An AI toolkit for libraries

Now that artificial intelligence (AI) tools are being widely used across academic publishing, how can we make informed assessments of these utilities? There is a need for a set of skills for evaluating new tools and measuring existing ones, which should enable anyone commissioning or managing AI utilities to understand what questions to ask, what parameters to measure and possible pitfalls to avoid when introducing a new utility. The skills required are not technical. Potential problems include bias in the corpus, a poor training set or poor use of metrics for evaluation. This article gives a quick overview of some of areas where AI tools are being used and how they work. It then provides a checklist for assessment. The goal is not to discredit AI, but to make effective use of it.

Keywords

AI; NLP; evaluation; metrics; research support

MICHAEL UPSHALL
Consultant

Introduction

A colleague walks up excitedly to you. 'I've just discovered a really cool AI app to speed up the submissions process for my articles – it does what we currently do by hand twice as quick, and all you have to do is to press a few buttons! Check it out!' How should you respond? The interface certainly looks well designed and appealing. There is not much information on the site, but the developers seem to have thought of everything. Do you feel qualified to give an opinion on this tool?

The aim of this article is to outline a framework for evaluating artificial intelligence- (AI-) based tools, without the need to have or to acquire detailed technical knowledge of how they were developed or any requirement to understand computing languages such as Python, or indeed any advanced maths. Nonetheless, the criteria for evaluation described here are crucial for the successful use of AI. The article argues that the human users may in some cases be better placed to evaluate the capabilities of a tool than the original developers, who quite possibly were not in a position to appreciate the context in which it would be used.

'a framework for evaluating artificial intelligence- (AI-) based tools'

AI tools can possibly provide a way to reduce the time taken to discover or to submit academic content; they have the potential to improve the quality of published articles by running more detailed and more accurate checks in advance of publication. However, it is not the intention of this article to provide arguments for or against the use of AI tools compared with human evaluation. Instead, the aim is to outline how such tools can be assessed in a real-world setting. There are already many tools making use of AI in our daily lives, but we don't always realize that AI is involved. For many of these tools, for example some of the components of the Google search engine, their introduction was without debate, and we have subsequently become accustomed to their strengths and weaknesses.

Like many new technologies, AI has been viewed through widely different perspectives during its long lifetime, dating back over 75 years to the 1950s[1] – from wild optimism to being written off. Unfortunately, both attitudes are wide of the mark, and neither extreme

is helpful for a balanced appraisal of AI tools. Repeatedly during AI's lifetime, many highly regarded thinkers have described AI either with glowing optimism, or expected AI to bring about disaster:

> 'Artificial intelligence will reach human levels by around 2029.' Ray Kurzweil.[2]
>
> 'The development of full artificial intelligence could spell the end of the human race.' Stephen Hawking, BBC Interview 2014.[3]
>
> 'At some stage therefore, we should have to expect the machines to take control.' Alan Turing, draft of lecture, 1951.[4]
>
> 'The internet makes us superficial.' Nicholas Carr, *The Shallows*, a 2010 book that has been cited 265 times.[5]
>
> 'With artificial intelligence we are summoning the demon.' Elon Musk.[6]

Perhaps more alarmingly, because they write from a perspective of widespread use of AI, several practitioners experienced in AI have recently written books that stress the negative implications of AI. Kate Crawford in her *Atlas of AI*[7] wrote about how AI creates low-paid work. Cathy O'Neill described the effects of AI bias on school selection panels in her *Weapons of Math Destruction*.[8] Erik Larson complains about the overextension of AI in an unthinking way in his *The Myth of Artificial Intelligence*.[9] You could be forgiven for thinking from these books that AI is an unmitigated disaster. Even the triumphs seem to turn out to be heavily qualified: we have been promised self-driving cars for many years but, apart from a relatively small number of controlled trials, self-driving cars have still not become commonplace. Perhaps this is simply technology evolving faster than we think in the long term but more slowly in the short term, or perhaps this is an inherent limitation of AI.

In the light of these warnings, should we be implementing AI tools or calling for more investigation? Should we abandon AI, or use it intelligently? The suggested solution here is to concentrate on a small subset of available AI tools and to follow a clear methodology for assessing them.

'Should we abandon AI, or use it intelligently?'

What AI tools consist of

The narrow, or limited, AI described here is based around a few components. Present-day AI for text tools for academic purposes typically comprises:

- a corpus
- a training set
- a test set
- an algorithm.

The 'corpus' is the body of content that you wish to analyse, for example, all scientific research articles published in the last 20 years. The corpus contains some information or characteristic that you wish to extract. A corpus need not only be text – there are corpora for facial recognition, for example, as well as the often referred to collections of images of cats and dogs.

The 'training' set is a subset of the corpus, which has been tagged in some way to identify the characteristic you are looking for. Thus, a training set for cats and dogs might be 100 images that were tagged by a human as one or the other. Another example is the Modified National Institute of Standards and Technology (MNIST) database of handwritten numbers,[10] which shows many examples of the range of styles used when humans write numbers by hand, see Figure 1.

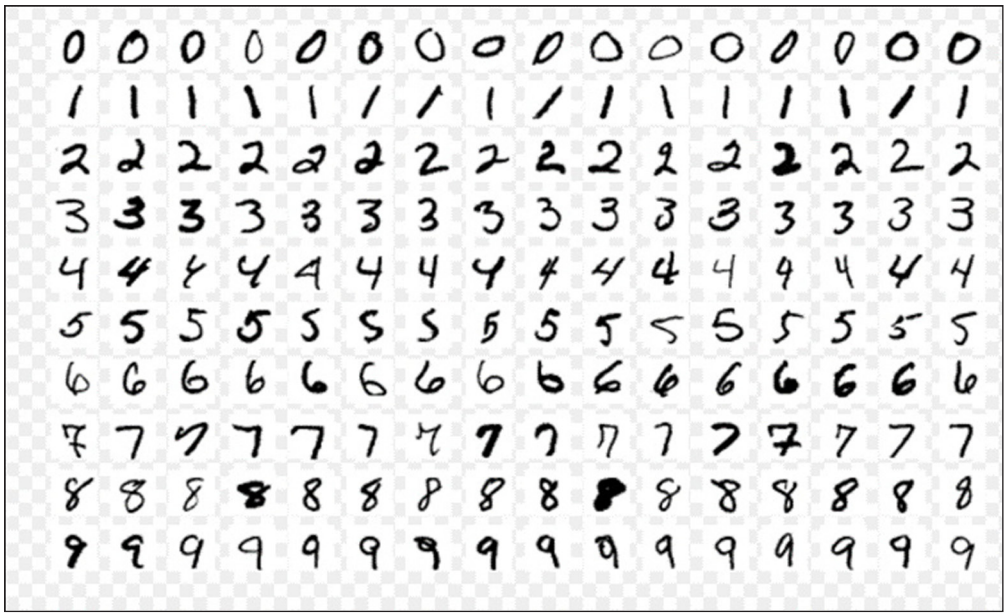

Figure 1. Example of the MNIST database of handwritten numbers

The 'test set' is the collection of documents to be used for trialling the algorithm, to see how successfully it carries out the operation. The 'algorithm' is simply the tool that looks at each item in the corpus and enables a decision to be made. An algorithm may be as simple (and frequently is as simple) as matching a pattern; so, for example, if you give the device ten handwritten examples of the numbers 0 to 9, then the machine is asked to find the closest match between the training set and the test set. A cookery recipe is an algorithm, as is a way of sorting documents – by date, or by subject. Much of computing is based around identifying the most effective algorithm to solve a specific problem, for example, how to sort a collection of numbers into numerical order.

Surveys suggest that the public has a low awareness that algorithms are being used.[11] Worse, there is a common misconception that when algorithms are used, they are the cause of any defects of the tool. One of the myths about present-day AI is that it is entirely about algorithms. If only algorithms were revealed in public, is the argument, then all the mysteries of AI would be revealed.[12] Reports in the media have tended to reinforce this misconception, accusing the algorithm of creating the problem; a 2019 study by the European Parliament on algorithmic accountability does not mention the term corpus.[13] Newspaper reports about using an algorithm to determine the results of public examination suggested the algorithm was the cause of the issue, rather than the (undocumented) way it had been implemented: 'We all remember the A-levels fiasco, when an algorithm decided what the results should be … the poorest students received worse marks'.[14]

'One of the myths about present-day AI is that it is entirely about algorithms'

More exactly, the success or failure of AI is as much based on the corpus as it is on the algorithm. If the corpus used has an imbalance of gender, ethnic group or geographical origin, then the algorithm will simply replicate that bias.

To summarize, artificial general intelligence raises many issues that, to be honest, are of little relevance to most present-day AI, even though they will keep leading-edge researchers busy for years. Narrow AI tools can, if implemented sensibly, greatly enhance our ability to carry out many of the tasks in the academic workflow. How these tools are selected and implemented is all-important. How can these tools – and the corpora they are based on – be evaluated? The role of the library is crucial in providing guidance on real-world selection, implementation and, finally, appraisal and metrics.

What is present-day AI?

For many years, AI researchers have been obsessed with creating AGI: artificial general intelligence. One algorithm could answer all the questions in the universe. The idea behind a universal algorithm is, in the words of AI researcher Pedro Domingos,

> 'If it exists, the Master Algorithm can derive all knowledge in the world—past, present, and future—from data. Inventing it would be one of the greatest advances in the history of science.'[15]

The idea of a universal general intelligence was widespread in the 1960s. It fell out of favour for several years, but traces of it are still evident today in research departments. According to Wikipedia,[16] there were 72 active AGI projects running in 2020, which indicates that many researchers continue to look for a unified solution via the use of AI, rather than making use of limited tools in specific contexts, which is what this article is concentrating on.

For the purpose of this article, the master algorithm will be ignored and the focus will be a smaller set of tools, typically employed for just one purpose. Formally, the tools described here make use of what is called 'supervised' or 'semi-supervised' machine learning.[17] Supervised means there is some human involvement in setting up the tool, usually in determining what the correct answers should be. 'Machine learning' (ML) means the use of a computer to follow a pattern, whether or not the pattern is identified by a human. 'Natural language processing' or NLP means the identification of patterns in spoken or written text.

Do we know that AI is being used?

This is a more fundamental question than might be imagined. There are many examples of AI tools in use without any mention that AI is being used, although, increasingly, the impact of the AI tool might be too subtle to notice. For example, in a Google blog post about BERT,[18] an ML technique for NLP, the benefit shown was simply the ability to link a preposition with a noun. Whereas earlier search tools tended to ignore prepositions and just focus on nouns, this more sophisticated tool was able to handle a question about a traveller from Brazil to the USA. It identifies a meaningful connection between the 'to' and the 'USA'.

'There are many examples of AI tools in use without any mention that AI is being used'

In social media and product literature, the term AI is frequently used as a buzzword to give the impression that a tool is more sophisticated than it really is. In practice, the kind of small-scale AI described above is very closely linked to 'string matching' or other well-established simple techniques. String matching means the use of a machine to identify instances of a sequence of characters in a text.[19] Eslami claims that once users are shown they are interacting with an algorithm, rather than with a human, they are reassured; there certainly appears to be widespread suspicion of an algorithm making decisions for a human. Not revealing that there is no human involved makes things much worse; the users feel cheated, because they were not told. Google search is an example where we as everyday users acknowledge that a perfect search experience is not possible, given the size and limitations of the corpus, and we tolerate the imperfections because we are not aware of any better alternative. As two researchers put it, 'College students AND professors might not know that library databases exist, but they sure know Google'.[20]

Can we combine the brain with technology?

Machines cannot think, but humans can. One way to assess AI tools is to determine what they are or are not good at. Some human activities lend themselves to automation more than others. Sorting a list into alphabetical or numerical order, for example, is an activity that a spreadsheet can do very easily, but humans can only do slowly and at a high error rate – partly because humans have a limited attention span and find it irritating to sort more than a few records. Does that mean the human brain is inadequate? Hardly, but it does imply that

human brains do not represent the ideal that all AI research is aimed at emulating. Similarly, humans have very poor information retention skills – we think ourselves clever if we can remember ten phone numbers. Miller's Law,[21] created by a Harvard psychologist, suggests that the number of objects a typical human can hold in memory is just seven.

What can we meaningfully ask of AI?

Questions we ask of AI tools may have different criteria to scientific research questions. The corpus-based approach using a training set, as described above, uses the process of inductive reasoning. This is the kind of thinking that states 'the sun rose yesterday, the sun rose today, so the chances are the sun will rise tomorrow'. Now, philosophers will argue that inductive reasoning is not scientific. Just because the sun rose yesterday does not mean the sun will rise tomorrow. We would like some external proof to enable us to sleep more peacefully. Inductive reasoning is well described by Eric Larson.[22]

However, for the purpose of AI tools as described here, inductive reasoning may be adequate, indeed ideal. The goal is to use existing evidence to predict a likely inference. Typically, we look to provide good quality results that are better than a human could achieve without the tool. 'Better' here meaning at least as good quality as a human but delivered faster, or better quality with no loss of time compared to a manual process, or both. Hence, for the purpose of AI in this context, you can ask an algorithm if the sun will rise tomorrow, and the machine will give you a workable answer for practical purposes.

To state that narrow AI tools make use of inductive reasoning may seem obvious, yet it is frequently ignored when humans assess the results of a machine-based process. For self-driving cars, an error rate of one in a thousand might mean abandoning the whole project. For spam checking and spell-checking, a much higher error rate may be good enough to use the tool.

Are we AI literate?

Long and Magerko[23] define AI literacy as 'a set of competencies that enable individuals to critically evaluate AI technologies, communicate and collaborate effectively with AI, and use AI as a tool'. Here is an essential role for the information professional. Millions of people use Google every day, but there is a difference between unthinking use and critical awareness. They further define over 30 relevant factors, of which just the first five are skills I believe to be essential to the assessment and recommendation of AI tools:

1. The ability to distinguish between tools that use or do not use AI.
2. Analyse differences between human and machine intelligence.
3. Identify various technologies that use AI.
4. Distinguish between general and narrow AI.
5. Identify problem types that AI excels at and problems that are more challenging for AI.

To be specific, the skills outlined here do not, I believe, require the ability to code. Given the increasing use of AI tools, it is becoming more difficult to distinguish tools based on human or machine judgement (skill 1). Perhaps this skill will eventually be replaced by skill 5, the ability to identify problem types that lend themselves to an AI-based solution.

'the skills outlined here do not … require the ability to code'

AI use in academic contexts

This section looks at some areas where AI tools are currently in use in the scholarly workflow.

Spell-check

The spell-check tool provided with many common word processors is an example of a widely used and generally accepted algorithm, or collection of algorithms. Users acknowledge (and frequently complain) that spell checkers do not detect all errors that would be detected by a human.

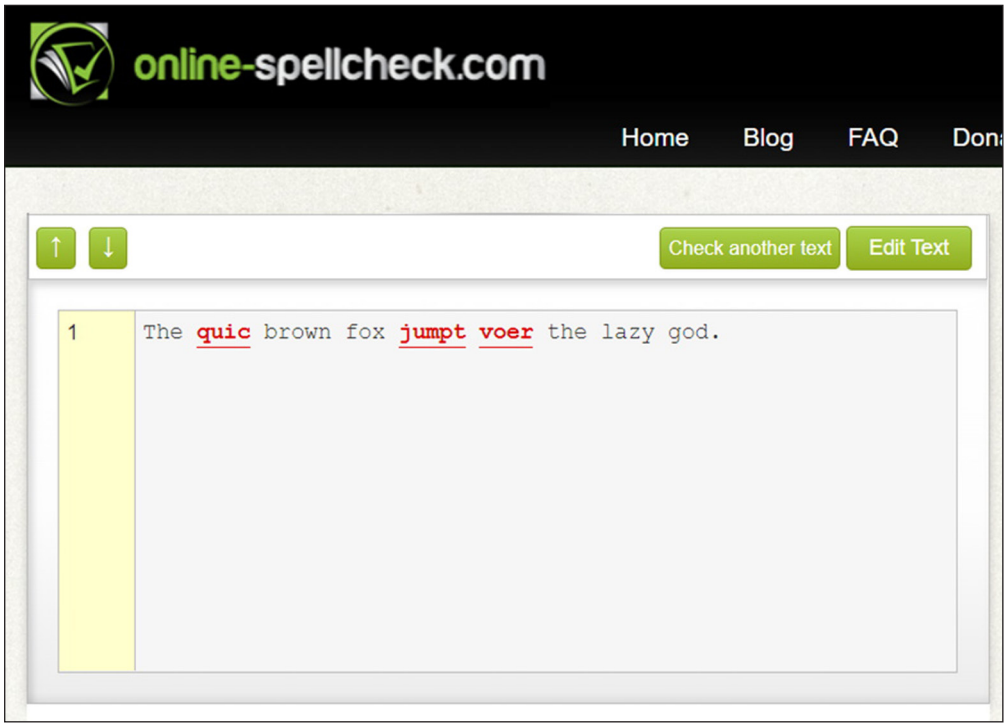

Figure 2. A typical spell checker displaying the limitations of a context-free tool[24]

Users of spell checkers have learned to live with their biggest drawback: that most spell checkers accept a word that corresponds with a term in the dictionary, even if it is the wrong word in context. As shown in Figure 2, the spell checker had no difficulty finding a misspelling for 'quick', but any English speaker would know that the last word should be 'dog', not 'god'. Nonetheless, the use of a spell checker can comprehensively and consistently identify transposed letters in words. The limitations are known and tolerated.

Spam check

Checking for spam e-mails is one of the most widespread uses of AI. Spam checks use a mixture of word- and phrase-checking to identify a likely spam message. A variety of checks are run, including:

- Is this an unfamiliar sender ID?
- Does the e-mail include terms such as 'offer' or 'bargain'?

As with spell-checking tools, spam checkers are imperfect but widely accepted, because the alternative, of repeatedly reviewing and deleting irrelevant emails, would make the use of e-mail difficult if not impossible. Users tolerate the small number of false positives (e-mails wrongly identified as spam). Jenna Burrell[25] differentiates various kinds of opacity in algorithms and reveals some interesting details about the criteria used to detect spam, but does not mention the corpus dimension of spam checking: e-mails from an e-mail address not in the individual's set of e-mail contacts is more likely to be spam.

Plagiarism detection

Plagiarism detection, such as Turnitin, Copyleaks and others, can use string or semantic matching, or both. The most common form is simply checking for string matches. A simple plagiarism check can be done against published articles by simply searching for a string, such as a full article title, in Google – the system typically finds a match (if one exists) in less than a second, see Figure 3.

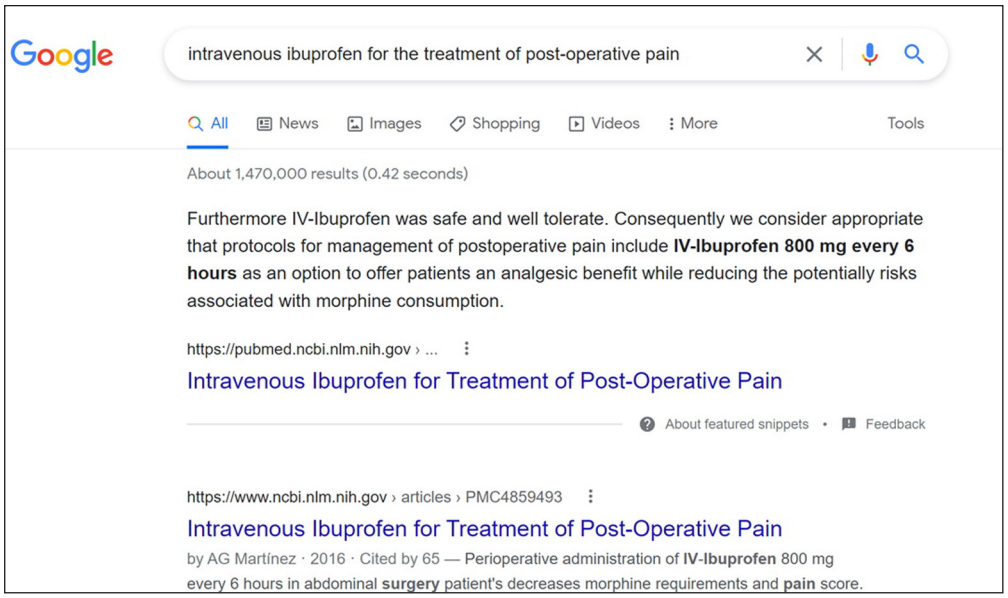

Figure 3. Google search for an article title

In recent years, plagiarism checks have become more sophisticated. While most common search engines can find strings of characters very efficiently, it is more challenging to find semantic matches. If the plagiarist uses the same ideas but changes a few of the words for common synonyms, the plagiarism is (currently) far less likely to be detected. This limitation does not prevent plagiarism tools being widely used by many academic publishers.

Discovery

One of the longest-established uses of AI is in content discovery. This can range from the simple recommender tool ('if you like X, you will like Y') to much more sophisticated recommenders that identify concepts in an article and match those concepts to other articles. Figure 4 shows an example from the Cambridge University Press content collection, linking book chapters to other book chapters and to articles:[26]

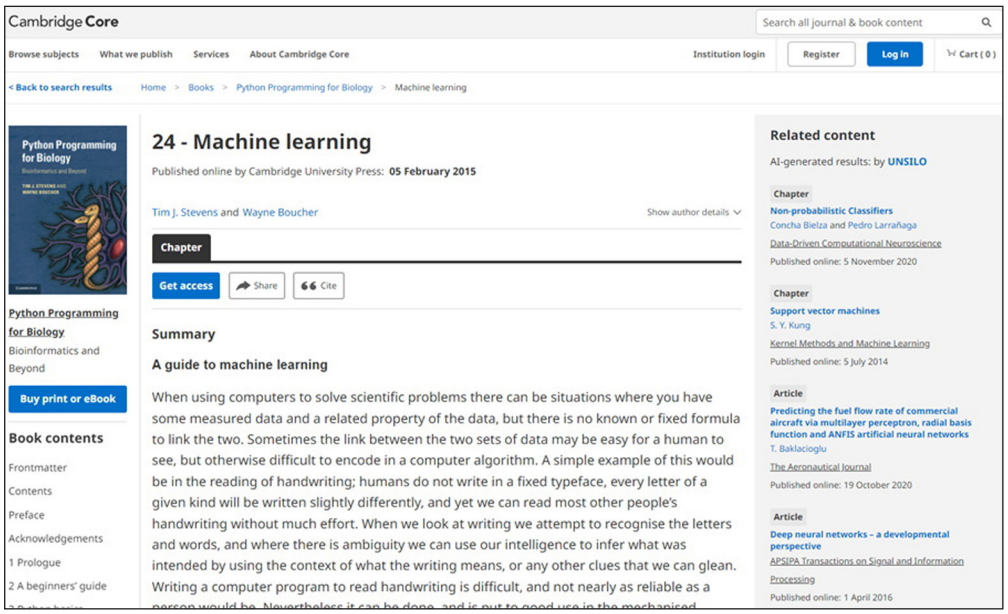

Figure 4. Example of a recommender system in action

Elaborations of the discovery tool include an alerting service, which finds new articles, essentially by replicating the search with a date filter, to look only for articles published on a subject in the last six months or six weeks on a topic. Recommender tools are common on most academic discovery platforms.

Impact

Citations are one way of assessing the relative worth of an article – if it has been widely cited, it must be significant. Citation indexes for academic journals were introduced in 1955 by Eugene Garfield.[27] But, of course, citations are contentious as indicators of quality. An article might be cited because the writer thinks the source article is incorrect. Citations for certain types of articles, such as review articles, are always higher than for research articles.

One reason why citations became widely adopted as a metric for article quality is that they can be counted. A human judgement ('is this article significant?') is thereby represented, however imperfectly, by an arithmetic tool. However, it is now recognized that a simple count of citations is unsatisfactory, and several modifications of the tool have been proposed, for example, the Hirsch or H-index.[28]

'AI tools have enabled a more sophisticated analysis of citations'

AI tools have enabled a more sophisticated analysis of citations. Tools are available, for example from scite.ai,[29] Scholarcy[30] (see Figure 5) and Semantic Scholar,[31] that not only identify citations, but show if the citation supports or refutes a statement.[32]

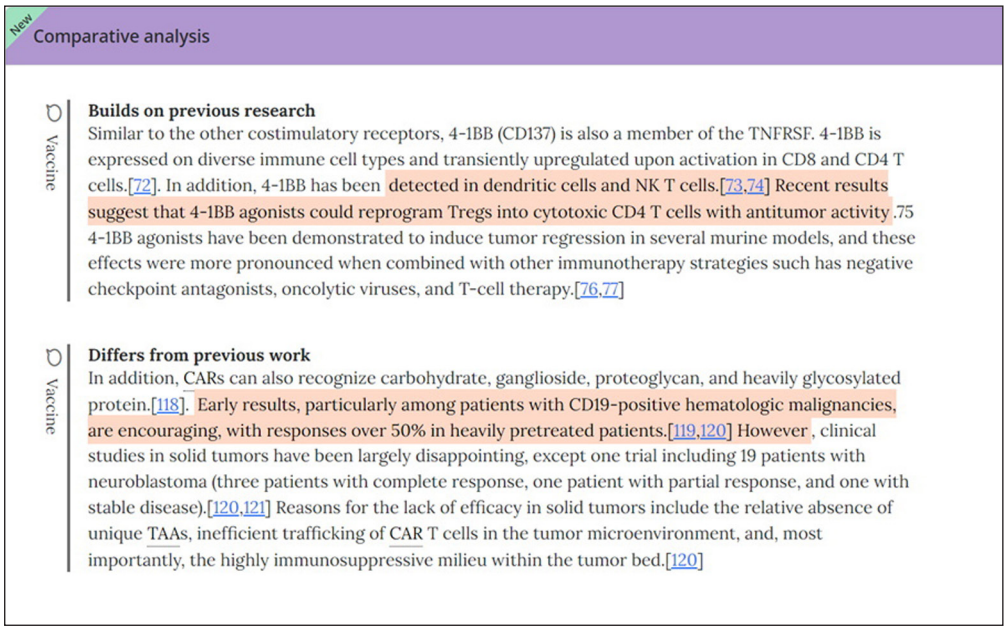

Figure 5. Categorizing citations into supporting or differing from earlier research[33]

Currently, few researchers will be aware of the wealth of tools available to them in this area.

The need for analytics

Any intervention in the academic workflow can only be assessed for robustness, efficacy and accuracy, if its impact is evaluated. This is as true for AI tools as for any other attempt to improve the process. Accordingly, both libraries and publishers have a responsibility to identify if, and how, AI tools are used, with an attempt to identify the impact of those interventions. However, many libraries do not carry out such studies. According to a 2021 survey of library analytics practice[34] the greatest barriers to data analysis (interpreted broadly to include bibliometrics, studies of user behaviour and such like) by libraries were:

- 61% lack of time
- 54% lack of expertise
- 52% lack of personnel

Unfortunately, all these justifications for inaction are ultimately self-defeating. If a poor-quality AI utility is adopted by the library, it will take more, rather than less time to manage its use and quite possibly lead to misunderstanding and corresponding negative feedback. Introducing a tool without capturing the data to assess its success cannot be a sensible procedure.

Automating metadata

Machines have difficulty with ambiguity, while humans are tolerant of inconsistencies and small errors. However, increased use of AI has partially resolved this distinction. Today, keying 'shakspeare' into Google results in the tool automatically suggesting the closest match from its index, see Figure 6.

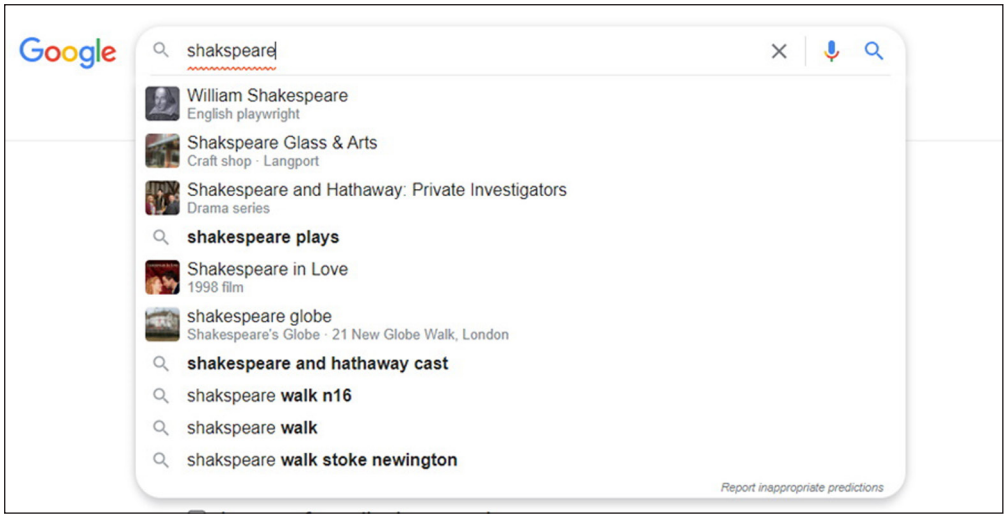

Figure 6. Google search 4 May 2022

This is rather similar to the way we tolerate a high level of incompleteness and errors in spoken dialogue – we guess what a speaker is trying to say. Similarly, the 'search ahead' feature in Google, see Figure 7, and other search engines attempts to guess what the user intends:

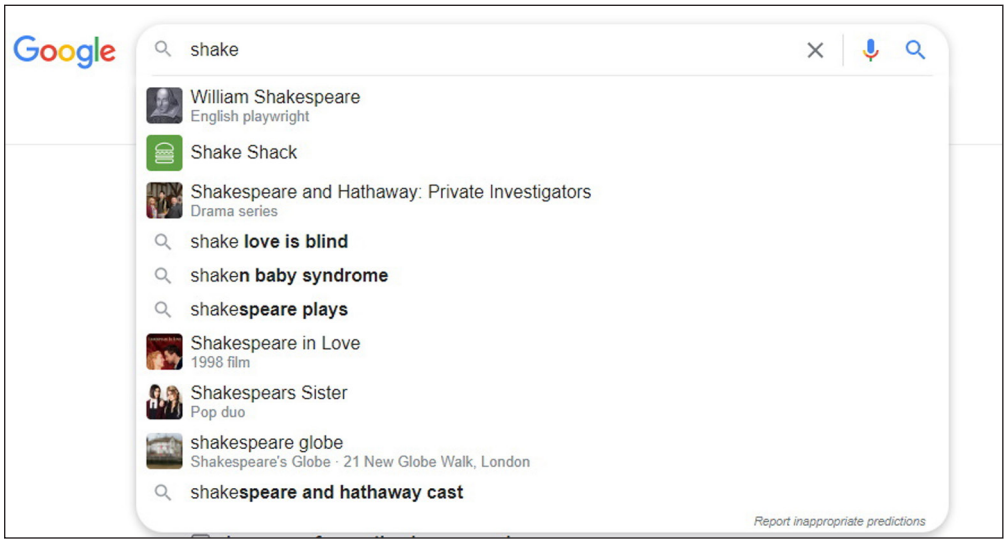

Figure 7. Google predictive search 4 May 2022

The system may be correct, or it may on occasion be wrong, but we tolerate its errors because most of the time it works and saves us effort. Humans have learned to live with the imperfect world of AI.

'Humans have learned to live with the imperfect world of AI'

Potential misuses of AI

Once developed, AI tools can be extended to domains where their validity is greatly reduced. There are several examples of this overextension of AI tools, with predictably irrelevant or meaningless results. A ranking website, Academic Influence, provides metrics for ranking departments, scholars and faculties with the slogan 'better rankings for a better education'. The site also contains a list of 'the most influential people

for the years 4000 B.C. to 2022', with Aristotle in first place, ahead of Plato, Marx and Kant. Shakespeare is listed in sixth place, with his plays and sonnets listed under his 'academic contributions'.[35] In this case, a tool developed for the comparison of current scholarship has been overextended back by several hundred years to a time before scholarly communications existed, see Figure 8 – and yet the site claims that their metrics are built using 'sanity checks': 'we make sure the rankings make sense by performing "sanity checks" against other independent information sources such as periodicals, journals, and global media outlets'.

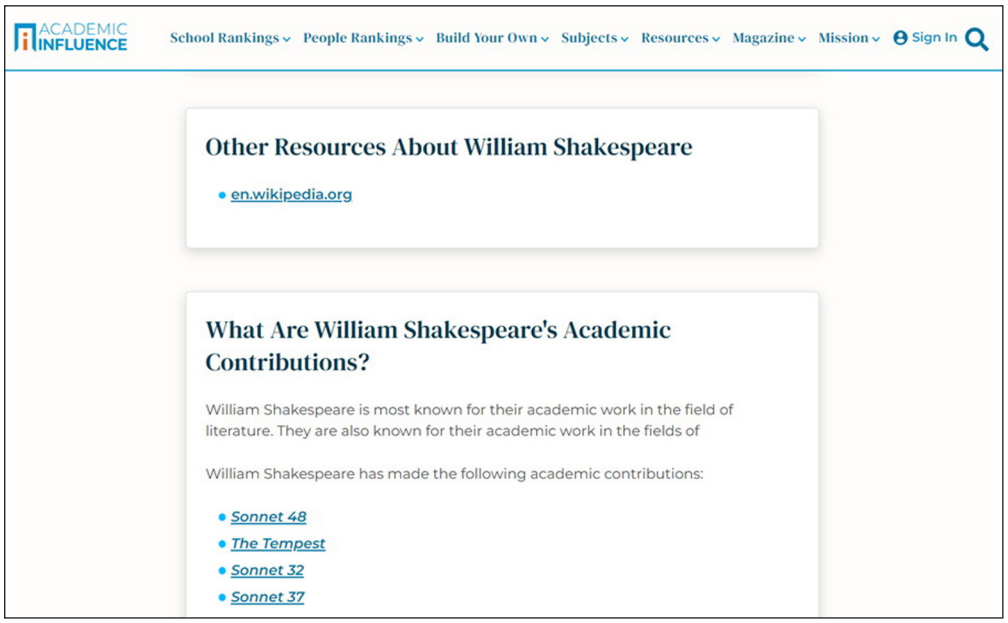

Figure 8. Entry for William Shakespeare from 'the 50 most influential scholars of all time'

Trying to establish the greatest thinkers in world history via an algorithm for researchers and academics is unlikely to produce trustworthy results.

The need for a sanity check

There is undoubtedly a need for sanity checking of AI-based tools, and humans are necessary to carry this out. Algorithms in narrow AI have no knowledge of the real world. A system that can differentiate images of cats from images of dogs could still not define what a cat or a dog is, nor could such a system identify any other animal species. However, it is tempting to apply an algorithm to a corpus way beyond a viable scope. However clever the algorithm, there is a need for a human to check that the results correspond with a common-sense view. What is meant by common sense here is quite specific, and quite limited. For example, consider an algorithm that applies subject tags to an academic article. This algorithm provides a probability ranking for the subjects physics, chemistry and politics. Using the algorithm, the following results were obtained, using a ranking between 0 and 1, see Table 1.

'A system that can differentiate images of cats from images of dogs could still not define what a cat or a dog is'

Content	physics	chemistry	politics
Article 1	0.65	0.54	0.12
Article 2	0.45	0.73	0.19

Table 1. Typical predictive scores for subject tagging

Clearly, article 1 looks to be more about physics than anything else, while article 2 seems to be obviously about chemistry. However, the algorithm has found traces of politics content in both articles. The determination of a suitable threshold, below which the user should state 'this article is not about politics', needs to be determined by common sense – or by using a subject-matter expert to identify what the threshold should be. In this case, the machine delivers a result, but in all probability the result for politics can be discounted.

One of the most widely used metrics for text-based AI is the F1 score,[36] which measures the mean of recall and precision, usually shown on a scale between 0 (no accuracy) and 1 (perfect accuracy). However, the F1 score has limitations which are easy to recognize using common sense. Harikrishnan[37] gives an example of a pregnancy test of 100 women, which identifies five as pregnant when they are not (false positive), and ten as not pregnant when they are pregnant (false negative). A machine-based algorithm that resulted in these figures would have an F1 score of 0.8, which in other contexts might represent an acceptable score, but would certainly not be adequate for a pregnancy test.

This is where the information professional has a key role. Some idea of context, of what is or is not required in the situation, is vital for ensuring that any tool delivers relevant results. Are the results relevant in context?

An example of the F1 score in use for subject tagging is described by Goh.[38] This study compares humans with a machine used to classify articles by subject. The machine outperforms the postgraduates by delivering a significantly more accurate set of tags. Even more impressive, it took one postgraduate two hours to classify 247 abstracts, compared to five seconds for a machine to complete the same exercise.

A comparison of the subject tagging between the machine and the humans shows immediately that the machine consistently delivers better results than the human taggers, but the most significant inference of the study is implied rather than explicitly stated. For the purpose of subject tagging, humans were found in this controlled study to have an average F1 score in the region of 0.5 or less, while the machine result was considered to be usable with an F1 score of around 0.7. While this figure at first glance seems poor (given that a perfect score would be 1), the implication is that the F1 score should be interpreted in context, not as an absolute measure. In other words, when comparing machine-based with human results, we should be considering relative, rather than absolute, measurements. If a machine delivers a better result than what is achievable by hand, it makes sense to adopt the machine solution immediately.

'If a machine delivers a better result … it makes sense to adopt the machine solution immediately'

Some sanity checks can be built into the tool itself. For example, a tool to identify peer reviewers could helpfully provide an indication when an article is submitted that is outside the corpus used to identify reviewers. A tool to identify the most relevant journal for an article could have a result of 'no suitable match found' if an article is submitted to the tool that is outside the subject areas of the corpus.

The corpus and bias

Another key role that information professionals can play in the evaluation of AI tools is the awareness of potential and actual bias. Any corpus contains bias. Bias is typically independent of any AI tools. All real-world data is inevitably biased. Even seemingly neutral collections reveal unconscious bias. For example, it might be assumed that PubMed, a collection of millions of scholarly articles published on biomedical topics over the last 50 years, would comprise a statistically valid corpus, yet there are more male than female authors in PubMed. Is this surprising? A study of gender disparity in medical articles[39] found this was the case over the last 20 years. Another article[40] shows a revealing graph of male and female authorship of articles in science journals since 1955. While the proportion of female authorship is growing, males continue to author the majority of science articles.

Of course, once bias is recognized, it is possible to take steps to work around this bias, but lack of awareness of bias means that there is an important role for information professionals when recommending these tools.

Corpus size

It is difficult to give absolute figures, but statistically based and reliable AI tools require at least several hundred documents in the training set, and a minimum of several thousand documents in the corpus. Depending on the goal, the training set may need to be larger if the question asked is less straightforward than differentiating between numbers or letters.

The algorithm

Any algorithm should be open in the sense that its basic methodology is clearly stated. This should not require any breach of commercial confidence for paid software and has the benefit that potential bias can therefore be revealed. The alternative, without any explanation of the methodology, is the dreaded 'black box', a tool that must be trusted rather than understood. An example of a methodology statement is 'we find peer reviewers by identifying similar articles to your submission. Then we identify the authors of those articles.'. This is how most AI tools to identify peer reviewers work, for example the Web of Science Reviewer Locator.[41] What constitutes a clearly defined algorithm is well explained on the 50 Examples page 'Background: Algorithms'.[42]

Manual checking

It is useful, but by no means a complete assessment, to try out the tool directly. By inputting an example or two with a known prediction, some idea can be obtained of the capabilities. It is surprising how frequently this technique can reveal assumptions on the part of the software developers that were not made clear when the tool was delivered.

Evaluation and metrics

Asking a couple of colleagues to have a look at the product is not a full evaluation. If using human criteria to evaluate the tool, give thought to suitable criteria for comparing human and machine performance – see, for example, Ewerth.[43]

When building or adjusting websites, A/B tests are often used. An A/B test is a randomized trial widely used in website development, in which two versions of a variable, such as a web page layout, are shown to different groups of users, and their resulting behaviour is measured. Neither group is aware of the trial or of the other version. In this way, like a randomized control trial, it is possible to identify which version has the greatest impact. A/B tests on the live site are the best way to evaluate the impact of any change.[44] This is because:

- evaluations are made with data rather than by guesswork
- the test gives the response of real users
- the results enable the estimation of metrics of success – what is an acceptable goal (rather than an absolute goal).

Similarly, for many AI tools, a methodology based on the A/B test is feasible, if complex, to provide a solid assessment. For example, a comparison of machine-based results and human-generated results could be carried out.

As for human evaluation, all humans are not equal at this task. Consider who are the best people to evaluate this tool. Should it be the person managing the process, or should it be the end user? It is a well-established principle in website design that the best way to evaluate software is to test it with real users, not with the software team who built the tool, or with the people tasked with managing the delivery of the tool.

When humans are used to evaluate a tool, there is the question of how many checks are required. If we ask the machine to use a training set of, say, 500 documents to determine the parameters for the exercise, does it make sense to judge the results by a human looking at two or three examples? What is the level of human agreement?

Credibility

What role does the information professional play in all this? Most of the criteria described above could be checked directly by the user, that is, the researcher, but most researchers have neither the time nor the knowledge to make measured comparisons of different tools. Without a solid analytical framework, humans tend to rely on instinct, which could be described as an internal assessment mechanism – they instinctively trust (or do not trust) a familiar methodology, or tools they have used before. The role for the information professional in all this is providing credibility: providing users with external validations that enable them to trust a tool and to deploy it with confidence. Researchers will, for the most part, look for an external validation of a tool that they can trust. The information professionals provide the credibility, based on their detailed evaluation.

'The role for the information professional in all this is providing credibility'

The AI toolkit: a framework for evaluation of AI tools

Here, in summary, is a toolkit for information professionals appraising any AI tool. Although making use of Long and Magerko's idea of AI literacy, the requirements here are much more specific.

Goal

1. What is a realistic goal? Expecting perfection for an AI utility is impossible. AI tools based on a training set cannot have 100% accuracy. Nonetheless, the accuracy they provide should be considerably greater than using humans for the same task.

Corpus

2. Is the corpus large enough? Is the training set large enough?
3. What are the start and end dates for the data in the corpus? Does this matter?
4. Who chose the corpus, when was it chosen and for what purpose? Details of the corpus used, like the data for a research article, should be publicly stated and accessible.
5. What is the corpus bias?
6. Is the tool likely to raise diversity, equality and/or inclusion issues?
7. Is personal data captured and reused?

Algorithm

8. Have the developers provided a single-sentence summary of the methodology behind the algorithm?

Evaluation and Metrics

9. Have I measured the current process before introducing any change, for example, time taken, number of errors?
10. Who to evaluate: end users or subject-matter experts, or both? Internal or external?
11. What metrics will be used to evaluate the tool? The F1 score, if used, must be interpreted in context.

Sanity check

12. Sanity check/common sense: Have the developers built in 'common-sense' limitations to prevent the algorithm being applied too widely? Am I asking a meaningful question? Is this a feasible exercise?
13. Does the tool provide feedback when a question is out of scope?
14. Based on the checks above, is the tool fit for purpose?

Dissemination

Is there easy-to-read documentation and guidance for new users that explains in simple terms how to use the tool and how it improves on current processes?

Feedback

Does the tool provide a feedback loop so it can be improved over time?

Conclusion

To make the best use of AI tools in publishing requires not only high-quality software, but also a critical awareness of the context in which the tool will be used. By following a template, and asking the right questions, those responsible for recommending and assisting in the take-up of AI tools can ensure a much higher success rate for this technology. All of us, often without realizing it, have developed in our everyday life an unconscious heuristic approach to working with AI and with those tools that make use of AI around us. If we ignore AI, we miss a host of benefits that can make us work more effectively. If we make use of AI tools uncritically, we risk discrediting a whole area of new technology. By using this framework, information professionals, without being developers, can become qualified to assess AI tools with confidence.

'By using this framework, information professionals … can become qualified to assess AI tools with confidence'

Abbreviations and Acronyms
A list of the abbreviations and acronyms used in this and other *Insight*s articles can be accessed here – click on the URL below and then select the 'full list of industry A&As' link: http://www.uksg.org/publications#aa.

Competing interests
The author has formerly provided services for UNSILO, a company mentioned in this article.

References

1. Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, 3rd edition (Pearson, 2016).
2. Ray Kurzweil, "A Wager on the Turing Test: Why I Think I Will Win," *Kurzweil, Kurzweilai.net*, April 9, 2002, https://www.kurzweilai.net/a-wager-on-the-turing-test-why-i-think-i-will-win (accessed 23 September 2022).
3. Rory Cellan-Jones, "Stephen Hawking Warns Artificial Intelligence Could End Mankind," *BBC News*, December 2, 2014, sec Technology, https://www.bbc.com/news/technology-30290540 (accessed 23 September 2022).
4. A. M. Turing, "Intelligent Machinery, A Heretical Theory*," *Philosophia Mathematica* 4, no. 3 (September 1, 1996): 256–60, DOI: https://doi.org/10.1093/philmat/4.3.256 (accessed 23 September 2022).
5. Nicholas Carr, *The Shallows: How the Internet Is Changing the Way We Think, Read and Remember*, Main-Re-issue edition (London: Atlantic Books, 2020).
6. Elon Musk, "Elon Musk: 'With Artificial Intelligence We Are Summoning the Demon'," *Washington Post*, October 24, 2014, https://www.washingtonpost.com/news/innovations/wp/2014/10/24/elon-musk-with-artificial-intelligence-we-are-summoning-the-demon/ (accessed 23 September 2022).
7. Kate Crawford, *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence* (New Haven: Yale University Press, 2021). DOI: https://doi.org/10.12987/9780300252392
8. Cathy O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy*, 01 edition (Penguin, 2016).
9. Erik J. Larson, *The Myth of Artificial Intelligence: Why Computers Can't Think the Way We Do* (Cambridge, Massachusetts: Belknap Press, 2021). DOI: https://doi.org/10.4159/9780674259935
10. Yann LeCun, Corinna Cortes, and Christopher J C Burges, "The MNIST Database," THE MNIST DATABASE, http://yann.lecun.com/exdb/mnist/ (accessed 23 September 2022).
11. Natalia Domagala and Hannah Spiro, "Engaging with the Public about Algorithmic Transparency in the Public Sector," *Centre for Data Ethics and Innovation Blog*, June 21, 2021, https://cdei.blog.gov.uk/2021/06/21/engaging-with-the-public-about-algorithmic-transparency-in-the-public-sector/ (accessed 23 September 2022.
12. Domagala and Spiro, "Engaging with the Public."
13. European Parliament. Directorate General for Parliamentary Research Services. *A Governance Framework for Algorithmic Accountability and Transparency* (LU: Publications Office, 2019), https://data.europa.eu/doi/10.2861/59990 (accessed 23 September 2022).
14. Rob Merrick, "Fears of Another A-Level-Style Fiasco as Scrutiny of Policies Made by Computer Are Ditched Following Brexit," *The Independent*, February 10, 2022.
15. Pedro Domingos, *The Master Algorithm: How the Quest for the Ultimate Learning Machine Will Remake Our World*, 1st edition (London: Penguin, 2017).

16. "Artificial General Intelligence," in *Wikipedia* (Retrieved 9 May 2022), https://en.wikipedia.org/w/index.php?title=Artificial_general_intelligence&oldid=1086964857 (accessed 23 September 2022).

17. "What Is Supervised Learning?," *IBM Cloud Learn Hub,* August 19, 2020, https://www.ibm.com/cloud/learn/supervised-learning (accessed 23 September 2022).

18. Pandu Nayak, "Understanding Searches Better than Ever Before," *Google* (blog), October 25, 2019, https://blog.google/products/search/search-language-understanding-bert/ (accessed 23 September 2022).

19. Motahhare Eslami et al., "'I Always Assumed That I Wasn't Really That Close to [Her]': Reasoning about Invisible Algorithms in News Feeds," in *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15 (New York, NY, USA: Association for Computing Machinery, 2015), 153–62, DOI: https://doi.org/10.1145/2702123.2702556 (accessed 23 September 2022).

20. Mary Shultz, "Comparing Test Searches in PubMed and Google Scholar," *Journal of the Medical Library Association: JMLA* 95, no. 4 (October 2007): 442–45, DOI: https://doi.org/10.3163/1536-5050.95.4.442 (accessed 23 September 2022).

21. G. A. Miller, "The Magical Number Seven plus or Minus Two: Some Limits on Our Capacity for Processing Information," *Psychological Review* 63, no. 2 (March 1956): 81–97, https://pubmed.ncbi.nlm.nih.gov/13310704/ DOI: https://doi.org/10.1037/h0043158 (accessed 27 September 2022).

22. Larson, *The Myth of Artificial Intelligence*.

23. Duri Long and Brian Magerko, "What Is AI Literacy? Competencies and Design Considerations," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (CHI '20: CHI Conference on Human Factors in Computing Systems, Honolulu HI USA: ACM, 2020), 1–16, DOI: https://doi.org/10.1145/3313831.3376727 (accessed 23 September 2022).

24. "Online-Spellcheck.Com," https://www.online-spellcheck.com/ (accessed 23 September 2022).

25. Jenna Burrell, "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms," *Big Data & Society* 3, no. 1 (January 6, 2016): 2053951715622512, DOI: https://doi.org/10.1177/2053951715622512 (accessed 23 September 2022).

26. "Cambridge Core, Recommender Links for Book Chapter," *Cambridge Core*, https://www.cambridge.org/core/books/abs/python-programming-for-biology/machine-learning/570B575C26034A8CB9A7AF7E17A795AB/ (accessed 23 September 2022).

27. Eugene Garfield, "Citation Indexes for Science," *Science* 122, no. 3159 (July 15, 1955): 108–11, DOI: https://doi.org/10.1126/science.122.3159.108 (accessed 23 September 2022).

28. J. E. Hirsch, 'An Index to Quantify an Individual's Scientific Research Output', *Proceedings of the National Academy of Sciences* 102, no. 46 (November, 15 2005): 16569–72, DOI: https://doi.org/10.1073/pnas.0507655102 (accessed 23 September 2022).

29. "Scite: See How Research Has Been Cited," scite.ai, https://scite.ai/ (accessed 23 September 2022).

30. "Scholarcy," Scholarcy|The long-form article summariser, https://www.scholarcy.com/ (accessed 23 September 2022).

31. "Semantic Scholar|AI-Powered Research Tool," https://www.semanticscholar.org/ (accessed 23 September 2022).

32. Marco Valenzuela, Vu Ha, and Oren Etzioni, "Identifying Meaningful Citations," in *Workshops at the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015, 6, http://ai2-website.s3.amazonaws.com/publications/ValenzuelaHaMeaningfulCitations.pdf (accessed 27 September 2022).

33. "Scholarcy".

34. "2021 Trends in Library Analytics," EBSCO Information Services, Inc., December 13, 2021, https://www.ebsco.com/blogs/ebscopost/2021-trends-library-analytics (accessed 23 September 2022).

35. "William Shakespeare," *Academic Influence*, https://academicinfluence.com/people/william-shakespeare-1 (accessed 23 September 2022).

36. "F-Score," *Wikipedia*, May 9, 2022, https://en.wikipedia.org/w/index.php?title=F-score&oldid=1086969326 (accessed 23 September 2022).

37. N.B. Harikrishnan, "Confusion Matrix, Accuracy, Precision, Recall, F1 Score," *Analytics Vidhya* (blog), December 10, 2019, https://medium.com/analytics-vidhya/confusion-matrix-accuracy-precision-recall-f1-score-ade299cf63cd (accessed 23 September 2022).

38. Yeow Goh et al., "Evaluating Human versus Machine Learning Performance in Classifying Research Abstracts," *Scientometrics* 125 (July 18, 2020): 1197–1212, DOI: https://doi.org/10.1007/s11192-020-03614-2 (accessed 23 September 2022).

39. Karla Bernardi et al., "Gender Disparity in Authorship of Peer-Reviewed Medical Publications," *The American Journal of the Medical Sciences* 360, no. 5 (November 2020): 511–16, DOI: https://doi.org/10.1016/j.amjms.2019.11.005 (accessed 23 September 2022).

40. Tyler Machado, Molly Callahan, and Eunice Esomonu, "Do Women Publish Less than Men in Scientific Fields?," *News @ Northeastern*, March 5, 2020, https://news.northeastern.edu/2020/03/05/do-women-publish-less-than-men-in-scientific-fields-turns-out-scientists-have-been-asking-the-wrong-question/ (accessed 23 September 2022).

41. "Web of Science Reviewer Locator," *Clarivate*, https://clarivate.com/products/scientific-and-academic-research/research-publishing-solutions/web-of-science-reviewer-locator/ (accessed 23 September 2022).

42. "Background: Algorithms," 50 Examples 1.0 Documentation, https://fiftyexamples.readthedocs.io/en/latest/algorithms.html (accessed 23 September 2022).

43. Ralph Ewerth et al., "'Are Machines Better Than Humans in Image Tagging?' – A User Study Adds to the Puzzle," *Advances in Information Retrieval*, ed. Joemon M Jose et al., Lecture Notes in Computer Science (Cham: Springer International Publishing, 2017), 186–98, DOI: https://doi.org/10.1007/978-3-319-56608-5_15 (accessed 23 September 2022).

44. Ron Kohavi, Diane Tang, and Ya Xu, *Trustworthy Online Controlled Experiments: A Practical Guide to A/B Testing* (Cambridge: Cambridge University Press, 2020), DOI: https://doi.org/10.1017/9781108653985 (accessed 23 September 2022).

Corresponding author:
Michael Upshall
Consultant, GB
E-mail: michael@consultmu.co.uk
ORCID ID: 0000-0003-1115-6847

To cite this article:
Upshall M, "An AI toolkit for libraries," *Insights*, 2022, 35: 18, 1–16; DOI: https://doi.org/10.1629/uksg.592

Submitted on 08 June 2022 Accepted on 18 July 2022 Published on 01 November 2022

Published by UKSG in association with Ubiquity Press.