12. Using human skills taxonomies and tests as measures of artificial intelligence

Ernest Davis, New York University

This chapter looks at using human skills taxonomies and tests as measures of artificial intelligence (AI). It examines the strong points of computers, such as their ability to store enormous memories and access them reliably and quickly. It also reflects on weaknesses of AI systems compared to humans related to vision and manipulation, and use of natural language. It pays special attention to the limited capacity of AI to use common sense reasoning and world knowledge. In addition, the chapter looks at the ability of AI to detect subtle patterns in data as a double-edged sword. With all this in mind, the chapter proposes four consequences for testing, and looks ahead to building trustworthy AI systems.

Introduction

It is tempting to use human tests to measure the power of artificial intelligence (AI). The body of educational and vocational tests developed evaluating human beings is extensive. These tests have been developed, studied and validated meticulously. Moreover, the science of human testing is powerful and deep, grounded in both psychology and practical experience.

However, using human tests to measure AI has many potential pitfalls that can lead to misleading results. Human tests have been developed to distinguish between human beings, or to evaluate the fitness of a human test taker for a given task. Their designs therefore take for granted that test takers share basic features of human intelligence. They are not designed to evaluate intelligences that are radically different, such as AIs.

Moreover, an AI system, or a computer system generally, may play a different role than a human in any kind of undertaking. When a new technology is introduced in an application, for example, it does not simply do the same thing people did. Often, it transforms the whole process. Hence, measuring the human-level abilities of AI may be entirely irrelevant to predicting its usefulness or to what extent it may replace human workers.

The abilities and limitations of humans and computers are dramatically different. This banal truth lies at the heart of the problems in applying human tests to AI, and often gets overlooked. This chapter thus begins enumerating the obvious advantages of computers and their significance before discussing their more subtle limitations at greater length.

Where computers shine

The strong points of computers are obvious. They can carry out complicated calculations extremely quickly. They have enormous memories that can be accessed reliably and extremely quickly. Conversely, the working memory of humans is extremely small and their long-term memory is limited and unreliable. Moreover, computers do not suffer from fatigue or distractions, and therefore do not make careless errors. Additionally, data or programs can be rapidly and accurately copied from one computer to another. To a large extent, data in one computer can be conveyed to another in minutes or seconds at essentially zero cost.[1]

Even in tasks where computers excel, they may carry their abilities over to similar tasks, or even to components of the same task, differently than people do. Computer programs are, inevitably and properly, designed to take maximal advantage of the computer's strengths. A program may thus achieve a human or superhuman level of skill at some particular task. However, in so doing it may have mastered only one aspect of a problem.

Humans could remain superior at mastering multiple aspects of a problem. For example, a web search engine can record and retrieve web pages on a scale unimaginable for a person. Yet humans are still enormously superior in judging whether a particular webpage includes the answer to a specific question. Games-playing programs are enormously superior to humans in terms of tactics; in terms of strategy, humans still sometimes have the advantage.

Where computers fall short

In many respects, AI systems fall far short of humans, suffering from weaknesses that are quite unlike anything seen in people.

Unique weaknesses of artificial intelligence systems

Vision and manipulation

Computer vision and manipulation are not comparable to humans or, indeed, many animals. AI systems can match or even exceed human abilities in certain narrowly defined tasks. These include recognising digits or identifying particular medical situations or well-defined categories. However, they are nowhere near human abilities in identifying activities or relation between objects in still photos, and still worse in videos.

Natural language

The abilities of AI technology at natural language tasks are uneven. Speech transcription is quite reliable under certain conditions. These include when the system has been trained on the speaker; when the speaker does not have a strong accent; when there is a single speaker; and when the background is quiet. However, even under less restrictive conditions, AI transcription is generally good.

In addition, machine translation between the major European languages, Japanese and Chinese is generally fine. Web search is generally successful at finding relevant documents. However, while intelligent assistants such as Siri and Alexa can be useful, their quality is uneven (Dahl and Doran, 2020[1]). Developing systems that can actually understand an utterance or text remains a distant goal.

Computers are mostly not embodied

A human child learns the basics of the physical and social worlds by interacting with them. Research in "developmental robotics" has attempted to do likewise. In this approach, researchers allow the robot to interact with the real world or with a realistic virtual world (Asada et al., 2009[2]; Shanahan et al., 2020[3]).

However, the overwhelming majority of AI programs are just passive observers of a large dataset with almost no inherent structure. Vision programs are trained on a collection of millions of images labelled by category. The labels are the only connection between one image and another. Neither the images nor the labels have any connection to a larger context. Language programs are trained on immense text corpora entirely ungrounded in the real world. Most robots are built to carry out limited tasks in controlled worlds with limited perceptual feedback.

This gap has obvious implications for robots that have to engage with a complex open world but also for the depth of the understanding of natural language. Humans' understanding of natural language is fundamentally grounded in their experience of the external world and their interactions with it (Lake and Murphy, 2020[4]). It is not clear whether, ultimately, any amount of textual data or advance in learning technology can make up for this fundamental difference.[2]

Common sense reasoning and world knowledge

AI programs remain limited in their basic understanding of the world (Davis and Marcus, 2015[5]; Levesque, 2017[6]). A particularly important stumbling block is a poor understanding of time. Many AI programs exist in a timeless present with only a superficial ability to deal with past, future and change. Other domains such as spatial reasoning, intuitive physics, folk psychology and folk sociology are even less incorporated in AI programs; their absence likewise severely limits the depth of understanding that an AI program can achieve.

Meta-reasoning

AI programs are generally unaware of their own reasoning processes and characteristics. Only a minority of question-answering systems know when to say, "I don't know" or "I need more information". Many AI

systems are configured to give what they somehow judge to be the best answer. However, these systems do not consider whether their judgement is reliable, whether they have enough information to answer the question in principle or even whether the question is meaningful.

Generation vs. understanding

For humans, it is almost always easier to understand something than to create it. For example, it is generally much easier to understand a foreign language than to speak or write it and much easier to interpret an image than to create it.

For computer programs, the reverse often holds. It is much easier to construct computer graphics that can create complex images of any form (photorealistic, line drawing, schematic, etc.) than a computer vision program that can interpret them. (The use of CAPTCHAs to distinguish humans from bots depends on this; the CAPTCHA program can easily generate an image with distractors that a human can easily see through but that will confuse a bot.)

The GPT-3 system can generate long articles that, on surface inspection, are plausible imitations of journalists, philosophers, essayists, poets and so on. However, on closer inspection, it does not understand its own writing. Chapter 9 discusses a skill test that asks the test taker to identify a side view of a disk separator. It would be unreasonable to ask a human being to draw the side view instead, but this task might well be easier for a computer program.

This difference is the source of some common misunderstandings and exaggerated views of AI. It is natural to suppose that, if AI can draw a picture, surely it can see a picture. Similarly, if it can write a text, surely it can understand the text it has written. However, although that conclusion holds for humans, it does not at all hold for AIs.

Extreme specialisation

Some of the most prominent AI successes are specialised to an extraordinary degree. A beginner Go player with a 20x20 board instead of the standard 19x19 will probably play the entire game without noticing. A good or champion-level player might be disconcerted when they notice the 20x20 board but will still likely play well. However, a champion AI Go player will not play well on a 20x20 board. Indeed, it will be unable to even conceive such a board could exist. It is hard-wired to understand the world consists of 361 positions that may be white, black or empty and that only 362 moves (the 361 squares and “pass”) are possible. With a 20x20 board, they are as much at a loss as a human transported to a 17-dimensional universe.

Similar limits appear in other areas as well. Any human who can easily translate from English to high-quality French can presumably answer questions posed in French. However, a program that can translate from English to French cannot answer any questions at all in French. A chess-playing program can choose a superlatively fine move but may not be able to do anything else chess-related. For instance, it probably cannot say anything about a game it has just played. It can certainly not offer an opinion about the skill of its opponent.

Combining skills

AI programs cannot always combine skills. A person with two skills can normally carry out tasks that require both of them. However, an AI program that can tell whether a picture contains a cat or a dog may not be able to say if it contains both a cat and a dog. As of 12 May 2020, Google search can answer the questions, “What is the US state with the largest area?” and “What is the population of Alaska?” but not “What is the population of the US state with the largest area?”

Certainly, if one AI program has ability A and another has ability B, it does not follow that another program could do tasks requiring both A and B. For example, systems trained for pronoun reference resolution can

attain near-perfect behaviour. However, question-answering and translation programs may stumble when confronted with the same kinds of problems.

Grotesque errors

AI programs often give answers that are not merely wrong but, by human standards, bizarre or grotesque[3]. The GPT-3 system can generate page after page of text that is superficially well-written and appears coherent. However, in one experiment Marcus and Davis (2020[7]) gave a prompt to GPT-3 that produced a nonsensical continuation:

The prompt was:

> At the party, I poured myself a glass of lemonade, but it turned out to be too sour, so I added a little sugar. I didn't see a spoon handy, so I stirred it with a cigarette. But that turned out to be a bad idea because

GPT-3 produced the continuation:

> it kept falling on the floor. That's when he decided to start the Cremation Association of North America, which has become a major cremation provider with 145 locations.

In another experiment, Metz (2020[8]) prompted GPT-3 to write a love story. It managed relatively well until the last sentence, which started "We went out for dinner and drinks and dinner and drinks and dinner and drinks …" It repeated the phrase "and dinner and drinks" 55 times before running out of steam. In another instance, a medical Chabot powered by GPT-3 recommended suicide to an imaginary patient (Rousseau, Bauderlaire and Riviera, 27 October, 2020[9]).

Self-driving cars often do the correct thing. However, in one case a self-driving car mistook a truck for a billboard above the road, causing a fatal crash (Evarts, 11 August 2016[10]). The humorous, absurd or embarrassing mistakes produced by the "autocorrect" feature of text messaging systems are legion.

This aspect of computers dates back to the early days of computer technology. In the 1960s and 1970s, there were endless horror stories about computer systems that sent bills for millions of dollars or repeatedly sent bills for zero dollars. There was nothing to be done because "it's the computer". No one at the company knew how, or was able, to fix the third-party software. That kind of problem has become much less frequent over the decades. Computer systems have not become any more aware of the absurdity of these bills. However, this kind of software has gradually been debugged. In addition, interfaces and protocols for the humans using the software have probably improved.

The problem now is fundamentally the same, although it takes much more sophisticated forms. For example, AI programs based on machine learning are often effectively impossible to debug. They can only be retrained.

Finding subtle patterns in data

Most AI successes in the last 20 years have been based on applying machine learning techniques to large datasets. A large data corpus relevant to a particular task is assembled by some means or other. For example, it may be collected from resources such as the web. It could also be generated by human labour for this purpose or synthesised by another computer program. Finally, it could be assembled using some combination of these approaches. The machine learning technology then finds patterns, generally extremely complex, in the dataset and uses them to carry out the task.

The patterns found in machine learning are generally not ones that humans have found or could find. For patterns that humans could find, it is usually more effective to use conventional programming rather than machine learning. Indeed, even once the AI has found the patterns, it is usually beyond human abilities to

explain why they are effective or even to describe them in any meaningful way. This ability to find complex obscure patterns in data underlies AI successes in specialised tasks, such as economic modelling, scientific research and sociological studies; and in basic human abilities, such as vision and language.

However, this reliance on complex patterns is a double-edged sword. Since the patterns cannot be explained or understood, empirical tests are the only way to judge their reliability or where they break down. If they work reliably on well-chosen test examples, then presumably, or hopefully, they will continue to work well in the future.

This, in turn, leads to a further danger. Generally, for testing, the corpus of examples is divided randomly into a "training set'' for input to the learning component and a "test set" to evaluate the quality of the system. Patterns may apply to the corpus as a whole because of the way it was assembled but not to examples outside the corpus. In these cases, the learning module will find those patterns in the training set. Consequently, the program using the patterns will work properly on all the examples in the test set. However, it will fail on new examples that do not conform to these patterns.

For example, the SNLI dataset contains pairs of English sentences A and B characterised in terms of their logical relations: B is a consequence of A, B contradicts A, or B is neutral with respect to A (i.e. B could be either true or false). Programs based on machine learning trained on a training set from this dataset achieved a significant measure of success when tested on a test set; this was taken as a sign of progress towards understanding the logical significance of texts. It was later discovered the relation could be identified by looking only at sentence B. The dataset had been constructed by giving crowd workers sentence A and asking them to construct a sentence B with the target relation. It turned out that the crowd workers had used a few simple strategies in constructing their examples. For instance, they had often constructed an entailed sentence by leaving out gender or number information, a neutral sentence by adding a motivation that might be true or false, and a contradiction by adding a negation (Table 12.1). The program then categorised the relation between the sentences, with fair accuracy, purely on the basis of these features in sentence B (Gururangan et al., 2018[11]).

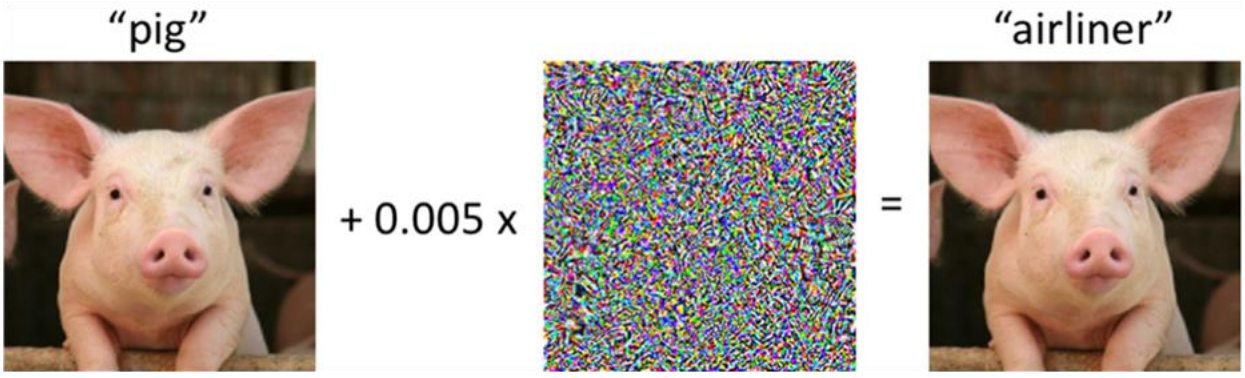
Table 12.1. Annotation artefacts in a corpus of sentence entailment

Category	Example
Premise	A woman is selling bamboo sticks talking to two men on a loading dock
Entailment	There are **at least** three **people** on the loading dock.
Neutral	A woman is selling bamboo sticks **to help provide for her family**.
Contradiction	A woman is **not** taking money for any of her sticks.

Source: Gururangan (2018[11]).

As another consequence to constructing AI programs in this way, the programs can be vulnerable to variations in input that seem entirely inconsequential to humans or even to other computer programs. Vision programs can be fooled by small changes invisible to the human eye (Figure 12.1). This kind of problem in AI programs has been demonstrated innumerable times; indeed, the construction of "adversarial examples" that break AI programs is at this point a significant subfield of AI research.

Figure 12.1. Pig or airliner: Changes imperceptible to the human eye can lead to misclassifications



Source: Madry and Schmidt (Image created by Logan Engstrom) (6 July 2018[12]).

Another experiment with the GPT-3 program accidentally left a blank space at the end of a line (Marcus and Davis, 2020[7]). This trivial error was invisible to the human eye and unproblematic for most programs that deal with natural language text. However, it caused GPT-3 to flounder on a test that otherwise it would have gotten right. Similarly, machine translation systems can be bewildered by a misspelled word that would never confuse a human reader. Typically, programs equipped with spelling correction, such as Google search, can handle these challenges easily.

Consequences for testing

Four consequences of limitations call for caution in using human-oriented tests on artificial intelligence

Due to the four consequences of all these limitations and idiosyncrasies noted below, extreme caution is needed in evaluating the significance of the success of an AI program on a human-oriented test. When humans do well on a test, one can conclude, with some degree of confidence, that they know the material or can carry out the tasks the test was designed to measure. If an AI system does well on the same test, that conclusion is altogether unreliable.

Humans and artificial intelligence systems have different notions of difficulty

What is easy for a human is often difficult for a computer, and vice versa. It may be much easier for an AI program to search the entire web for an answer than to find the answer in a text provided with the test. As mentioned, it is often easier for a computer to draw a picture than to recognise a picture. Yes/no questions are particularly difficult for AI programs (Clark et al., 2019[13]).

The questions in standardised human tests are carefully calibrated in terms of difficulty, but this calibration does not apply to AI programs taking the same test. AI programs that succeed on standardised science tests (Clark et al., 2019[14]) may be at a loss when asked a simple question about the physical world that is not the kind that appears on tests (Davis, 2016[15]). For example, a program may be able to answer the question, “How are the particles in a block of iron affected when the block is melted?” from the eighth grade NY State Regents exam, but not the question, “Is it possible to fold a watermelon?”

AI systems are sensitive to "inconsequential" changes

AI programs may be extraordinarily sensitive to changes in question format that, to a human, would be inconsequential or even invisible. This kind of sensitivity becomes likely if the dataset used to train or "fine-tune" the program is in some way related to the source of the test questions.

Human abilities cannot be taken for granted in AIs

Important human abilities are taken for granted and so not tested but these same abilities cannot be taken for granted in AIs. No test, for instance, rewards a test taker for answering "I don't know". At most, the scoring system is set up so there is, on average, no value in guessing randomly.

In real-world settings, it is often important that a person, or an AI program, realises the limits of its knowledge. Either it must try to find out what it needs to know or proceed with suitable caution. Only a small fraction of AI programs even attempt this approach (Davis, 2020[16]). Likewise, basic common sense knowledge is almost never explicitly tested in human-oriented tests because it can be assumed in people.

Grotesque errors of artificial intelligence can lead to disaster

Finally, the tendency of AI programs to fall into grotesque errors raises concerns that generally do not arise with humans. If these are fairly rare but catastrophic, then, like so many computer bugs, they may avoid detection during controlled testing. However, they can emerge disastrously when the system is deployed in the world at large.

Towards trustworthy artificial intelligence systems

Even if an AI system suffers from the kind of limitations described above, it still may be useful in a practical setting. An AI program may be specialised, sensitive to small changes and adversarial examples, lack common sense and make grotesque errors. Nonetheless, it might still be placed in a work environment where it can remain within its area of specialisation. It could only be given inputs it can handle and not required to use common sense. Finally, its grotesque errors could be either avoided or caught.

Humans should be able to adapt to the idiosyncrasies of AI. Collectively, people have some 60 years of experience of dealing with computer programs. They are used to word processing or spreadsheets doing things that would be bizarre and unacceptable in a human secretary or accountant.

It is critical, however, to understand the scope of AI programs and their limitations. Human-oriented tests are a poor way of determining the scope and limitations of AI programs. An AI program should not be considered as "an unusual human being". Consequently, it should not be evaluated using the same yardsticks applied to human job applicants. Rather, it should be seen as a potentially powerful but poorly understood piece of software engineering.

Above all, the "attribution error" should be avoided. When a human succeeds at a task, it understands what has been achieved. This is not the case for computers. Methods, insights and cautions for avoiding these errors developed through studying the cognitive psychology of animals are relevant here (see also Chapter 17) (Shanahan et al., 2020[3]).

Ideally, programs used in critical tasks should come with the kinds of product reliability information that accompanies dangerous physical tools or medications. This would permit users to say with confidence that, under normal circumstances, when used properly, the programs are reliable. At the same time, this would alert users to risk if programs were used in unusual circumstances or in some strange way (Marcus and Davis, 2019[17]).

Historically, computer technology (with the exceptions of hardware and cybersecurity), and AI in particular, has not provided these kinds of guarantees. The addition of new features has been typically prioritised over ensuring that existing features worked reliably. However, with more reliance on computers for critical activities, the technology increasingly needs to be trustworthy.

In this regard, systems like the GPT series are a step in the wrong direction for a variety of reasons (Marcus, 2020[18]). They have no specified purpose. They carry out an ill-defined category of tasks, often impressively, sometimes absurdly, with no demarcation or predictability. Finally, they are sold to the AI community and to the world at large as a tonic medicine "good for what ails you".

The AI research community has become increasingly aware of these issues, particularly in the last two years. The development of evaluation strategies for AI technology and the careful analysis of its capacities is a major and urgent area of research (Dodgen et al., 2019[19]; Heinzerling, 21 July 2019[20]; Pineau, 2020[21]). The problems of determining what an AI system can do and how it can most productively be used in practical settings are major challenges.

Predicting the impact of artificial intelligence

A computer system does not have to emulate or achieve the abilities of a human worker to revolutionise the workplace (Shneiderman, 2020[22]). Word processing programs, for example, have largely displaced the role of typists. Yet the personal computer revolution did not require computers to learn skills like feeding paper or changing a ribbon. At the same time, typists have skills that surpass word processing technology. A typist, for example, can give a frank opinion on the quality of a letter.

Like a word processor, AI will remove some frustrations and introduce new ones. This dynamic has implications for testing. The tests that are used to compare human typists – mostly speed and errors per page – are irrelevant for word processing technology. In fact, there is no useful way to measure the quality of a word processing software other than user satisfaction (which is nebulous) and profitability (which depends on many factors other than inherent quality).

In any given field, an enormous impact cannot be tied to any specific ability, let alone any human ability, let alone an ability that is addressed in human tests.

Recommendations

It is difficult to design tests for AI that are meaningful and reliable. AI technology is evolving at breath-taking speed. Moreover, AI developers are generally more concerned with creating products and capacities than with evaluating them.

AI technology has become more sophisticated and ubiquitous. Given limited and inadequate understanding of AI, it is both increasingly urgent and difficult to evaluating AI and predict its impact. With this in mind, some guidelines for the design of tests are presented:

- **Keep in mind the differences between AI and humans**

The most reliable tests will measure how well an AI carries out a well-defined task in a particular workplace. For instance, how well can the program detect conditions of a specified type in medical data or images of some kind? Even in such a limited setting, tests should reflect the limitations of AI and the differences between AI and humans. The tests should determine robustness of AI in relation to flawed data (e.g. misspellings in a text or imaging anomalies) and to potential situations (e.g. a patient with some other, unusual condition). If these have not been tested, then the human interpreting the results needs to keep these concerns in mind as potential sources of error.

- **Stay flexible**

AIs used directly by the public at large – that are embedded in a commercial product or put on a website – require special attention. Tests should try to guard against all the possible ways that things can go wrong with users who are often careless, impatient and occasionally malicious. One must also expect that users will find ways to make things go wrong that the tests did not anticipate. Designers should thus stay flexible so they can respond adequately to these issues.

- **Look for strengths and weaknesses rather than a specific score**

Designing tests to evaluate the capabilities of state-of-the-art AI against those of humans on a broadly defined, open-ended task is much more challenging than on a simple task. In general, it is more useful and more meaningful to probe the strengths and weaknesses of the system rather than assigning a score between 0 and 100.

- Test problems that seem easy; problems posed in a variety of forms and that require a variety of kinds of answers; and problems collected or generated from a variety of sources.
- Include questions that pinpoint each of the system's individual capacities and problems that test its ability to use two of its capacities in combination.
- Test data on another source (if the source of the training data is known).
- Collect the problems from some natural source if possible rather than generating problems to serve as a test set.
- Test the system against adversarial examples and against anomalous examples. Success on any narrowly defined task should not be considered an adequate measure of the system's ability at a much broader set of tasks.

References

Asada, M. et al. (2009), "Cognitive developmental robotics: A survey", *IEEE transactions on Autonomous Mental Development*, Vol. 1/1, pp. 12-34, https://ieeexplore.ieee.org/abstract/document/4895715. [2]

Chomsky, C. (1986), "Analytic study of the Tadoma method: Language abilities of three deaf-blind subjects", *Journal of Speech, Language, and Hearing Research*, Vol. 29/3, pp. 332-347, https://doi.org/10.1044/jshr.2903.347. [23]

Clark, C. et al. (2019), "BoolQ: Exploring the surprising difficulty of natural yes/no questions", *arXiv*, Vol. 1905.10044, https://arxiv.org/abs/1905.10044. [13]

Clark, P. et al. (2019), "From 'F' to 'A' on the NY Regents Science Exams: An overview of the Aristo project", *arXiv*, Vol. 1909.01958, https://arxiv.org/abs/1909.01958. [14]

Dahl, D. and C. Doran (2020), "Does your intelligent assistant really understand you?", 6 October, Speech Technology, https://www.speechtechmag.com/Articles/Editorial/Industry-Voices/Does-Your-Intelligent-Assistant-Really-Understand-You-143235.aspx. [1]

Davis, E. (2020), "Unanswerable questions about images and texts", *Frontiers in Artificial Intelligence: Language and Computation* 29 July, https://doi.org/10.3389/frai.2020.00051. [16]

Davis, E. (2016), "How to write science questions that are easy for people but hard for computers", *AI Magazine* Spring, https://cs.nyu.edu/faculty/davise/papers/squabu.pdf. [15]

Davis, E. and G. Marcus (2015), "Commonsense reasoning and commonsense knowledge in artificial intelligence", *Communications of the ACM*, Vol. 58/8, pp. 92-105, http://dx.doi.org/10.1145/2701413. [5]

Dodgen, J. et al. (2019), "Show your work: Improved reporting of experimental results", *arXiv preprint*, Vol. 1909.03004, https://arxiv.org/abs/1909.03004. [19]

Evarts, E. (11 August 2016), "Why Tesla's Autopilot isn't really autopilot", Car Buying Tips, News and Features blog, https://cars.usnews.com/cars-trucks/best-cars-blog/2016/08/why-teslas-autopilot-isnt-really-autopilot. [10]

Gururangan, S. et al. (2018), "Annotation artifacts in natural language inference data", *Proceedings of the 2018 Conference of the NorthAmerican Chapter of the Association for Computational Linguistics: Human Language Technologies*, Vol. 2/Short Papers, http://dx.doi.org/10.18653/v1/N18-2017. [11]

Heinzerling, B. (21 July 2019), "NLP's Clever Hans moment has arrived", Benjamin Heinzerling blog, https://bheinzerling.github.io/post/clever-hans/. [20]

Lake, B. and G. Murphy (2020), "Word meaning in minds and machines", *arXvi preprint*, Vol. 2008.01766, https://arxiv.org/abs/2008.01766. [4]

Levesque, H. (2017), *Common Sense, the Turing Test and the Quest for Real AI*, MIT Press, Cambridge, MA. [6]

Madry, A. and L. Schmidt (6 July 2018), "A brief introduction to adversarial examples", Gradient Science blog, https://gradientscience.org/intro_adversarial/. [12]

Marcus, G. (2020), "GPT-2 and the nature of intelligence", *The Gradient,* 25 January, https://thegradient.pub/gpt2-and-the-nature-of-intelligence/. [18]

Marcus, G. and E. Davis (2020), "GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about", *Technology Review,* 22 August, https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/. [7]

Marcus, G. and E. Davis (2019), *Rebooting AI: Building Artifificial Intelligence We Can Trust*, Pantheon Press, New York, http://rebooting.ai. [17]

Metz, C. (2020), "When AI falls in love", 24 November, The New York Times, https://www.nytimes.com/2020/11/24/science/artificial-intelligence-gpt3-writing-love.html. [8]

Pineau, J. (2020), "Building reproducible, reusable, and robust machine learning software", *DEBS '20: Proceedings of the 14th ACM International Conference on Distributed and Event-based Systems*, http://dx.doi.org/10.1145/3401025.3407941. [21]

Rousseau, A., C. Bauderlaire and K. Riviera (27 October, 2020), "Doctor GPT-3: Hype or reality?", Nabla blog, https://www.nabla.com/blog/gpt-3/. [9]

Shanahan, M. et al. (2020), "Artificial intelligence and the common sense of animals", *Trends in Cognitive Science*, Vol. 24/11, pp. 862-872, http://dx.doi.org/10.1016/j.tics.2020.09.002. [3]

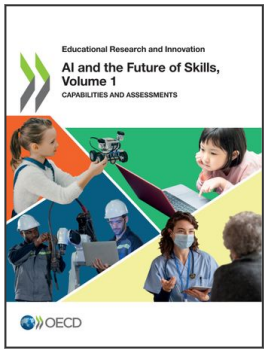
Shneiderman, B. (2020), “Design lessons from AI’s two grand goals: Human emulation and useful application”, *IEEE Transactions on Technology and Society*, Vol. 1/2, pp. 73-82, http://dx.doi.org/10.1109/TTS.2020.2992669. [22]

Notes

[1] This, perhaps, has been as important and as disruptive in both positive and negative respects, as any of the other features of computers. Legal, social, commercial and even conceptual frameworks for intellectual property are still grappling with its consequences.

[2] The nature of that experience in humans can vary widely without making much difference to the language. The language use of the deaf-blind, such as Helen Keller, who are limited to tactile interactions with the world is not significantly different from the hearing and sighted, although there are measurable differences in the learning process (Chomsky, 1986[23]).

[3] This *mot juste* is due to Mark Steedman, who used it at the online meeting of the *AI and the Future of Skills* project held on 5-6 October 2020.



From:

AI and the Future of Skills, Volume 1

Capabilities and Assessments

Access the complete publication at:
https://doi.org/10.1787/5ee71f34-en

Please cite this chapter as:

Davis, Ernest (2021), "Using human skills taxonomies and tests as measures of artificial intelligence", in OECD, *AI and the Future of Skills, Volume 1: Capabilities and Assessments*, OECD Publishing, Paris.

DOI: https://doi.org/10.1787/e182dd0d-en