Deep limitations? Examining expert disagreement over deep learning

Carla Zoe Cremer¹

Received: 3 July 2020 / Accepted: 1 March 2021 / Published online: 26 June 2021
© The Author(s) 2021

Abstract

We investigate expert disagreement over the potential and limitations of deep learning. We conducted 25 expert interviews to reveal the reasons and arguments that underlie the disagreement about the limitations of deep learning, here evaluated in respect to high-level machine intelligence. Experts in our sample named 40 limitations of deep learning. Using interview data, we identify and explore five crucial, unresolved research subjects that underpin this scholarly disagreement: abstraction, generalisation, explanatory models, emergence of planning and intervention. We suggest that such origins of disagreement can be used to form a research road map to guide efforts towards overcoming the limitations of deep learning.

Keywords High-level machine intelligence · Deep learning · Expert elicitation · Generalisation · Planning · AI forecasting

1 Introduction

Deep learning has led to remarkable progress in artificial intelligence (AI) research [1]. We define deep learning here as using steps of differentiable computations for end-to-end training of multi-layered artificial neural networks. Deep learning has significantly advanced performance across traditional research areas such as image recognition, natural language processing and game play [2]. Its successful application across different domains has led to some debate and speculation on whether deep learning might be the key to fulfilling the ambition articulated by McCarthy et al. in 1955: to describe every aspect of learning and feature of intelligence so precisely that a machine could be made to simulate it [3]. Some researchers now proclaim that domain-general artificial intelligence could be achieved within a decade [4] and that deep learning will be central to its development [5].

Knowing whether or not such drastic progress is to be expected within short timeframes matters. Advanced AI will bring many challenges [6–10] ranging from privacy to safety concerns. Deep learning plays a role in this question: if currently used methods face few fundamental limits, progress is likely to be faster. But if limitations, as, for example,

articulated by Cantwell Smith [11] remain a challenge, progress is likely to be slower. Investigating the limitations of deep learning matters to prepare for the challenges posed by advanced AI ahead of time.

Foremost, the question is academically and strategically interesting in its own right. Different research teams strategically decide where to place their focus. They thereby place bets on what techniques are most likely to lead to results—be those insights into the nature of intelligence or engineered applications. Their choices can reflect the relative credence in the potential of deep learning. Take, for example, two companies which explicitly aim to engineer intelligence. OpenAI places a focus on scaling existing deep learning approaches [12–15] and emphasises the importance of increasing computing resources [4, 16] DeepMind in contrast is paying more attention to transferable lessons from the study of biological systems [17–19] and previous paradigms such as symbolic approaches [20, 21].

Surveys that captured the differences between expert expectations by asking experts for quantitative predictions [22, 23] on high-level machine intelligence (or HLMI defined in [22] as “when unaided machines can accomplish every task better and more cheaply than human workers”) show that indeed many experts do not rule out continued and drastic progress within the next decade and century. Grace et al. [22] asked 352 machine learning experts (21% of the 1634 authors who were contacted at NIPS and ICML 2015) to estimate the probability of reaching HLMI in future years. The aggregated responses provided

✉ Carla Zoe Cremer
carla.cremer@philosophy.ox.ac.uk

¹ Future of Humanity Institute, University of Oxford, Oxford, UK

a widely reported [24, 25] 50% chance of reaching HLMI within 45 years (from 2016). Yet a closer look (Fig. 1 in [22]) shows strong disagreement between experts. Forecasts by individuals are widely dispersed, ranging from high confidence of attaining HLMI within a decade, to attributing only a 10% chance of HLMI until the end of the century. Predictions differ so widely that they suggest quite different futures and actions. Without knowledge of the sophistication of arguments underneath these predictions, we cannot discern who's estimate is likely to be more accurate. Quantitative surveys provide no insight into the substance behind predictions.

Previous studies also chose a narrow selection of experts. It is clear who is an expert on machine learning, but less clear who is qualified to answer question on the nature of intelligence. While the sample in [23] was wide ranging in academic discipline (but still weighted towards computer science), [22] only surveyed attendees of machine learning conferences. This excludes much of the available expertise on intelligence (artificial and biological) from neighbouring disciplines and subfields. And as Page [26, 27] and Landemore [28] argue, the diversity between prediction models in a sample can make a substantial difference to prediction outcomes. Lastly, these surveys rely on an expertise that is not verifiably held by the experts they surveyed. Experts who did not receive training in forecasting tend to have poor predictive judgement [29–31], even in their field of expertise and especially for long timescales. This might question the reliability of previous quantitative surveys on AI progress.

Our approach addresses the limitation of quantitative surveys in three ways: first, by refraining from demanding quantitative predictions of AI experts, second, by diversifying the expertise in our sample size by expanding the notion of an AI expert to expertise with relevance to AI, including subfields of neuroscience, cognitive science, philosophy and mathematics and third, by focusing on the extraction of reasons and arguments behind expert disagreement.

Our aim is neither to predict what deep learning will or will not be able to do, nor when HLMI will be achieved. Both are questions that will be answered by research, not forecasts. We use the concept of HLMI only to stimulate the discussion on deep learning limitations. Our aim is to identify those research projects that will most likely advance our understanding of intelligence. We do this by mapping the disagreement without resolving or picking a position within it. The hope is to show that we can clarify, foster and use debate for the overall advancement of progress in AI—a goal common to both sides.

We conducted 25 expert interviews resulting in the identification of 40 limitations of the deep learning approach and 5 origins of expert disagreement. These origins are open scientific questions that partially explain different interpretations by experts and thereby elucidate central issues in AI

research. They are: abstraction, generalisation, explanatory models, emergence of planning and intervention. We explore both optimistic and pessimistic arguments that are related to each of the five key questions. We explore common beliefs that underpin optimistic and pessimistic argumentation. Our data provide a basis upon which to construct a research agenda that addresses key deep learning limitations.

This paper makes several novel contributions. First, we systematically collect a list of significant limitations of deep learning from a diverse set of experts. Second, we make these expert estimates legible and transparent, by revealing the reasons and arguments that underlie them. We thereby reduce the information asymmetries between insiders and outsiders to the debate. Third, we make use of our map of expert disagreement to identify central open research questions which can be used build a strategic research agenda to advance progress on current limitations.

This paper outlines the methodology, followed by the results which include a list of limitations and the analysis of arguments derived from interviews. The analysis has a description of common beliefs associated with optimistic and pessimistic viewpoints, a description of five origins of disagreements with evidential excerpts from interviews. We conclude with a discussion section which notes the limitations of this investigation and outlines the proceeding use of expert disagreement in the construction of AI research agenda.

2 Methods

We used the Consolidated Criteria for Reporting Qualitative Research (COREQ) checklist for reporting qualitative research [32]. Our sample selection does not aim to be representative of the frequency distribution of expert opinion [33]. It targets a diversity and variety of expert arguments, not a report on how frequently a view is held. Our aim is to display the arguments in the most objective way possible and to minimise any interference with how arguments are perceived. We want arguments to stand for themselves. To reduce biases and inference we made four methodological choices. First, we stay in line with previous surveys and choose to present positions anonymously to reduce any biases which the recognition of names might introduce. However, we do make transparent what level and area of expertise is found in our sample of experts. Second, we did not supplement each argument or limitation with potentially relevant references from the literature. This ensures that the arguments are not weighted by more or less numerous references. Moreover, an unbiased, full literature review across all disciplines we surveyed is outside the scope of this single study. We do not aim to present only correct arguments but to report on all reasonable arguments brought forward in

the academic community, so that the expert community as a whole, rather than one subjective view can, with clarifications and analyses such as this one, converge on the correct arguments over time. Third, we supplement arguments with quotations from interviews that represent this viewpoint directly. Note these quotations are only a concise selection of the evidence found in the interview transcripts. Forth, we do not give a verdict on the correctness of either viewpoint in the discussion section of this paper.

To ensure consistency between what experts were assessing, we provided the same prompt to all experts: the limitations of deep learning in respect to giving rise to HLMI. We avoided the ambiguous and often colloquially used term of “AGI” and generated the following definition of HLMI: a general or specific algorithmic system that collectively performs like average adults on cognitive tests that evaluate the cognitive abilities required to perform economically relevant tasks. Our definition differs from that of Grace et al. [22] in that we ask for cognitive skills needed to do human-level economic tasks and ignore the economic cost of performing the task. Note that cognitive tests of this kind, specifically for algorithms, do not yet exist. The fact that tests for HLMI as well as HLMI itself are currently an ill-defined fiction does not render our study less useful, but instead provides a degree of freedom for experts to choose what particular cognitive skills to assess deep learning on. This is reflected in the different level of abstractions, ambition and relative available scientific understanding of different limitations that were named (see, e.g. adversarial attacks, vs conscious perception).

We combined the non-probabilistic, purposive method and the stratified sampling method, as described by [34] in chapter six. Stratified sampling recognises distinct subcategories within the sample population and samples within subcategories. We sampled 25 experts from the following subcategories: disciplines (cognitive science, mathematics, philosophy and neuroscience), rank (degree attained) and sector (industry or academia). We selected those disciplines because their expertise is relevant to the study of intelligence. Cognitive and neural sciences examine biological intelligence, while subfields of mathematics and philosophy study formal notions of intelligence. We aimed to include experts across sectors because experts under different incentive structures (academia vs industry) might have different perspectives. We followed purposive sampling within subcategories: selecting experts with particular relevance (institution, expertise, research focus). We sampled researchers within subcategories by targeting those that had given a relevant talk at a conference, worked for an organisation aiming to engineer or study HLMI or had written related journal articles. Experts that were geographically close and/or familiar with the notion of HLMI (had publicly spoken about advanced AI or were recommended to us by other

interviewees) were preferred. The majority of interviewees were familiar with the notion of HLMI and located in London, Oxford or Cambridge in England and San Francisco or Berkeley in the USA. Experts were approached via email or personally at conference venues and, if receptive, met at an office or conference venue. No participant dropped out of the study, but several female researchers (but no male researchers) declined the invitation to be interviewed. One out of all the interviewed experts was known to the interviewer in a personal capacity.

The sample covered eight researchers in machine learning with specialisation in, for example, natural language processing, interpretability, fairness, robotics, AI progress metrics and game play; seven researchers in computer science all with specialisation in AI, specifically robustness and safety, progress metrics, natural language processing, machine learning, computational models of consciousness, symbolic AI and causal representation; two researchers in cognitive psychology with specialisation in developmental psychology and the cognition behind concepts and rationality; three philosophers with specialisation in comparative cognitions between animal brains and machine learning, philosophy of mind, of AI and of causality; two mathematicians with specialisation in machine learning, optimisation, Bayesian inference and robustness in AI; two computational neuroscientists, with specialisation in neural network applications to neural theory and one engineer with specialisation on AI and computation.

Some interviewees had interdisciplinary backgrounds, such as having worked in both philosophy, computer science and animal cognition. Interviews were conducted in 2019 to early 2020 with seven professors, nine postdoctoral, senior researchers in academia, six researchers (with at least Masters degrees and several years of research experience) in companies or institutes and three PhD candidates. Interviewees sometimes held positions in both academia and industry or had done so in the past. They held the listed position at the time of the interview. The sample is diverse in institutional prestige and expert seniority with deliberate inclusion of junior researchers. Our sample has a bias towards white, male researchers (23/25), reflecting a prevalence of males in higher academic positions and a male prevalence in the discipline of computer science.

We conducted individual semi-structured interviews. Semi-structured interviews use an interview guide with core questions and themes to be explored in response to open-ended questions to allow interviewees to explain their position freely [33, 35]. Each participant was provided information about the purpose of the interview, signed a consent form and was given an extended version of the given definition of HLMI (see Appendix 1). Interviews lasted 30–60 min, were recorded and conducted in English. Notes were taken during and after the interviews,

using the recordings. During interviews, only interviewee and interviewer were present. Interviews, note-taking and interview coding were all done by one person. No repeat interviews were carried out. The author devised a questionnaire (Appendix 1) as a guide, with questions like: What do you believe deep learning will never be able to do? Does image recognition show that neural networks understand concepts? Why can we not yet automate the writing of discussion sections of scientific papers? Do you see limitations of deep learning that others appear not to notice? In response to these and similar questions, all interviewees named their perceived limitations. Note that the questions were posed in recognition of the expertise of the interviewee. For example, a philosopher was not asked about the adversarial attacks in machine learning and a machine learning expert was not asked about cognitive development in children. We used this interview data to identify issues that play a central role in the disagreement between experts in our study. We identified these issues following guidelines for conventional, inductive content-analysis. A content-analysis is the “subjective interpretation of text data through the systematic classification process of coding and identifying themes or patterns” [36]. This approach does not utilise preconceived theories and uses both an analysis of manifest (literally in the text) and latent (implied) content [37]. All limitations named by interviewees are manifest interview content and were collated. The list of perceived limitations provided in the results section lists all limitations, with shortened, paraphrased explanations by the author. They are deliberately not ordered into categories and are anonymised. We conserved each expert’s preferred terminology, despite recognising that some limitations may turn out to refer to the same problem under different names.

We provide brief examples of how we implement the traditional coding schemes. Some limitations of content analysis approaches can be found in [38]. We combine coding of verbal designations (e.g. does interviewee use word “abstraction?”), scaling (e.g. is argument provided by interviewee a more pessimistic than optimistic argument?), simulation of hypothesis testing (e.g. does text support or refute the hypothesis that “abstraction” is an origin of the disagreement?). Variables, as required by methodological instructions, in our study are the origins (e.g. “abstraction”) of disagreement. Each variable is thus found in one of two instantiations: a pessimistic [e.g. “artificial neural networks (ANNs) do not abstract, thus have limits.”] and an optimistic (e.g. “ANNs do abstract, thus have potential”) instantiation of an issue (e.g. “abstraction in ANNs”).

We categorised arguments into pessimistic and optimistic instantiations and highlighted recurring themes in accordance with previous studies mentioned above. Themes occurred across interviews. They could be a theory, a justification, an open question, an experiment or a study. We

identified themes that were used as justifications in both more optimistic and more pessimistic arguments to make two opposing points. They are our variables. If both pessimistic and optimistic arguments would make use of the same variable to support at least two different positions, we would highlight this variable as an origin underpinning the disagreement. We demonstrate the methodology of identifying origins and propose their use in generating research projects that can address algorithmic limitations. This study was approved by the Ethics Committee for the School of the Humanities and Social Sciences at the University of Cambridge.

3 Results

3.1 Interview data

In Table 1 we present 40 limitations of deep learning as currently perceived by experts (at the time of the interviews). Experts differed on estimates of the number and severity of limitations. All interviewees held the view that deep learning is useful.

This set should be understood as a temporary best estimate of the true number and nature of limitations. Empirical progress will determine whether the number of limitations is actually larger (including currently unknown limitations) or much smaller (some limitations overlap and are not independent). Many limitations appear specific to deep learning. Others apply to AI research generally. Many perceived limitations already receive significant attention by researchers. Interviewees recognised this, but considered the problem insufficiently solved at the time of their interview.

3.2 Common beliefs: scale and insight

We observed common beliefs amongst both the optimistic and pessimistic viewpoints. The perceived limitations of deep learning related to the wider debate about the possibility and timeline of engineering intelligence. An experts’ position in the disagreement depended on how numerous and difficult they thought the limitations to be. Interviewees showed a nuanced consideration of the complexity of the question at hand. They rarely took a definitive stance and mostly differed in how much credence they attributed to a particular position being correct and in how much credence they assigned over plausible ways of achieving HLMI. One inclination, however, united interviewees across the optimism–pessimism spectrum: each believed their own view could more accurately reflect and incorporate the uncertainties inherent in the study

Table 1 Limitations of deep learning according to experts

<i>Representation</i> Insufficient ability to construct representations of the environment that are sufficiently abstract to allow easy and efficient storage, manipulation and generalisation of representations	<i>Common sense</i> Insufficient knowledge and access to an essential set of priors to help gradually build common sense knowledge, similar to a set of beliefs and facts that are useful across domains of tasks
<i>Concept formation</i> Insufficient ability to comprehend, manipulate and formulate abstract concepts for dimensionality reduction or language use	<i>Object permanence</i> Insufficient ability to represent the existence of objects even if objects are out of sight
<i>Causality</i> Insufficient ability to detect causal direction of how data was produced and inability to generalise learned causal relations	<i>Using unobservables</i> Insufficient ability to account for unobservable phenomena, to navigate and explain an environment, such as representing causation or subatomic particles
<i>Meta-learning</i> Insufficient ability of learning how to learn: choosing, on the basis of meta-data about learning tasks, what algorithmic adaptation learns best in each domain-specific learning task	<i>Cross-domain</i> Generalisation Insufficient ability to apply what is learned in one domain to the next domain
<i>Hierarchical decomposition</i> Insufficient ability to discern composition and modularity of observations and decompose tasks into sequential, hierarchical subtasks	<i>Overfitting</i> Insufficient ability to prevent a too close fit to and a memorisation of the dataset, hindering generalisation to other datasets
<i>Disentanglement</i> Insufficient ability to discern separable constituents of observations and separate, recombine and recognise those constituents in other contexts	<i>Variable binding</i> Insufficient ability to attach context-dependent instantiations to a placeholder symbol
<i>Adversarial attacks</i> Insufficient ability to protect against adversarial data	<i>Architecture-search</i> Insufficient ability to design neural networks automatically, using neural networks that search, select or generate the most suitable architecture of a learning network, its hyperparameters and model, to fit a learning task
<i>Grammar</i> Insufficient ability to construct and decompose sentences according to grammatical rules	<i>Analogical reasoning</i> Insufficient ability to detect similarity between abstract representations of observations across domains
<i>Mathematical reasoning</i> Insufficient ability to search, identify and develop mathematical proofs and follow logical deduction in reasoning	<i>Reading comprehension</i> Insufficient ability to detect narratives, semantics, themes and relations between characters in long texts and answer questions about the text without accessing the document itself
<i>Uncertainty estimation</i> Insufficient ability to represent and consider epistemic and model uncertainty. Insufficient representation of what is unknown and unknowable both for exploration and safety	<i>Visual question answering</i> Insufficient ability to answer open-ended questions about the content and interpretation of images
<i>Reinterpretations</i> Insufficient ability to recategorise, reassign and relabel previously stored data without retraining, in light of new information and theories	<i>Theorising and hypothesising</i> Inability to propose general theories and testable hypotheses, detect differences between theory and reality. Alter models to incorporate new data to approximate reality. Make inferences via exclusion, deduction and propose experiments
<i>Catastrophic forgetting</i> Insufficient ability to retain previously stored information as new data is incorporated into the neural network and weights are adjusted	<i>Flexible memory</i> Insufficient ability to recognise reusable knowledge, gained in one environment such that it will be stored in memory rather than forgotten, with stored representation being flexible enough to be updated as new knowledge is gained. The usability of stored information must be recognised in new environments and hence retrieved
<i>Efficient learning</i> Insufficient ability to learn a lot from little amounts of data	<i>Scientific interpretability</i> Insufficient understanding of neural network dynamics, which hinders research progress on, for example, memory storage or catastrophic forgetting
<i>Continual learning</i> Insufficient ability to learn continuously, instead of learning only once, resulting in fixed knowledge thereafter	<i>Active learning</i> Insufficient ability to follow a self-directed, autonomous, novelty-seeking exploration of environments that supports safe but efficient search, data collection and learning
<i>Inaccessible data</i> Data-dependent, gradient-descent-based approaches are inefficient when relevant data is either hard or expensive to access, protected or undifferentiable	<i>Dynamic data</i> Insufficient ability to learn from changing data streams. Most important data input changes continually over time since most phenomena change over time
<i>Brittle environments</i> Insufficient ability to navigate brittle, highly irregular, complex environments, instead of regularised environments with clear reward signals and short reward feedback loops	<i>Misguided data collection</i> Insufficient knowledge of what experience moulds and supports the formation of useful representations and effective learning in intelligent agents. It is unclear what data needs to be collected, observed and explored to support learning
<i>Generating valuation functions</i> Inability to understand and follow new rules immediately, without training time. Generate new valuation function from scratch to follow newly given rules	<i>Missing data</i> Data-driven approaches are ineffective in domains where data is strictly not existent, e.g. historical data
<i>Environmental pressure</i> Insufficient knowledge of the environmental pressures that support the genesis of intelligent agents. Complex environments are thought to play a role in shaping intelligence, but it is unclear what exact environmental pressures do so	<i>Scalability</i> Insufficient ability to use the same algorithms and retain the same performance while scaling the analysis from few to many features, using proportional resources. Adding additional features to a dataset requires providing disproportionately more data, parameters and computational power to retain performance
<i>Limits of simulation</i> Inability to learn all important knowledge in simulation. Not all relevant experience can be simulated, e.g. simulation of a human	<i>Metric identification</i> Inability to identify an accurate metric of task success, such that optimising for the measured quantity accomplishes the task, rather than the metric
<i>Conscious perception</i> Insufficient knowledge of possible functionalities of consciousness for higher-order reasoning agents, and replicability of consciousness through deep learning	<i>Context-dependent decisions</i> Insufficient ability to switch between and optimise towards time and context-optimal goals, just like humans optimise for their different physiological, social and psychological needs in a context-appropriate manner

of intelligence.¹ Emblematic of that view is the following quote of one (more optimistic) interviewee:

¹ “Partly, I’m just more uncertain about research progress: I won’t write off [...] a field [that is an alternative to deep learning] just because right now it doesn’t seem to be progressing very quickly.” (Participant 10). “[The compute focused view] seems overconfident to me—I just have more uncertainty around it. I’m not necessarily willing to take a bet that compute is going to get us to general intelligence.” (P25).

“There’s a bunch of world models, some of which predict 50% success in engineering HLMI in 10 years. Someone predicting a 0.1% probability of success has insufficient model uncertainty. They are over 99% sure many world models are wrong, even though they usually strongly agree we should be highly uncertain about progress over long time frames.” (P8).

Optimists attributed higher probabilities to reaching HLMI in shorter timespans than pessimists. Optimists were often impressed by how well trained ANNs generalise to test data. They attributed the success and potential of ANNs to their ability to identify useful representations of the environment, without using preconceptions based on human domain knowledge. Optimists stressed how much progress resulted from augmenting data and computational resources and warned not to underestimate potential performance gains derived from scaling existing methods.² They suggested qualitative differences, like perceptibly new skills such as reasoning and planning, might emerge from quantitative scaling.³ While pessimists interpreted limitations such as e.g. grammar or disentangled representations as insights missing, optimists either saw these as achievable within the deep learning framework or not essential for intelligence.

Optimists believed a trial-and-error research approach could lead to rapid progress towards HLMI, even without substantial improvements in the theoretical understanding of deep learning.⁴ They found it plausible that all foundational insights have been discovered and incremental improvements of deep learning techniques could suffice to build HLMI.⁵ Deep learning, they posited, may have stumbled upon the core components of intelligence.⁶

“I think existing techniques would certainly work given evolutionary amounts of compute. [...] A model [of how to achieve HLMI could be]: it’ll require some ingenuity and good ideas but it will be business as usual and it will rely on much more compute.” (P8)
 “[For instance] using convolutional attention mechanisms and applying it to graphs structures and training to learn how to represent code by training it on

² “We underestimated how far this goes [...] Just more of the same could work.” (P18) “I’m surprised when I’m not surprised anymore. I’m constantly underestimating the rate of progress.” (P5) “There could be one large network to approximate everything.” (P9) “We’re going to have a lot more compute, even in ten years.” (P14).

³ “If the environment is complex enough, you can’t use heuristics anymore and you will develop general reasoning abilities” (P9) “We’re seeing higher level strategies that no-one seems to have an explanation for. I don’t see any particular reason why it should not continue.” (P21).

⁴ “We know that one can get minds without understanding because evolution produced it.” (P21).

⁵ “No idea what it would be if not neural nets.” (P9) “I think it’s very unlikely neural nets will be totally replaced by a paradigm shift, but I can imagine something meaningful gets integrated with neural nets like external memory. That’s the largest kind of paradigm shift that appears plausible to me in the next 10 years.” (P8).

⁶ “What I think [ANNs] can learn in terms of representations and pattern are a superset of the things and patterns that I think brains can learn. Gradient descent is probably more powerful and a more fundamental method.” (P21) “It does seem plausible to me that all what intelligence is, is some incredibly complicated pattern recognition. [Deep learning] seems [to match] quite a lot of [the] things that we think intelligence to be.” (P25).

GitHub corpora...that kind of incremental progress would carry us to [...] superintelligence.” (P21).

3.2.1 Insight

Experts tending towards pessimism also shared beliefs. They considered deep learning useful but expected that essential insights are missing⁷ and paradigmatic shifts in research methods may be required. Missing skills for example include generalising, language and reasoning, and using abstract, disentangled representations. Pessimists often drew from their understanding of animal intelligence and stressed the difficulties of studying intelligence.⁸ Pessimists seldom suggested that deep learning captures central components of intelligence. They rarely believed that data and computing availability alone will lead to the emergence of new skills.⁹ Scaling of deep learning is no solution to them because they deemed it infeasible or inefficient. Pessimists believed new algorithmic innovations must overcome the problem of deep learning requiring disproportionate additions of data for each new feature.¹⁰

Previous successes were seen as only a weak indicator for future performance. The low hanging fruits of deep learning applications might soon have been harvested.

“Those people who say that’s going to continue are saying it as more of a form of religion. It’s blind faith unsupported by facts. But if you have studied cogni-

⁷ “If we had the right idea, probably current compute will be sufficient.” (P10).

⁸ “I believe that we can get sufficiently intelligent behaviour out of neurons that are significantly simpler than biological neurons. [...] But exactly which details need to be preserved is an empirical question. Anybody who says they have the answer today is being religious rather than scientific.” (P23) “I think [the engineering view] is a gross underestimation of the problem. [...] There’s a bunch of strange priors that you have, that maybe are really important. There’s strange stuff that happens in your peripheral vision for when you track the velocity of objects [...] maybe it’s not important but maybe it is. [...] Defining a task that tests the bit that we know about...I have no idea how big a slice of the problem it is, but I would make the assumption that it’s a small slice [...]. I would assume that we don’t know.” (P2) “AI is really, really hard. Making an intuitive concept precise is really hard. How do you take a rough and intuitive and vague concept and make it precise without losing that which you tried to capture? Because if you knew the precise nature of that which you’re trying to capture it would not have been vague and intuitive!” (P3).

⁹ “I would not agree with [this] statement [that AGI is a serious possibility in five years]. With keeping what we’re doing we’re not going to get to AGI. Maybe on the sort of problems that OpenAI is tackling, where you can simulate things, but on the natural language side, things are lot more messy” (P7).

¹⁰ “You need increasingly large amount of data to get deep learning to do better. It’s not linear. [...] Deep learning-based language tech uses trillions of words of training data and that to me is not scaling, it’s the opposite.” (P23).

tion, if you have studied the properties of language... [...] you recognise that there are many things that deep learning [...] right now isn't doing." (P23).

"My hunch is that deep learning isn't going anywhere. It has very good solutions for problems where you have large amounts of labelled data, and fairly well-defined tasks and lots of compute thrown at problems. This doesn't describe many tasks we care about." (P10).

Pessimists pointed out norms and practices in research communities (e.g. ineffective benchmarking and unpublished negative results) that could delay progress towards HLMI.¹¹ They particularly note the lack of a scientific understanding of deep learning techniques¹² and think that trial and error approaches have limited power in navigating researcher towards an ill-defined goal like HLMI.¹³

"If you think you can build the solution even if you don't know what the problem is, you probably think you can do AI" (P2).

3.3 Origins of disagreement

We identify key origins and scientific questions that underpin expert disagreement about the potential of deep learning approaches to achieve HLMI, which are: abstraction, generalisation, explanatory models, emergence of planning and intervention. These are scientific questions with incomplete evidence, about which experts propose different hypotheses or interpretations and thus end up disagreeing. For each origin, we show what arguments lead to pessimistic or optimistic views on deep learning. Origins of disagreement depend on perceived limitations: disagreement exists because limitations persist. For instance, ANNs are currently limited in representing higher-order concepts. This creates uncertainty about whether they are capable of doing so—an uncertainty within which the disagreement presides. Disagreement can be resolved as solutions are found and uncertainty is reduced.

We present a non-exhaustive list of origins that can be used to make progress towards expert agreement. They are a

subset of the origins of disagreement and do not map the dispute or limitations exhaustively. We identified an origin by noting when both optimistic and pessimistic experts referred to the same underlying issue to support opposing positions. Each section states the open question that gives rise to the disagreement and proceeds by paraphrasing pessimistic and optimistic arguments that experts reported as reasons for their position. Note that arguments given by different experts may be mutually exclusive but still point in the same direction. Quotes from interviews were provided as evidence if they captured the argument succinctly.

3.3.1 Abstraction

Do current artificial neural networks (ANNs) form abstract representations effectively? (Table 2).

3.3.2 Generalisation

Should ANNs' ability to generalise inspire optimism about deep learning? (Table 3).

3.3.3 Explanatory, causal models

Is it necessary, possible and feasible to construct compressed, causal, explanatory models of the environment as described in [39] using deep learning? (Table 4).

3.3.4 Emergence of planning

Will sufficiently complex environments be sufficient in enabling deep learning algorithms to develop the capacity for hierarchical, long-term reasoning and planning? (Table 5).

3.3.5 Intervention

Will deep learning support and require learning by intervening in a complex, real environment? (Table 6)

4 Discussion

4.1 Summary

Scientific uncertainty generates expert disagreement. A lack of data leads experts to make reasonable, but opposing interpretations of, and extrapolations from existing data. Limitations contribute to these origins. Both the uncertainty about what skills are required for HLMI and whether deep learning can support each prerequisite leads to disagreement. This disagreement can guide research efforts aimed at overcoming limitations.

¹¹ "There's one sense in which people are quite happy to get points amongst our research community for beating benchmarks performance and are not thinking about why do we care about this benchmark? [...] Another problem is papers only report the best performance they got after a lot of fine-tuning." (P10) "Because there's so much trial and error we don't really know which problems we've solved [...] you don't know the limits of how far this is going to go." (P25).

¹² "In most of machine learning, what works is not exactly what the theory is telling you. [...] We don't really understand ourselves how the deep models work. [...] Can we really control it in a scientific manner and not just an engineering manner?" (P11).

¹³ "With trial and error over what is almost an infinite problem space? Seems hard." (P2).

Table 2 Abstraction

Optimistic arguments	Pessimistic arguments
ANNs form abstract representations of their input in intermediate neuronal layers. There is no reason why ANNs, with a rich enough data source and computing resources, should not scale to generating higher-level abstractions (“The question is always- do we find the right representation? [...] Deep learning is searching for a good representation in the intermediate levels [and it] might have found a way of making better representations.” (Participant 13) “Read some of the outputs from GPT-2. It’s hard to say what the system knows but it’s a fairly high level of abstraction” (P21))	The ability to form and manipulate abstract representations is mostly lacking in ANNs now. (“[ANNs] are not doing much [abstraction] at all right now.” (P13) “The ability to manipulate abstract reasoning is currently lacking but there is no reason why this should be so.” (P2))
Abstract, symbol-like representations can be formed on the basis of available techniques, such as attention mechanisms (“Transformers and attention mechanisms [are] going to be very powerful in symbolic [...] representations.” (P21) “Symbolic representation [...] seems like something that [...] emerged from distributed representation.” (P8))	New algorithmic insights are required to achieve higher order abstractions, language, reasoning. (“If we’re able to solve the age-old knowledge representation problem via meta learning then [good], but if we can’t solve it then I expect other paradigms are eventually going to overtake deep learning.” (P10))
The non-discrete representations formed in ANNs might be similar to what we call intuition in humans This powerful process of knowledge representation underlies much of human reasoning. This non-discrete representations in ANNs could be the foundation of reasoning, concept formation and abstraction (“Neural nets seem to solve problems in a way that looks like intuition. A neuron can correspond to a concept like a sentiment and concepts are built up into higher level concepts. It seems plausible that with a bit of work ‘intuition’ could be the foundation for reasoning as concepts build up and become sufficiently abstract.” (P8))	Abstraction in ANNs is dissimilar from abstraction in humans. (“The best game-playing systems certainly learn some kind of abstractions, but we don’t really know anything more specific than that. They certainly don’t seem to be learning any abstractions that look human-like.” (P3) “We still haven’t got a good way of doing object recognition in video.” (P10))
Similarities between ANNs and brains are encouraging: ANN abstractions work similarly to visual cortex abstractions, biological neural patterns can be understood as vector representation and since human intelligence developed through culture, a human brain without culture would look similar to ANNs now. (“A human raised in total isolation from other humans would be a lot less ‘smart’. A lot of the sophistication of humans comes from our ability to accrue skills and knowledge through culture. I believe this leads us to overestimate our innate capabilities. ‘Cavemen’ were similarly good learning machines as homo sapiens today, even though when we picture them they seem a lot less intelligent.” (P8))	Top-level reasoning protects humans against adversarial attacks, which ANNs are vulnerable to. (“We basically made almost no progress [on adversarial attacks and this could be] a fundamental problem of deep networks or models [...] which do not have some direct knowledge of the structure of your problem.” (P10) “What fools humans is very different from what fools machines.” (P12))
ANNs are in fact superior in abstracting useful representations because they can do it with any data, whereas humans only do this in domains that they are evolutionarily adapted to. (“Deep learning has higher performance in entirely alien tasks.”(P21))	ANNs only co-locate surface features, so a node activation in response to stimuli does not mean ANNs understand the concepts shown. (“Our computer vision systems do not need to see the dog to say that it’s likely that we will find a dog in the image.” (P16))
	Researchers use increasingly more data but still approach a performance plateau in computer vision, which shows that we have not found the right representations and are memorising data. More data and computation will not make up for ill-suited representations. More of the same will not get a difference in kind. (“We don’t want to learn all the images via memory. [...]. Computation by itself does not do anything.” (P16))

4.2 Limitations

Our study too has limitations. Our focus on deep learning as it is now, means that one must be careful to use this data for estimates of AI progress beyond a few years. As our understanding of deep learning improves, its definition will change. Indeed, even now, interviewees had significant disagreement over the definition of deep learning. Deep learning might soon signify algorithms different from those discussed here, or of course, AI progress could occur without deep learning.

This study cannot show how many of the listed limitations are true limitations or how fundamental they are. Several limitations appear to overlap (e.g. representation and variable- binding), partially due to the high level of abstraction on which participants named limitations. A full

literature review is beyond the scope and non-interference goal of this paper.

We noticed that experts might be using different notions of key terms that factor in the discussion (e.g. abstraction, priors and generalisation) which might result in significant semantic disagreement. We do not resolve this here and encourage researchers to expand our analysis, to define terms or suggest additions to this list of origins. Our list of origins is by nature subjective and thereby would improve as more interpretations and viewpoints are added. Similarly, even though a debate-based research agenda will reduce expert disagreement, it is unlikely to lead to total agreement. As we mention, some expert differences depended on different interpretations of available data, different views on the nature of intelligence and different views on what

Table 3 Generalisation

Optimistic arguments	Pessimistic arguments
<p>ANNs ability to generalise evokes optimism about deep learning. Deep learning algorithms generalise well from training to test sets, if test data lies within the training distribution. This may be all that is needed. Humans too, are bad at reasoning outside training distributions (“[Neural networks] are impressive in their ability to generalise.” (P4). “Humans don’t generalise all that well.” (P9).)</p> <p>ANNs might not need to generalise across distributions either. The key to good performance is resource-intensive training on a wide enough distribution and a complex environment that encompasses most experiences (“I think people believe that we will train neural nets that will generalise really well, but I mostly think: not really. I also think humans don’t generalise all that well. [...] To the extent that you want to get AGI out of neural nets you’re going to have to train it on a sufficient complex environment. [...] I do think this is basically the approach that OpenAI and DeepMind are taking: give it a sufficiently complex environment and it will do well”. (P9))</p> <p>Deep learning is highly general and an all-purpose problem-solving tool, thus applicable to many datasets and domains. With deep learning, researchers discovered how to replicate the essence of intelligence: forming compressed representations and detecting invariance patterns in complex environments (“Deep learning might have stumbled upon a way of making better representations and that is explains why they generalise so much better than previous systems.” (P13). “It’s extremely general, it has long been shown that ANNs are computationally universal. [...] People [are] making enormous progress and it’s largely been a question of adjusting the shape of the computation. [...] What I think [ANNs] can learn in terms of representations and pattern are a superset of the things and patterns that I think brains can learn. [...] If you look at the systems from the right angle you will see they have something extremely universal.” (P21))</p>	<p>Deep learning cannot generalise across domains and is highly sample inefficient in comparison to humans “[ANNs] are just absolutely terrible in any new domain”. “[ANNs] are horrendously sample inefficient. (P9)” “Gradient descent is a very unprincipled optimisation algorithm [...] I don’t think you’re going to get efficiency out of deep learning.” (P10).)</p> <p>Proposals for reducing sample inefficient learning will likely fail: meta-learning for instance would shift (but not reduce) the burden of computational costs towards training at the meta-learning stage. Continued efforts to reduce sample inefficiency have led to minimal progress towards human-like learning, and that gives reason to be pessimistic.</p> <p>Humans generalise across domains, using compressed representation and analogical reasoning. Features like variable-binding, grammar, symbolic reasoning and decomposition are missing in deep learning, which explains its sample inefficiency and inability to learn well outside of training distributions.</p>

intelligence must and could be capable of. Some disagreement will likely remain. But our goal here is to point out how to use these difference not to minimise them.

Because we conducted interviews, our sample size was constrained and led to around about 25 of expert argumentation. It is necessarily not a full representation of all expert arguments. Future research could expand the list of arguments and sample even more sub-disciplines, increase the level of detail provided and investigate the validity of the arguments provided by experts. Finally, and foremostly we hope researchers will utilise these origins to generate specific experiments, concise definitions, benchmarks and a collective research agenda to test the hypotheses that underlie the theories of intelligence presented here.

4.3 Progress in artificial intelligence

Expert disagreement offers a rich landscape upon which to construct a research agenda to overcome the current limitations of deep learning. A research agenda guided by expert disagreement can (a) define key terminology in origins of disagreement, (b) dissect origins into tractable, feasible

research questions, (c) collect optimistic and pessimistic hypotheses on origins, (d) specify experiments that could falsify either of the hypotheses, (e) generate benchmarks and competitions that encourage coordinated experimentation on origins and (f) conduct experiments and falsify hypotheses.

We give some examples of the types of research questions which, if addressed, can contribute to progress on reducing uncertainty. For instance, developing tests to distinguish if deep reinforcement learning agents learn heuristics or planning, identifying games which can only be solved using high-order reasoning and planning to test the emergence-hypothesis, developing different measurements for degrees of abstraction and concept formation in ANNs, test the correlation between levels of abstractions formed and computing resources, define what prediction performed by an ANN would indicate that “model-building” has been achieved, defining desirable levels and types of generalisation that liken human generalisation and investigating the causes and extent of unexpected, unspecified, emerging skills in predictive language models.

Coordinated experimentation will reduce the uncertainty that gives rise to origins of disagreement and advance

Table 4 Explanatory, causal models

Optimistic arguments	Pessimistic arguments
Models or complex environments will likely emerge from predictive, generative deep learning algorithms (“We didn’t specify learning to count as an objective for GPT-2, it just learned it.” (P11))	Exploiting correlations through predictive, generative algorithms may not be enough to understand and navigate complex environments. HLMI will have to understand causal relations and explain how and why the observed data was produced.
Predictive training likely requires the agents to learn to reason about the world. (“You can imagine some kind of disembodied reasoning AI and it can know quite a lot of profound things even though it learns from reading texts. Imagine it has lots of data and it just tries to model that data and in order to do that it has to learn how to reason about the world.” (P14))	Human reasoning is helped by using rules and logic, such as grammar and the laws of physics. (“If you want true natural language understanding [...], it can’t easily be expressed with a neural network blob.” (P7) “There’s a lot more structure to brains than deep learning assumes” (P13) “[What research] I think is important is the difference between induction and deduction.” (P2).) These skills are only mimicked by current systems. (“It’s not surprising [that] GPT-2 learns conventions for how we use language. [They] built a system that is really good at mimicking language” (P25))
Predictive language models already learn skills, like counting, without targeting this skill explicitly by the loss function. Causality too, could be such a skill learned via predictive training. (“There’s a possible way to get to AGI quickly by simply scaling up language models. [...] A GTP-2 learns something about the world just by trying to predict the next word. And maybe you can really learn quite a lot that way especially if you have humans help nudge you in the right direction” (P14). “Having a predictive training objective [already] gives you some rudimentary world modelling” (P11). “Causal model would be learned and [are] not specified by the loss function.” (P11))	Such rules enable fast exploration of counterfactuals. Having a compressed model that allows agents to manipulate concepts internally to make predictions about the world and explore counterfactuals, is important for generalising, theorising and planning. (“We need an architecture that allows you to play around with internal concepts” (P13))
ANNs appear to already have the building blocks to generate compressed world models: They form abstract representations from high-dimensional data. They can navigate complex environments, like in Dota 2 and StarCraft II, where they show emerging reasoning and strategising behaviour (e.g. encircling the enemy), suggesting some world modelling of the environment is happening. (“I don’t see prediction vs hierarchical compositional processes as being a dichotomy.” (P23))	Since data is always finite, agents must use theory to make correct inferences beyond available data (“The deepest problem is that deep networks are actually very shallow about what they model about the world. [...] There is overwhelming evidence from cognitive science: one of the things that makes humans different from other animals: We have an incredible ability to posit things about the world that are unobservable. Not just unobserved but unobservable. Personality traits, disposition, causation. A deep network is not going to generate this [...] The reason we posit unobservables is because it allows us to generalise, transfer, prediction, control. I want to see interesting, deep level predictions.” (P3))
	Agents trained to play Dota 2 and StarCraft II do not construct causal models of complex environments: they make no long-term predictions or generalisations to new game environments, meaning they probably do not learn rules

agreement between experts. This research agenda should extend beyond computer science research, as it will benefit from interdisciplinary efforts in, for example, psychology, philosophy or animal cognition as demonstrated by [40–42]. Collaborations in which pessimists set the tasks that would make them more optimistic and in which optimists try solve the task using deep learning, could be fruitful. We suggest expert disagreement provides a road map for progress towards artificial intelligence.

Appendix 1: Interview procedure and questionnaire

Introduction given to interviewee

You can take as much time as you like before you answer a question. We want to collect informed opinions. If you do not know the answer to a question, please say so and we will move to another question. You can speak to me with on

highest technical level and assume that I understand what you’re talking about. If I don’t understand a term or the details of a technique, I will ask. Please name your title at your workplace, your workplace, field and expertise.

Definition human level machine intelligence

This interview is about exploring the potential and limitations of connectionism and in particular deep learning, for building high-level machine intelligence. We are interested in understanding deep learning in reference to high-level cognitive skills across domains of intelligent cognition. We are therefore not asking whether deep learning can, or will be, economically useful. We are not concerned with problems of consciousness, moral patienthood, etc. Instead, we are interested in the potential of machines to reach high-level intelligence on the basis of deep learning techniques.

According to Hutter and Legg, intelligence measure an agent’s ability to achieve goals in a wide range of environments. Here’s a more detailed definition: high-level machine

Table 5 Emergence of planning

Optimistic arguments	Pessimistic arguments
<p>If given sufficiently complex environments, data, computational power and run-time, it is possible that deep reinforcement learning agents will make macro-strategic, long-term, hierarchical plans. An agent in complex environments should develop general reasoning capacities because that is the best way to accomplish goals. Reasoning, planning, even cooperation, might emerge as a by-product if the environment is complex and computational resources abundant (“If the environment is complex enough, you can’t use heuristics anymore and you will develop general reasoning abilities” (P9) “You hope you might get it as a by-product, like in order to fight each other in the game they need to develop planning and cooperation.” (P14))</p> <p>Agents playing StarCraft II/Dota 2, showed some use of macro strategy, e.g. when encircling the enemy. We lack tests to exclude the possibility that agents developed general reasoning in these games: we cannot distinguish between planning or heuristics (“We’re seeing higher level strategies that no-one seems to have an explanation for. I don’t see any particular reason why it should not continue. It will take more compute to learn layer-after-layer higher-level representations. What we’ve been seeing is that higher and higher levels of abstraction and representation emerge spontaneously in ways we don’t understand.” (P21))</p> <p>But games played so far can be won without general reasoning and might not require the agent to develop reasoning even if it could. Future games might necessitate and foster planning (“You don’t need sophisticated reasoning to play this game. [...] For the right game with enough compute we might get sophisticated strategies”. (P14) “The lack of ‘thick’ concepts and abstractions will be most problematic just when the context changes in some important way, which doesn’t ever happen within these games / simulated environments.” (P3))</p>	<p>Planning, reasoning and hierarchal decomposition of strategies into subtasks is how humans accomplish goals in complex environments. Deep reinforcement learning agents do not show such skills. Simply adding more computing resources has not changed this so far (“We [provide] much compute and hope for more reasoning and strategy but we seem to see the same old problems” (P14) “There’s not been significant progress on a solution that doesn’t require human hand holding or unrealistic access to the environment”. (P10))</p> <p>Game play of agents in StarCraft II/Dota 2 are likely heuristics learned via trial and error. The agents made errors that humans who reason would never make and these heuristics did not generalise to slight variations of the game. Reasoning would have allowed generalisation.</p> <p>Agents won games via micro, not macro-play, which machines are understandably better at than humans.</p> <p>Higher-order reasoning has not emerged and planning is not even required to win such games that are actually very simple (“Standard RL benchmark tasks [...] are really solvable by [...] gradient free search over linear policies [...]. These tasks are much simpler than they actually seem.” (P10) “We don’t have a clear way of testing whether something is doing general reasoning or is learning heuristics” (P9)).</p>

intelligence (HLMI), describes an algorithm that performs as well as an average human in many economically relevant tasks, including patient diagnosis, speech writing or cleaning. Such tasks are only achievable given cognitive skills. Psychologists have developed a variety of tests that characterise a human’s cognitive skills. We thus define HLMI as an algorithm that is able to achieve average performance on the vast majority of these cognitive skill tests. They include IQ tests and sociality assessments which test skills like: memory recall, pattern and story completion, verbal comprehension, analogical, mathematical, analytical, fluid and spatial reasoning, social intent, one-shot learning, and can be extended to testing theory of mind, creative imagination and divergent thinking. An algorithm should perform in most of those standard tests as well as an average human would, to be considered a HLMI.

We do not assume these algorithms to be economically viable, widely spread or readily usable by all actors. We do not make assumptions about the actors involved. It could be companies, academics for state-actors. We do not specify whether the algorithm is general or specific. We do not assume that it must be one agent that achieves all tasks. Each task could be solved by a different agent/algorithm, and we would still describe these collective capabilities to be HLMI.

Questionnaire

What is deep learning (DL)?

Why and when does deep learning work?

Can you think of a task/skill that you think might be hardest/most implausible to do with ANNs?

Is there anything you believe ANNs will never be able to do?

I will now describe several intelligence tests taken from the cognitive sciences. Please tell me what cognitive skills are necessary, to score as highly as an average adult human on this test.

Intention: An artificial agent has a pencil lying in front of it on the table. The agent observes a human taking a piece of paper into his hands. The human then starts looking around, searching his pockets. Will the agent understand that the human is looking for a pen to write with?

Exclusion inference: “An artificial agent is given four objects. It is told to choose the blimb. It has not learned the concept of a blimb, but it knows that three of the four objects are not blimbs. Can it pick the blimb without trial and error?”

What skills are involved?

Table 6 Intervention

Optimistic arguments	Pessimistic arguments
<p>Deep learning approaches could lead to HLMI by a combination of learning from available Internet data and learning necessary precursor knowledge in simulations (“The most plausible story for short term HLMI probably does not involve a lot of robotics. Two things work out: systems can learn lot from the internet [and] training in simulation [is] sufficient to take you to some real world task.” (P14) “I would imagine that everything that my mind experiences can be simulated for an AI.” (P13))</p> <p>Embodiment via robotics is likely not necessary for HLMI (“You can imagine some kind of disembodied reasoning AI and it can know quite a lot of profound things even though it learns from reading texts. Maybe you get a system that doesn’t know what it’s like to tie a shoe lace but there’s many things that are much more important and someone who uses the system could really have a big impact on the world especially if someone uses it to improve the next version of AI.” (P14) “I’m not convinced you will get profound breakthroughs by being embodied and you could do just as well in a virtual environment. Eventually it will [in simulation] discover ‘if I will something, my body does something’. [...] That might be the first step in a long chain to figure out causality.” (P13))</p> <p>Good natural language prediction models, like those already achieved on the basis of ANNs, will help to learn a lot from available Internet data. Internet data is sufficiently vast. (“You might learn really useful things from reading the internet.” (P14) “You can get a lot of videos of people interacting and talking, so I can imagine something like GPT-2 [...] imitating normal social reaction.” (P13))</p> <p>Learning from the real world, rather than from simulations might be an infeasible research avenue for deep learning since humans can’t generate enough data. Learning in simulation like it is already being done will likely be the most efficient option and it won’t require learning with and from humans.</p>	<p>Manipulating our environment was a key factor in human evolutionary cognitive development and it is key to child development, which means it might be important for HLMI. (“If you can’t do active learning you will not understand unobservables. It involves intervention. [...] It would be bizarre if our cognition didn’t show any influences of our active engagement with our world in our concepts and the way that we understand it.” (P3) “Intelligence is an act in the real world.” (P6)) Learning key concepts may require robotics (“as children learn to manipulate their environment you see massive shifts in the way that they conceptualise the world.” (P6) “We haven’t got a good way of doing object recognition in videos.” (P10))</p> <p>It is unclear what some particular algorithm can learn from just reading Internet data. For instance, without precursor knowledge or the ability to experiment with how data is produced, data does not indicate to the observer, the causal direction through which this data was produced. Causality is crucial to understand environments.</p> <p>It is unclear whether all important priors can be learned in simulation. It is unknown whether all important precursor knowledge learned in simulation will transfer to real environments. (“unsupervised learning with very little data [...] might require totally new techniques. [Currently] there’s a lot of hand engineering of environments to make things simpler. [...] That’s another reason why I like robotics, because you can’t get rid of real complexity.” (P10)). Some essential learning experiences cannot be simulated (“By definition we can’t simulate human-level intelligence” (P14) “It’s not like babies learn to be intelligent by watching a lot of intelligent behaviour.” (P 13).)</p>

When do you think we can we do it? Can this skill be achieved using ANNs?

Emulation versus Imitation: “The artificial agent has never seen a person open a door. A person demonstrates opening the door for a single time, but happens to drop the key, before unlocking. The artificial agent repeats the exact actions after a single demonstration, but does not drop the key”.

World model: The “Alternative Use Task” is used to assess divergent thinking and creative potential in humans. A human is given an object such as a “Paperclip” and must come up with as many unusual use cases for the object as possible (e.g. “pinch a hole into paper, by bending the steel wire”).

More creative humans come up with more items that are more original and non-obvious. Responses are rated by other humans by originality (statistically uncommon), fluency (quantity), flexibility (how many different categories) and elaboration (detail).

Human subjects in divergent thinking tests are given the task: “Come up with as many problems as possible that could arise between you and your parents”.

The artificial agent is placed into a virtual office space that it has never seen in this exact configuration. You instruct: “Move to the white cupboard in the left corner and fetch the old-looking notebook out of the second drawer”.

Companies such as (<https://scinote.net>) are automating the writing of the introduction of scientific papers. Why can we not automate the writing of the discussion part?

If your company would have to place a bet on a technique, in the aim of creating high-level machine intelligence (in at least one domain), would you invest all your money and time on research-based ANNs work architectures or would you explore other methods. Why?

Can any ingredients of intelligent behaviour or any skill (or the processes that enables intelligent behaviour), not easily be described by a mapping from feature inputs to target outputs?

Researchers at OpenAI have stated publicly that short term AGI/HLMI is a serious possibility. “It is not possible

to determine a lower bound to progress in the near term and maybe the current way of progress will actually lead us to AGI”.

Do you agree or disagree?

If you disagree, why do you think proponents of such claims are missing/get wrong?

What do you think he would have to see/learn about, so that he would agree with you?

What must a chatbot be able to accomplish in the next two years so that you would maybe agree?

A lot of human behaviour and intelligent behaviour is mathematically and formally ill-defined. (Examples: creativity, social interaction and emotion.) Does that pose a problem for near term HLMI?

Which components to human cognitive skills do you think we don't need to build in order to engineer an intelligent agent? Which components might be redundant, side effects or irrelevant?

Expectation: “If no one has published XXX within the next 2 years, using a connectionist model, I will be really surprised.”

Please fill in XXX.

How would this change your mind?

What type of problem does the research community have to crack so that you start to believe HLMI could be achieved in 10 years, using connectionist architectures? You can name any domain you like.

What type of problem does the AI community have to crack in the next five years so that you start to believe HLMI could be achieved twice as quickly as you think it can be achieved now?

Do you have a hunch or intuition about hard barriers that deep learning will hit, that no one else seems to speak about?

What is your most plausible story for short term HLMI on the basis of deep learning.

What has to happen so that you think we should not continue approaching a task (e.g. one-shot, compositionality, planning, long-term memory, transfer learning) using ANNs?

Why, due to reasons related to the technique themselves, could progress toward HLMI using current techniques slow down? Assuming funding, talent, compute, etc. continues to progress?

What is so fundamental to the success of deep learning so that it is here to stay?

What are the tasks that you confidently think we will solve using ANNs? (it doesn't matter when we will solve them).

When would you stop trying to improve algorithmic performance (in a particular domain) using the connectionist approach? What would count as a “connectionist failure” to you?

What is the type of data you would have to see to believe that an HLMI algorithm (in a particular domain) cannot be built by scaling ANNs?

What result/published paper would you have to see in the next two years, in any domain of your choice, to think that we could reach HLMI 10 years earlier than you thought?

What papers/techniques you think look most promising to increase data efficiency of ANNs work training?

Is backpropagation here to stay?

Brains are able to do object recognition using much less energy than ANNs. What are the reasons for this difference?

Free generalisations of universals from little data—why are ANNs not doing this? Will ANNs ever be able to do this?

In biological and physical systems, we find inherent, unavoidable trade-offs.

What trade-offs are inherent to all ANNs works?

E.g. in physics: Between temporal and spatial resolution/Heisenberg uncertainty principle Biology: Virulence trade-off hypothesis, which is the trade-off between transmission rate and virulence (damage to host) of a virus.

Here are some of the points of critique that deep learning received: data inefficiency, object capture, causal reasoning, transfer learning, one-shot learning, meta-learning, common sense, concept formation, etc. Many researchers believe that these are current problems of deep learning but that that deep learning researchers will solve them eventually. Are there any current issues in deep learning that you don't expect us to make great improvements on? Why? What indicates that, for any of the above current limitations, we are unlikely to make progress?

Do you hold an opinion about deep learning that you think most other people don't hold?

Do you know of a valid criticism against DL, which you believe the community would not like to hear?

Is there a benefit or a potential of DL that you believe most researchers dismiss too easily?

Did AlphaGo/AlphaZero change your thinking about how long it might take to reach HLMI? Did it influence your estimates of what you think is difficult to achieve and what is easy to achieve?

Can every cognitive task be described with a utility function?

How many cost functions does the brain have? How do they interact?

How should we estimate the “computational power” of brains?

What ingredients to intelligence are, you think, easier than most cognitive science researchers assume?

What makes scalability in DL so possible? What features are responsible for its ability to scale or what features prevent it from scaling?

What can we not learn in a simulation?

What cannot be learned from observation, what needs interaction?

Theoretically, ANNs can approximate any function. This theoretical result presumes access to infinite time and memory. Why is this an important result for making practical engineering choices in the real world?

Do you think the ability of ANNs to generalise is surprisingly good or surprisingly bad?

Do you believe ANNs understand the concept of a dog? If yes, why do you believe this?

What ingredients to intelligence are, you think, harder than most machine learning researchers assume?

Can every cognitive task be described with a utility function?

Supervised learning, which has worked very well thus far, is able to learn well when target features are discrete categories. Can all cognitive skills be translated into a discrete categorisation problem?

What is the role of abstraction in intelligence and do ANNs do this?

What, according to you, are the most fundamental hurdles that we face in deep learning?

What are the skills, which we have not achieved, that you think are fundamental to a lot of ingredients of intelligence?

Appendix 2: Participant information sheet

Participant information sheet

Title of study: Expert Elicitation on Limitations of Deep Learning Contact:

Funded by: Berkeley Existential Risk Initiative Conducted at: University of Cambridge.

Purpose: Collect arguments from experts in computer science, cognitive science and philosophy, about the fundamental limitations of deep learning, via interviews.

Participation: Involves answering a series of questions regarding the potential and limitation of deep learning. The question will not require you to disclose sensitive information about yourself, your work or your employer.

No known risk is associated with participation.

Your participation is entirely voluntary and you can withdraw from the project at any time without prejudice, now or in the future.

You can indicate your preference regarding anonymity below.

Collected data will be the recorded interviews and notes from the interviews and will be stored securely, in compliance with University guidelines. The audio recordings will not be published.

Interviews will be processed manually. The transcripts/notes may be shared anonymously with other researchers for the purpose of research only.

Some parts of your responses may be included (anonymously if indicated on consent form) in a final publication that summarises the arguments collected over all interviews. The full interviews themselves will not be published.

An academic publication venue may require us to provide a list of the respondents. This list will not be published.

If you, at any point, have any further questions, please email:

Interview consent form

If you consent to being interviewed and to any data gathered being processed as outlined above, please print and sign your name, and date the form, in the spaces provided.

Please indicate, by ticking ONE of the boxes below, whether you are willing to be identified, and whether we may quote your words directly, in reports and publications arising from this research. (2.)

I/and my employer (cross out which does not apply) may be identified in reports made available outside the research and publication review teams, and in publications.

Neither I, nor my employer, may be identified in reports made available outside the research teams and publication review teams, nor in any publications. My words may be quoted provided that they are anonymised.

Neither I, nor my employer, may be identified in reports made available outside the research teams and publication review teams, nor in any publications. My words may not be quoted.

Please print your name:

Signature: ...

Acknowledgements We report no conflicts of interest. This work was financially supported by the Berkeley Existential Risk Initiative (BERI). The work was begun while in residency at the Leverhulme Centre for the Future of Intelligence and the Centre for the Study of Existential Risk and completed at the Future of Humanity Institute at the University of Oxford. Thanks to Seán Ó hÉigeartaigh and Shahar Avin for supervision on this project and to Luke Kemp, Jess Whittlestone and Alexa Hagerty for comments on drafts.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

- LeCun, Y., Bengio, Y., Hinton, G.: Deep learning. *Nature* **521**, 436–444 (2015). <https://doi.org/10.1038/nature14539>
- Jordan, M.I., Mitchell, T.M.: Machine learning: trends, perspectives, and prospects. *Science* **349**, 255–260 (2015). <https://doi.org/10.1126/science.aaa8415>
- McCarthy, J., Minsky, M.L., Rochester, N., Shannon, C.E.: A proposal for the Dartmouth summer research project on artificial intelligence, august 31, 1955. *AI Mag.* **27**, 12–12 (2006). <https://doi.org/10.1609/aimag.v27i4.1904>
- Sutskever, I.: Ilya Sutskever at AI Frontiers 2018: Recent Advances in Deep learning and AI from OpenAI. Progress towards the OpenAI mission LINK (2018). <https://sqlandsiva.blogspot.com/2019/01/day-202-ilya-sutskever-at-ai-frontiers.html>. Accessed 5 Apr 2021
- Ord, T.: *The Precipice: Existential Risk and the Future of Humanity*. Bloomsbury, London (2020)
- Liyanage, H., Liaw, S.-T., Jonnagaddala, J., Schreiber, R., Kuziemsky, C., Terry, A.L., de Lusignan, S.: Artificial intelligence in primary health care: perceptions, issues, and challenges: primary health care informatics working group contribution to the yearbook of medical informatics 2019. *Yearb. Med. Inf.* **28**, 041–046 (2019). <https://doi.org/10.1055/s-0039-1677901>
- Frey, C.B., Osborne, M.A.: The future of employment: how susceptible are jobs to computerisation? *Technol. Forecast. Soc. Chang.* **114**, 254–280 (2017). <https://doi.org/10.1016/j.techfore.2016.08.019>
- Crawford, K., Dobbe, R., Dryer, T., Fried, G., Green, B., Kaziunas, E., Kak, A., Mathur, V., McElroy, E., Sánchez, A.N., Raji, D., Rankin, J.L., Richardson, R., Schultz, J., West, S.M., Whittaker, M.: AI now report 2019.
- Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Allan Dafoe, A.: The malicious use of artificial intelligence—forecasting, prevention, and mitigation. <https://arxiv.org/ftp/arxiv/papers/1802/1802.07228.pdf> (2018)
- Russell, S.J.: *Human Compatible: Artificial Intelligence and the Problem of Control*. Penguin (2019) ISBN:978-0-525-55861-3
- Cantwell Smith, B.: *The Promise of Artificial Intelligence: Reckoning and Judgement*. MIT Press, Cambridge (2019). ISBN:9780262043045
- Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners. *ArXiv200514165 Cs*. (2020)
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. *ArXiv200108361 Cs Stat*. (2020)
- Hernandez, D., Brown, T.B.: Measuring the algorithmic efficiency of neural networks. *ArXiv200504305 Cs Stat*. (2020)
- McCandlish, S., Kaplan, J., Amodei, D., Team, O.D.: An empirical model of large-batch training. *ArXiv181206162 Cs Stat*. (2018)
- Amodei, D., Hernandez, D.: AI and compute. *OpenAI Blog*. <https://openai.com/blog/ai-and-compute/> (2018). Accessed 17 Dec 2020
- Hassabis, D., Kumaran, D., Summerfield, C., Botvinick, M.: Neuroscience-inspired artificial intelligence. *Neuron* **95**(2), 245–258 (2017). <https://doi.org/10.1016/j.neuron.2017.06.011>
- Banino, A., Barry, C., Uribe, B., Blundell, C., Lillicrap, T., Mirowski, P., Pritzel, A., Chadwick, M.J., Degris, T., Modayil, J., Wayne, G., Soyer, H., Viola, F., Zhang, B., Goroshin, R., Rabinowitz, N., Pascanu, R., Beattie, C., Petersen, S., Sadik, A., Gaffney, S., King, H., Kavukcuoglu, K., Hassabis, D., Hadsell, R., Kumaran, D.: Vector-based navigation using grid-like representations in artificial agents. *Nature* **557**, 429–433 (2018). <https://doi.org/10.1038/s41586-018-0102-6>
- Banino, A., Koster, R., Hassabis, D., Kumaran, D.: Retrieval-based model accounts for striking profile of episodic memory and generalization. *Sci. Rep.* **6**, 31330 (2016). <https://doi.org/10.1038/srep31330>
- Garnelo, M., Shanahan, M.: Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Curr. Opin. Behav. Sci.* **29**, 17–23 (2019). <https://doi.org/10.1016/j.cobeha.2018.12.010>
- Cranmer, M.D., Xu, R., Battaglia, P., Ho, S.: Learning symbolic physics with graph networks. *NeurIPS 2019 ArXiv190905862 Astro-Ph Phys. Phys. Stat.* (2019)
- Grace, K., Salvatier, J., Dafoe, A., Zhang, B., Evans, O.: Viewpoint: when will AI exceed human performance? Evidence from AI experts. *J. Artif. Intell. Res.* **62**, 729–754 (2018). <https://doi.org/10.1613/jair.1.11222>
- Müller, V.C., Bostrom, N.: Future progress in artificial intelligence: a survey of expert opinion. In: Müller, V.C. (ed.) *Fundamental Issues of Artificial Intelligence*, pp. 555–572. Springer International Publishing, Cham (2016)
- Technology Review, experts predict when artificial intelligence will exceed human performance. <https://www.technologyreview.com/2017/05/31/151461/experts-predict-when-artificial-intelligence-will-exceed-human-performance/> (2017)
- Gray, R.: How long will it take for your job to be automated? BBC (2017). <https://www.bbc.com/worklife/article/20170619-how-long-will-it-take-for-your-job-to-be-automated>
- Hong, L., Page, S.E.: Groups of diverse problem solvers can outperform groups of high-ability problem solvers. *Proc. Natl. Acad. Sci.* **101**(46), 16385–16389 (2004). <https://doi.org/10.1073/pnas.0403723101>
- Page, S.E.: *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools and Societies*. Princeton University Press, Princeton (2008)
- Landemore, H.: *Democratic Reason*. Princeton University Press, Princeton (2017)
- Tetlock, P.E., Gardner, D.: *Superforecasting: the Art and Science of Prediction*. Crown Publishers, New York (2015)
- Armstrong, J.S.: The seer-sucker theory: the value of experts in forecasting. Postprint version. Published in *Technology Review*, **82**(7), 16–24. http://repository.upenn.edu/marketing_papers/3 (1980)
- Chang, W., Chen, E., Mellers, B., Tetlock, P.: Developing expert political judgment: the impact of training and practice on judgmental accuracy in geopolitical forecasting tournaments. *Judg. Decis. Mak.* **11**(5), <http://journal.sjdm.org/16/16511/jdm16511.pdf> (2016)
- Tong, A., Sainsbury, P., Craig, J.: Consolidated criteria for reporting qualitative research (COREQ): a 32-item checklist for interviews and focus groups. *Int. J. Qual. Health Care.* **19**, 349–357 (2007). <https://doi.org/10.1093/intqhc/mzm042>
- DeJonckheere, M., Vaughn, L.M.: Semistructured interviewing in primary care research: a balance of relationship and rigour. *Fam. Med. Community Health.* **7**, e000057 (2019). <https://doi.org/10.1136/fmch-2018-000057>
- Krippendorff, K.: *Content Analysis: An Introduction to Its Methodology*. SAGE, Thousand Oaks (2013)
- Jamshed, S.: Qualitative research method-interviewing and observation. *J. Basic Clin. Pharm.* **5**, 87–88 (2014). <https://doi.org/10.4103/0976-0105.141942>

36. Hsieh, H.-F., Shannon, S.E.: Three approaches to qualitative content analysis. *Qual. Health Res.* **15**, 1277–1288 (2005). <https://doi.org/10.1177/1049732305276687>
37. Kondracki, N.L., Wellman, N.S., Amundson, D.R.: Content analysis: review of methods and their applications in nutrition education. *J. Nutr. Educ. Behav.* **34**, 224–230 (2002). [https://doi.org/10.1016/S1499-4046\(06\)60097-3](https://doi.org/10.1016/S1499-4046(06)60097-3)
38. Busch, C., Lynn, T., Kellum, R.: *Content Analysis*. Colorado State University, Fort Collins (2012)
39. Lake, B.M., Ullman, T.D., Tenenbaum, J.B., Gershman, S.J.: Building machines that learn and think like people. *Behav. Brain Sci.* **40**, 253 (2017). <https://doi.org/10.1017/S0140525X16001837>. (Epub 2016 Nov 24)
40. Buckner, C.: Empiricism without magic: transformational abstraction in deep convolutional neural networks. *Synthese* **12**, 1–34 (2018). <https://doi.org/10.1007/s11229-018-01949-1>
41. Crosby, M., Beyret, B., Halina, M.: The Animal-AI Olympics. *Nat. Mach. Intell.* **1**, 257–257 (2019). <https://doi.org/10.1038/s42256-019-0050-3>
42. Krishnan, M.: Against interpretability: a critical examination of the interpretability problem in machine learning. *Philos. Technol.* **33**, 487–502 (2020). <https://doi.org/10.1007/s13347-019-00372-9>

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