Rather, they will first look to explain the erroneous value; were the experimental conditions wrong, were the measurements mistaken, etc. Repeatability of experiments is critical to the validity of science. Even if the failures were real, scientists would tend to hold on to an existing theory, embellishing it with special cases as necessary, until it would eventually "collapse under the weight of its own improbability". And even then, until a new theory could be found—one that covered all the troublesome cases as well as the "normal" cases—the existing theory would continue to hold sway. The notion of a critical experiment, one that clearly falsified a theory and so forced practitioners to switch to a new paradigm, was a convenient fiction—a rewriting of history. Rather, as subsequent studies in the Philosophy of Science (by Lakatos, Kuhn and Feyerabend, among others) have demonstrated, scientific progress is much messier and more subject to social forces than many, including Popper, had realised. The idea that multiple theories can usefully co-exist is particularly significant, especially considering that none of them need necessarily be complete or even correct. Indeed, it becomes clear that the very notion of a single completely correct theory may be an unattainable ideal.

None of this, however, detracts from Popper's insight into the nature of scientific theories; the fact that they are fallible, and forever subject to revision. It is this fallibility, coupled with the idea that all scientific theories are actually conjectures that can and must be tested and replaced if found wanting, that Deutsch sees as necessary for progress towards true AI. The question of how and whence conjectures come, is now of crucial importance. Indeed, part of Deutsch's claim, is that creativity and, in particular, the ability to create new explanations (new conjectures), is what is necessary for agents to be able to explain everything. This seems correct. Consider Deutsch's example, the explanation that the points of light we see in the night sky are actually very distant suns. Surely no amount of induction would generate such an idea. It requires an intellectual leap—a flash of inspiration—of the sort that happens so rarely, that the humans who actually perform it (and are able to convince others of it) are often regarded as geniuses.

Model Building Made Easier?

The task of an individual agent parallels that of the scientist. As we have already seen, agents sense the world (make observations) and use this information both to form models, and as input for existing models which then generate predictions to guide future actions. If we already had a completely correct model, this latter task would be relatively easy. The problem is that we don't, and acquiring it in the first place corresponds to inductivism, which is impossible without infinite time, a luxury that real world agents most definitely do not have.

The alternative, fallibilism, places no such constraints on the correctness of its model, indeed, it isn't even restricted to a single model. Fallibilism allows multiple, possibly incomplete, possibly erroneous models to coexist. What matters—what is absolutely crucial to fallibilism—is that the "incorrect" models can, sooner or later, be removed from consideration leaving the better ones to guide the agent's future actions. This is Darwinian "survival of the fittest" (most useful) at an altogether different level. Two questions now arise: how can an agent generate new hypotheses, and, how can it select relevant models and weed out the "bad" ones?

How can an agent know which hypotheses are the "bad" ones, how can it know it is wrong? In one sense it can't. All it can do is predict what it might expect to happen in a given situation and then sense what actually happens. If what happens is what it expected, its model seems sound, but if something else—something unexpected—occurs then clearly its model is not completely correct. It can't immediately know what is wrong with its model, but it does now have additional information about the situation, which it can take into account in the future. At higher levels (i.e. levels involving language and abstract thought), it is possible that the agent could run some simulations "in its head", determine whether or not the results were consistent with its expectations, and "label" the hypotheses accordingly so that they are easily included or excluded from future consideration. Only those hypotheses which appear reasonable need be tested.

This, then, provides an insight into how agents can construct hypotheses/models in the first place. Having remembered (stored) as much as they can of what happens, including any actions they make, they match the new (sensory) inputs against the existing ones. The process of doing this is essentially abductive inference, guided by top-down expectations from existing hypotheses[[8]](#footnote-8) (the memories created when the agent first encountered something different). Those hypotheses (memories) that recur and so prove useful, live to fight another day, whilst the others may eventually die off or at least not be included in subsequent computations. This, of course, is not enough for AGI, but it does get us started. It only takes a few small miracles to get from here to the other major component we need, which is language. Once an agent can name (and so group) arbitrary concepts, it can more easily detach itself from the here-and-now and contemplate currently unsensed things. Much of what constitutes (high-level) creativity is simply the result of chance combinations of the present sensory experience with memories (knowledge) of things previously named. It is this simple "conjunction" (joining) of ideas that is the spark. Some prolific inventors are known to adopt this approach, systematically combining concepts at random and investigating the consequences.

Of course there are other, more sophisticated ways in which agents can generate new hypotheses, including, abstraction, analogy and what we might refer to as pseudo-induction. By pseudo-induction I mean extracting "rules" from collections of observations, exactly as with induction, but with the "understanding" that the results are merely hypotheses not universal laws. This then allows us to continue to make use of induction without the epistemic worries, which is exactly what scientists and most of the machine learning community actually does.

Analogy is obviously very closely related to modelling and explanation, and was also proposed in a response to Deutsch by (Wiedermann, 2013). Analogical reasoning comes in many varieties (Bartha, 2013). In essence, it involves selecting corresponding objects/states/processes between two systems (or perhaps more correctly, between models of two systems)—one of which we are usually familiar with—, noting that most of these relations hold true and from that inferring that others, whose status is currently unknown, are also true. Of course, there is no obvious reason that this should be so, hence analogy, like induction, is an invalid form of inference. It is, however, very common. In one sense we use analogies whenever we do something in a slightly different context. For example, each restaurant we visit is different, but we can abstract and draw analogies between them, such that we are usually successful in getting the food we want. More interesting and creative forms of analogy involve mappings between two completely different domains. A classic example of this would be modelling atomic particles as billiard balls. This works fine, until we realise that in some situations particles actually behave more like waves (another analogy) than billiard balls. Unfortunately, physicists were unable to find an analogy that provided the necessary intuitions, leaving them with the so-called, wave-particle duality. A similar situation is now being played out in quantum physics, for which there seems no suitable analogy at all[[9]](#footnote-9).

Finally, we should not forget that explaining everything should include not only what exists, but also what might exist. The ease with which we can construct mental models of fictional situations and worlds, is the basis of art and literature. And, of course, one of the biggest contributors to creativity is undoubtedly the social interactions that make up our culture. It is "chance" encounters[[10]](#footnote-10) with others, which provide further opportunities to combine or slightly modify things in new ways, that ultimately drives art, science and society.

Of Knowledge and Truth

This, of course, leaves us with the question of what knowledge is; what counts as knowledge when anything can be wrong. What we have in our heads are "beliefs"—so the story goes—, and only "justified true beliefs" are real knowledge; false beliefs clearly aren't and true beliefs may be merely accidental —there is no way to "know" whether they are knowledge or not, even if one seems perfectly justified in holding them (the Gettier problem, see (Nagela, Marb, & Juan, 2013)).

Peirce suggested that "knowledge", in a fallibilist world, was what scientists would ultimately come to believe "at the limit of enquiry". Whilst clearly correct to some extent, it is not a particularly helpful definition, for even if we appear to have reached the limit of inquiry and scientists have a consensus, they could still conceivably be wrong. History has numerous examples of beliefs that everyone for hundreds of years would have sworn were true, but which ultimately turned out to be mistaken (or at least not entirely correct), e.g. The Earth is flat, or the Earth is the centre of the universe. Despite Peirce's belief in fallibilism, this view of knowledge and truth still has echoes of inductivism.

A better answer may be the more "pragmatic" (fallibilist) one. Simply put, knowledge in the (old) sense of absolute true facts, does not make sense and must be replaced with a less certain version. This new sense dictates that knowledge is always subject to revision and that there may well be multiple alternative, equally valid, visions of the world. This is not to imply that anything goes; knowledge and truth are still tempered by "reality"—be it the reality of a formal mathematical system, a fictional world set in another time or on another planet, or the "actual" world we all appear to inhabit.

Concluding Remarks

This paper attempts to understand and respond to Deutsch's recent critique of AI research. It began by trying to clarify some fundamental ideas and terminology related to AI agents. Given this understanding of what agents are and how they may function in the world it became clear that Deutsch makes a valid point when he says that fallibilism is the only basis for creating genuine AI. This is because the choice of philosophical approach affects the ease with which sophisticated agents can create and maintain models. However, when Deutsch claims that AI research has failed to make any progress exactly because it lacked the proper philosophical approach, he is only partially correct, as there are indeed a number of research programs based on fallibilist assumptions. Work by Hutter, Friston, and even Floridi, show both the spark of a new understanding and the difficulty of the undertaking.

Despite calling his approach Universal Induction, Hutter (2005) acknowledges that inducting universal laws is impossible in practice. His elegant theoretical approach to AI thus relies on (what we termed) pseudo-induction, building on work by Smolonsky and Kolmogorov to effectively provide, as he puts it, "a gold standard" for AGI (Rathmanner & Hutter, 2011). Unfortunately, like Deutsch's multiverse approach to quantum theory, it involves infinities that make it uncomputable (echoes of inductivism again, perhaps?) In contrast, Andy Clarks' recent survey paper on Predictivism (Clark, 2013) describes a more realistic approach to AI, one obviously in line with fallibilism (see also his reply to comments (Clark, 2013)). In the paper he examines the work of Friston (2008) & (Friston, Adams, Perrinet, & Breakspear, 2012), whose hierarchical neural network approach to AI sees top-down expectations being compared with input signals to feed-forward error signals (effectively the reverse of most ANN work to date). In another break from tradition, Floridi's (2011) Philosophy of Information offers a semantic view of information as well-formed, meaningful and truthful data. He (mistakenly) views information as being contained wholly "in the message" as it were, rather than in the interpretation of the message by an agent (which is what our analysis would suggest—see also (Adriaans, 2013). In this and in his Levels of Abstraction and the Correctness Theory of Truth, Floridi seems to adopt an absolute observer's perspective, clearly at odds with the fallibilist view presented here. And yet, if one takes a slightly different perspective, as outlined for example in Davenport (2012) (2009) & (1997), many of these ideas can align.

Certainly, we are still a long way from developing an AGI, but we do have a better understanding of the problem (and its difficulty). Sadly, a lot of time and research effort is wasted simply because we lack a common vocabulary and approach, surely an indication of a field in a pre-scientific stage. Perhaps Deutsch's "outsider's intuition" can help AI converge on the path to explaining everything.

References

Adriaans, P. (2013). *Information.* (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy (Fall 2013 Edition): http://plato.stanford.edu/archives/fall2013/entries/information/

Bartha, P. (2013). *Analogy and Analogical Reasoning.* (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy (Fall 2013 Edition): http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy/

Clark, A. (2013). Are we predictive engines? Perils, prospects, and the puzzle of the porous perceiver. *Behavioral and Brain Sciences*. doi:10.1017/S0140525X12002440

Clark, A. (2013). Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*, 181–253. doi:10.1017/S0140525X12000477

Davenport, D. (1997). Towards a Computational Account of the Notion of Truth. *TAINN'97.* Retrieved from http://www.cs.bilkent.edu.tr/~david/papers/truth.doc

Davenport, D. (2009). Revisited: A Computational Account of the Notion of Truth. In J. Vallverdu (Ed.), *ecap09, Proceedings of the 7th European Conference on Philosophy and Computing .* Barcelona: Universitat Autonoma de Barcelona.

Davenport, D. (2012). Computationalism: Still the Only Game in Town. *Minds & Machines*. doi:10.1007/s11023-012-9271-5

Davenport, D. (2012). The Two (Computational) Faces of AI. (V. C. Müller, Ed.) *Theory and Philosophy of Artificial Intelligence; SAPERE*.

Deutsch, D. (2011). *The Beginning of Infinity.* Penguin Books.

Deutsch, D. (2012, October 3). *Creative blocks The very laws of physics imply that artificial intelligence must be possible. What's holding us up?* Retrieved June 21, 2013, from Aeon Magazine: http://www.aeonmagazine.com/being-human/david-deutsch-artificial-intelligence/

Floridi, L. (2011). *The Philosophy of Information.* Oxford University Press.

Friston, K. (2008). Hierarchical Models in the Brain. *PLoS Comput Biol, 4*(11). doi:10.1371/journal.pcbi.1000211

Friston, K., Adams, R. A., Perrinet, L., & Breakspear, M. (2012). Perceptions as hypotheses: saccades as experiments. *Front. Psychology, 3*(151). doi:10.3389/fpsyg.2012.00151

Goertzel, B. (2012). *The Real Reasons We Don't Yet Have AGI.* Retrieved from http://www.kurzweilai.net/the-real-reasons-we-dont-have-agi-yet

Hutter, M. (2005). *Universal Artificial Intelligence: Sequential Decisions based on Algorithmic Probability.* Springer. doi:10.1007/b138233

Kaufman, A. S. (2000). Tests of intelligence. In R. J. Sternberg (Ed.), *Handbook of intelligence.* Cambridge University Press.

Legg, S. (2008). *Machine Super Intelligence (PhD. Thesis).* Lugano: Faculty of Informatics of the University of Lugano. Retrieved from http://www.vetta.org/documents/Machine\_Super\_Intelligence.pdf

Nagela, J., Marb, R., & Juan, V. S. (2013). Authentic Gettier cases: A reply to Starmans and Friedman. *Cognition*.

Popper, K. R. (1959). *The Logic of Scientific Discovery.* New York: Basic Books.

Rathmanner, S., & Hutter, M. (2011). A Philosophical Treatise of Universal Induction. *Entropy, 13*, 1076-1136.

Vickers, J. (2013). *The Problem of Induction.* (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy (Fall 2013 Edition): http://plato.stanford.edu/archives/fall2013/entries/reasoning-analogy/

Wiedermann, J. (2013). The Creativity Mechanisms in Embodied Agents: An Explanatory Model. *IEEE Symposium Series on Computational Intelligence (SSCI 2013).* Singapore: IEEE.

Woodward, J. (2011). *Scientific Explanation.* (E. N. Zalta, Ed.) Retrieved from The Stanford Encyclopedia of Philosophy (Winter 2011 Edition): http://plato.stanford.edu/archives/win2011/entries/scientific-explanation/

1. While agents commonly have well-defined physical boundaries, there is no obvious need for such a restriction, c.f. distributed agents and the extended mind hypothesis. [↑](#footnote-ref-1)
2. Nothing in what follows actually depends on there being a physical reality; there is simply no way we can prove that an external world exists, or that it has existed for more than a few seconds prior to the present moment, nor that it is not a figment of the individual's imagination. [↑](#footnote-ref-2)
3. Agents may not know what these states are, and initially may have them fulfilled either accidentally or purposefully through the actions of other agents, such as parents. [↑](#footnote-ref-3)
4. What the states are, how they are mapped, and how they are recognised and interpreted are also important questions. [↑](#footnote-ref-4)
5. Of course, the model's states and their interpretation may well change between implementations. [↑](#footnote-ref-5)
6. Not every material is necessarily suitable. [↑](#footnote-ref-6)
7. Some people include abduction and other uncertain inferences under the general heading of induction. In this paper I use induction in the narrow sense outlined above, rather than adopting this broader sense. [↑](#footnote-ref-7)
8. There is an exact parallel here with the idea from the Philosophy of Science that all experimentation is carried out within the context of a particular theory. [↑](#footnote-ref-8)
9. The inability to find a suitable analogy has led some physicists to suggest that intuitions provided by realist analogies (models) are unnecessary, and that the highly abstract mathematical formulations (also models) are sufficient in and of themselves. Deutsch disagrees and devotes a large part of his book to explaining his realist model of a quantum mechanical universe. [↑](#footnote-ref-9)
10. Modern society promotes individual development by explicitly creating such encounters, in the form of schools, museums, exhibitions, concerts, etc. [↑](#footnote-ref-10)