

UNSILO AI in Academic Publishing Survey 2019

1 November 2019

Summary

Between July and September 2019 UNSILO carried out a survey of academic publishers. They were asked questions about their current and expected use of AI, following which they were asked questions about their attitudes to AI. The results of the survey were discussed at two panel sessions at the Frankfurt Book Fair. Each session was chaired by Toby Green, former head of OECD Publishing. Among the main findings of the survey are:

- Over two-thirds of publishers are using at least one AI tool
- One-third of publishers are building their own AI tools in-house
- 45% of publishers not currently using AI tools plan to introduce them in the coming twelve months
- AI is currently being used primarily to enhance staff efficiency, rather than to create new functionality
- Only 10% of publishers using AI check for bias.

Background

Every few years a new technological innovation creates a disruption in the world of academic publishing. From print to online delivery, XML workflows, the introduction of software tools for manuscript submission and workflows have all transformed the life cycle for the creation and dissemination of scholarly content. AI and machine learning are in the process of disrupting established systems yet again. But is AI a threat or an opportunity for academic publishing? Is this technology a supplement to or a replacement for human decision-making?

To gain a clearer impression of the extent of this change to current systems, UNSILO carried out a survey, between July and September 2019, UNSILO ran a survey exploring attitudes to the use of AI, and how AI is being used in academic publishing. Respondents (total=82) focused on publishers. The results were announced at the Frankfurt Book Fair in two panel sessions, chaired by Toby Green, former head of publishing at the OECD. The first panel discussed opportunities and challenges of AI in academic publishing, while the second panel looked at trust, bias and other related issues in AI.

The justification for introducing more automation to the scholarly workflow should be obvious. More than 3,000 new academic journal articles are published every day, and yet many of the tools for processing content through the academic workflow continue to be very manually based. For many aspects of the workflow, tasks have been outsourced to an offshore provider, such as companies in India, Asia, and the Philippines. However, such outsourcing has proved only to be a temporary fix. Labour costs continue to rise, even in areas with lower GDPs, and in the longer-term, a more automated solution will be necessary.

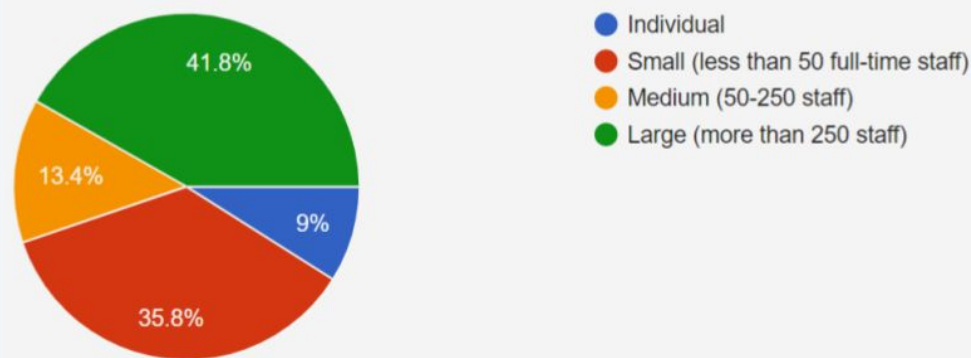
Nonetheless, fears remain. In the wider world, questions are increasingly asked about machine-learning algorithms and their effect on society. Hence, the survey looked at such topics as:

- Can you compare the accuracy of human v machine operations?
- How can we check for bias? Who should carry out these checks?
- Do we relinquish all responsibility for editorial processes by adopting AI tools?
- Can we provide human curation, for example feedback on AI decisions?
- Can we reduce time to publication without compromising quality?
- Can AI tools answer the same questions as humans?

Who responded?

Size of organisation

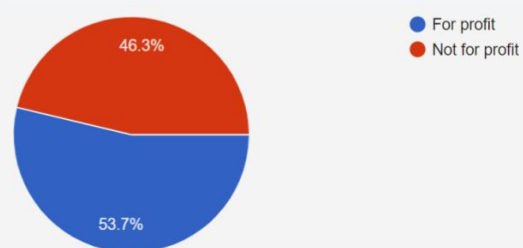
Most respondents are from either small (<50) or large (>250) organizations.



The bulk of the respondents were publishers (43%), followed by software vendors and developers (15%) and information professionals (15%). Organisations they work in tended to be either large (42%) or small (less than 50 staff) (36%).

Type of organisation

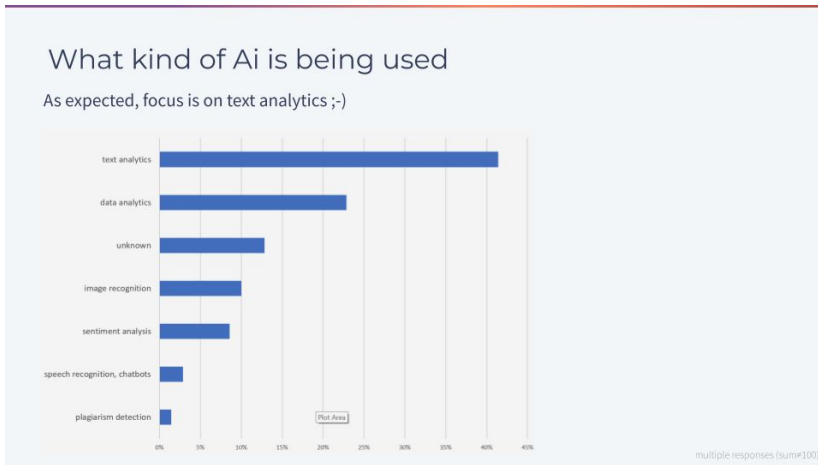
Respondents are equally divided among for-profits and nonprofits.



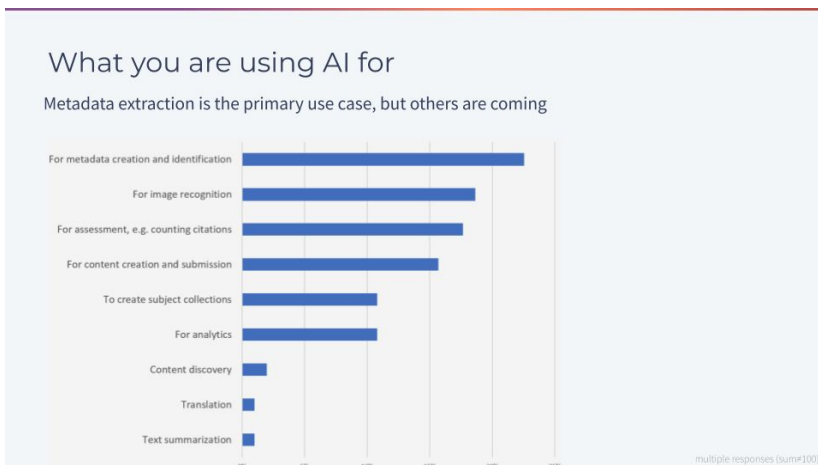
Respondents were almost evenly split between for-profit and not-for-profit organisations.

Findings: how AI is currently being used

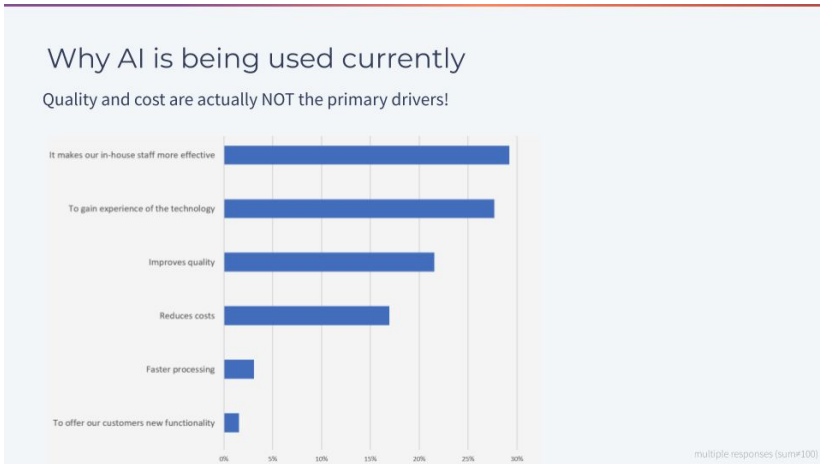
First, we looked at publishers currently using AI tools, and we asked what those tools were being used for.



Perhaps unsurprisingly, text analytics was the biggest single use, with over 40% of publishers who were using AI were using it for text analytics; although, perhaps significantly, more than 10% of respondents were unable to state what the AI was being used for, which suggests that although AI might be in use, it is often the preserve of a technical team that does not communicate much about what it is actually being used for. By “text analytics”, publishers typically mean such tasks as automatic tagging, identification of entities, and identification of metadata such as title and author.



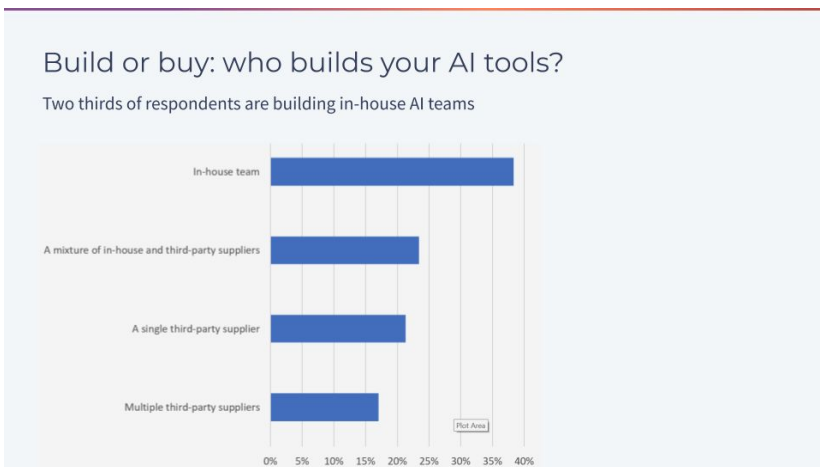
A more detailed follow-up question highlighted metadata creation and identification as the main goal of AI tools.



As for the justification of why AI is being used, making in-house staff more effective was the most common reason specified, followed by a desire to get more experience of the technology. As Toby Green pointed out in the panel sessions, using AI to create new functionality for user needs is not mentioned. Clearly AI is intended to be used to improve current processes than a transformation of the researcher experience – although such a transformation may come later.

Findings: who is carrying out the AI work

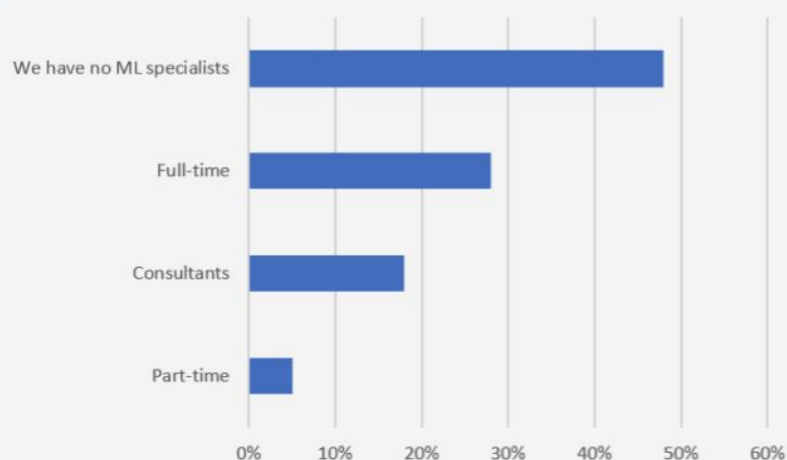
This area considered the extent to which organisations are developing new skills, or obtaining those skills from outside the organisation.



When asked the “build or buy” question, the largest single response was from organisations using an in-house team to build AI tools, but in fact the situation appears to be fairly evenly balanced between in-house and out-of-house teams. Where an external supplier is used, the preference seems to be to use a single third-party supplier rather than multiple suppliers (presumably for reasons of simplicity).

Do you employ machine-learning specialists?

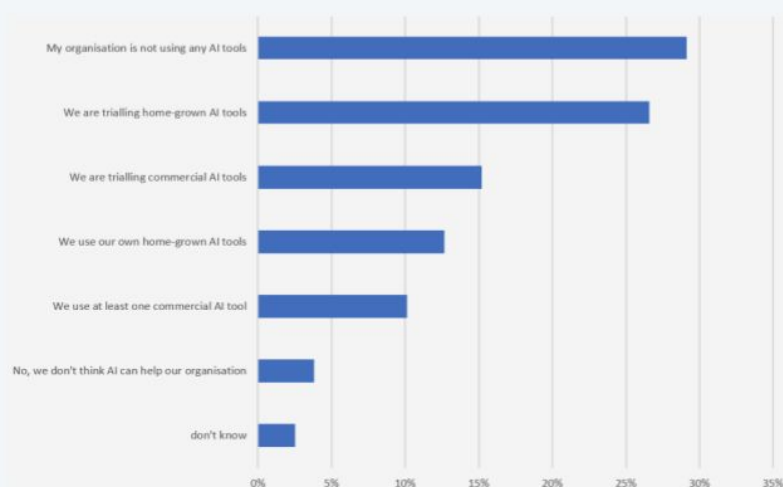
About half of the industry employs ML specialists



About half the industry has no dedicated machine-learning staff. This suggests a landscape where AI tools are either purchased ready-made, simply being configured to work with existing platforms and tools via APIs, for example, or not used at all. Around 28% of organisations employ full-time machine-learning staff, which could suggest an attempt to build a strategic AI solution, but since there is no further information about what tools these full-time staff are creating, it might also be that they are working on small-scale, self-contained initiatives that do not represent a global strategy. It is difficult in a survey to clarify whether an AI solution is strategic or small-scale, and of course it may be that as long as the tool is still being evaluated, the answer may not be clear.

Are you using AI tools?

But most organizations are only in a trial stage



This question was aimed at identifying the maturity of the market. Are publishers using or simply trialling AI tools? When asked to give details about the use of AI tools, almost 30% of publishers stated they are not using any AI tools. Just over a quarter were trialling home-grown tools, with a smaller number (around 15%) trialling commercial tools.

It is worth noting that AI is already ubiquitous in business software, and one of the panellists pointed out that most publishers are already using some AI tools without necessarily being aware of it, for example, in the area of plagiarism checking, or email spam detection.

What AI techniques are being used?



This question produced some surprising answers. There was a wide spread of responses, but machine learning and NLP were joint leaders, ahead of rule-based tools. As publishers could answer yes to more than one of the techniques, the total is above 100%, but even so, it is remarkable that machine learning is in equal first place. While NLP tools are today widespread, the introduction of machine learning in academic publishing is much more recent, even if the term is often in the news headlines. Perhaps a cynic might say that because of its topical interest, it is top of people's attention.

Reasons for not using AI

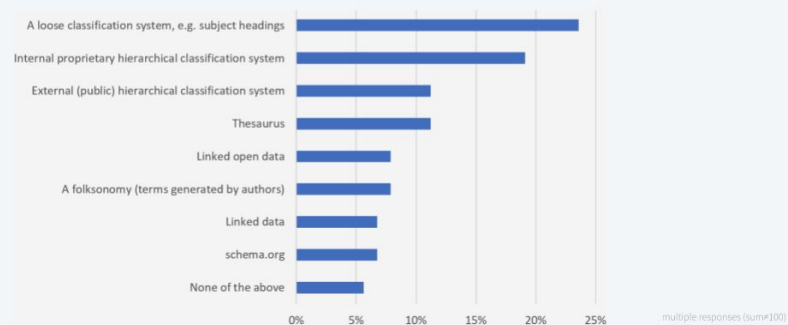


Publishers not using AI give as their main reason for not doing so the lack of specialist AI-trained staff, although responses were quite broadly divided to this question. There does not seem to be one dominant reason for not using AI; instead, multiple factors play a part. The biggest single reason for not using AI, a lack of specialist trained staff, was only stated by 18% of respondents, and perhaps reveals a technology at a fairly early stage of implementation, with users feeling they need to understand it in detail before putting it into production.

AI and taxonomies

Taxonomies and classification

There is little sign of industry-wide use of shared classification systems such as Linked open data



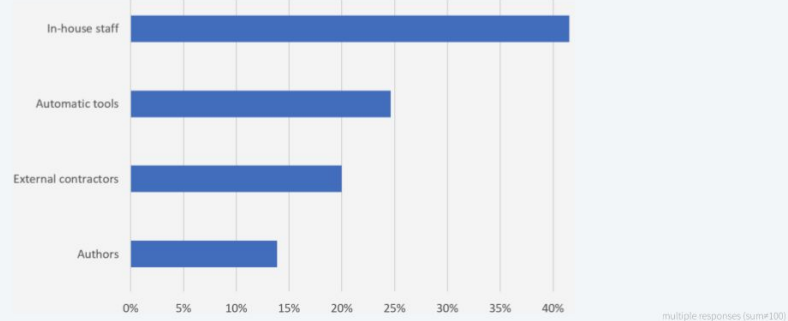
The centrality of taxonomies and ontologies in the academic publishing cycle has come under examination with the introduction of new technologies. Some kinds of AI work with taxonomies, for example using a training set to expand the tagging of documents to a larger corpus, while others (notably unsupervised machine-learning) do not require any pre-existing taxonomy. Hence, we included a question to learn more about publishers' use of taxonomies, as well as identifying if publishers are using linked data or schema.org as part of their AI strategy. These tools, however, came a long way down the list of taxonomies in use by publishers; more than 40% of respondents were using subject headings or some kind of classification system. It could be argued that if only 7.5% of publishers are using linked open data, there will not be much open data available to be shared.

The most widely used taxonomy was a publisher's own internal classification system. This, of course, will not necessarily be compatible with any other classification system, and suggests that providing compatible and shareable metadata with other publishers is not a priority for most publishers.

Who adds the metadata?

Who adds the metadata?

Publishers still spend valuable time carrying out tasks of low strategic importance, such as adding metadata



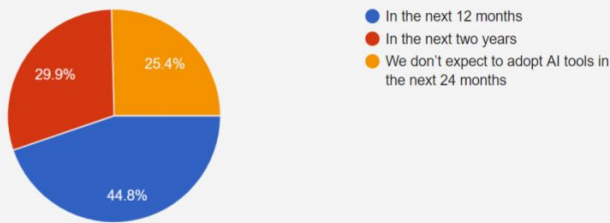
For a publisher to use metadata, it has to be created. Remarkably, the work of creating the metadata remains the task of in-house staff. This is an area where one would imagine either outsourced labour and/or automated tools could be making in-roads, but they clearly have not overtaken human tagging activities.

Findings: Take-up of AI, opportunities and challenges of using AI tools

The findings show a steady take-up of AI tools by publishers. Two-thirds of respondents are currently using AI tools, and only 3% of respondents felt that AI could not benefit their activities in some way.

Timescale for adoption of AI tools, for non-AI users

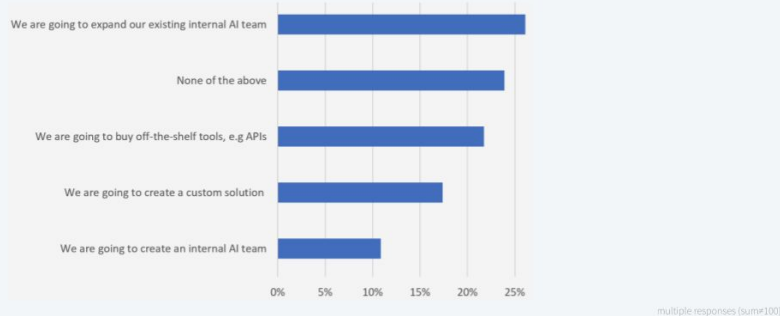
There is some sense of urgency...



This question was to identify what publishers are going to do next with AI. The story here is a very positive one. When publishers not currently using AI were asked about when they planned to introduce AI-based processes, around 45% of publishers expected to do so in the coming 12 months – this clearly suggests introducing AI tools is something of a strategic imperative.

Existing users: What will your next steps in AI be?

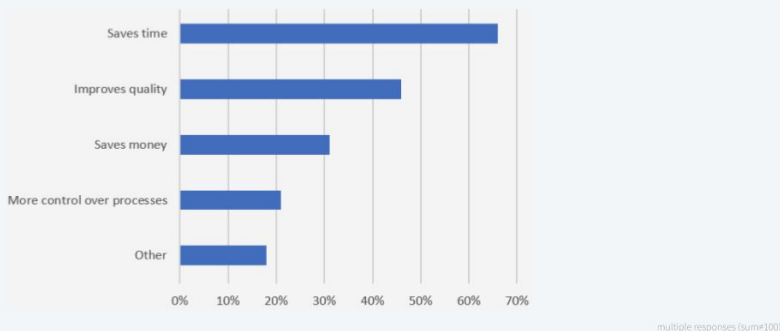
Companies using AI are following up on their investment



How the publishers plan to expand their AI capability is interesting: do you build or buy? Around one third of publishers plan to create or to build an in-house team, which suggests an emphasis on keeping skills internal to the organisation, rather than buying in external tools.

Primary perceived Benefits of AI

Saving time the biggest perceived benefit



The primary perceived benefit of AI was that it could save time. This could be seen as evidence of a new realism among publishers, since the thinking is presumably to apply AI tools to relatively straightforward processes that could be completed faster with the aid of a machine, such as the identification of relevant articles for a manuscript submission, or finding potential peer reviewers who have authored papers on similar topics to a manuscript submission. These are tasks that a human could carry out, but not as quickly as with a machine doing some of the looking up.

Other stated benefits of AI

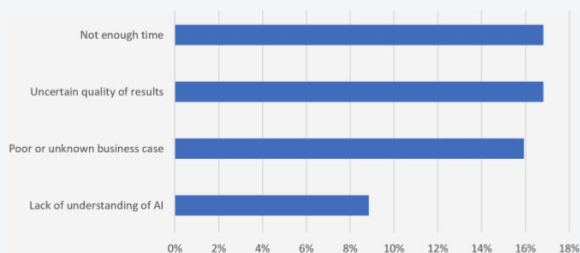
AI will generate new types of value

- Enable large data analysis tasks
- Enable products we've not considered possible before
- Should create new engagement with R&D
- Enable humans to focus on more creative tasks
- Power new business models for consumption
- Provide an evidence base to support human decision-making

When asked to provide more details about what benefits AI tools might provide, respondents suggested several ideas. Here are signs of using AI in a more strategic way, and perhaps it is not surprising that these ideas tended to be stated by just one or two respondents.

Existing users: The biggest factor holding you back?

Companies using AI may have a hard time building the business case



Note that this question was only asked of existing users of AI tools; these are the publishers who you would expect to understand the opportunities and the challenges. Yet there was no single dominant answer, although there were several reasons stated by just one or two respondents (not shown here). Interestingly, a lack of understanding of AI was the least chosen option, something that may be surprising for a technology still in its infancy. Responses stating lack of time suggest that AI tools may not have presented their business case successfully enough, since new technology should either provide new functionality or save time, or both, and clearly neither has been identified from this question, even with existing users of AI.

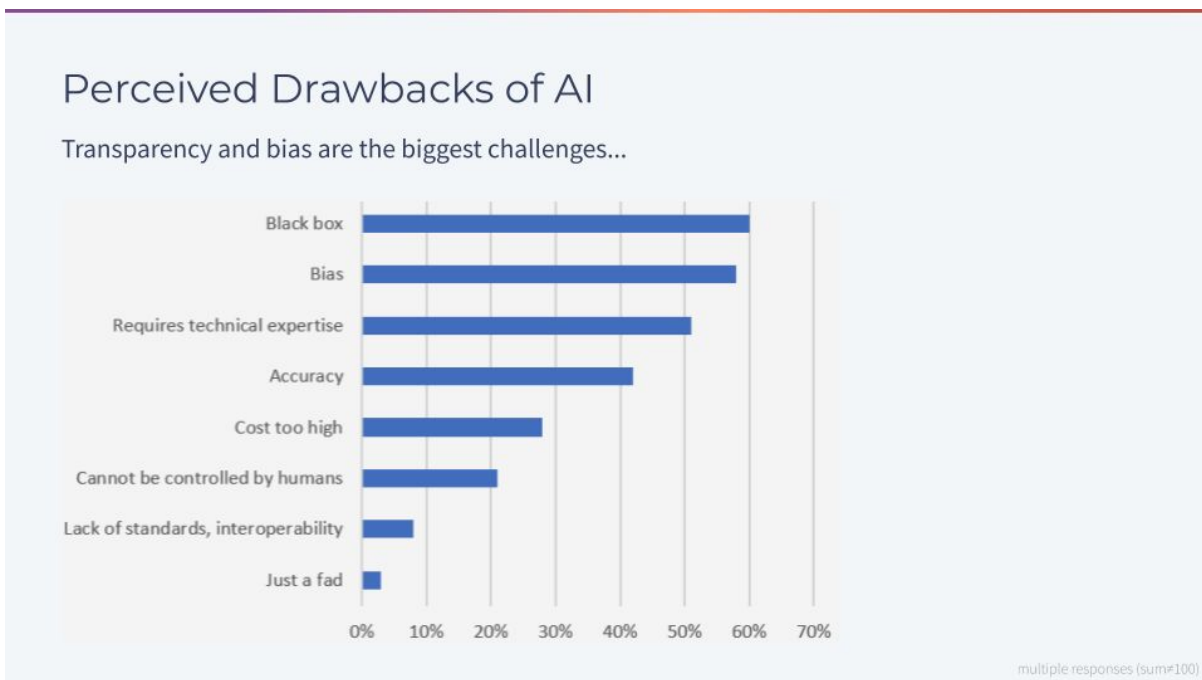
Findings: Trust, bias and accuracy

Perhaps inevitably, a question was included about bias and accuracy. Yet, for all the coverage in the media about questions of bias and privacy being topics of major concern, few of the respondents seemed to be taking action about these issues.



These responses suggest that AI tools, are not checked exhaustively for possible issues. Accuracy is the main check, but fewer than 10% of publishers check their AI tools for bias, and only 8% for privacy and compliance with GDPR. Of course, it could be that the AI tools in use were exhaustively checked for these aspects before their roll-out in an operational setting, but it is increasingly understood that questions of bias and accuracy may be as much related to the underlying data as to the algorithms in use. Hence each use in a new domain or environment should be validated.

Drawbacks of AI tools



When asked to state the drawbacks of AI, not surprisingly, bias ranked very high (over 50%), but the highest drawback was stated to be “black box” – the fear that humans cannot understand the reasoning behind the AI tool (this applies especially to algorithms that do not use a training set, but is also a charged levelled against tools that use a training set curated by humans). Interestingly, the cost of implementing AI, while still significant, ranked considerably below these other objections.

Perhaps the moral here is that software vendors should take more trouble to explain how their tools work, and/or provide some means for humans to provide feedback to help qualify the result.

Conclusion

Overall, take-up of AI in academic publishing shows a steady progress, but the survey seems to suggest that ‘transformational’ uses of AI are still rare. For the most part, AI tools have been introduced to improve workflow and efficiency.

The largest publishers have already been experimenting with AI tools. But just a tier below these, even mid-size publishers have difficulties over the introduction of AI, as they will not have the resources to create in-house expertise in AI. Yet publishers claim a major drawback to implementing AI is the lack of trained in-house staff. It seems unlikely there is a simple solution to this problem. Either publishers need to invest much more in internal AI resources (slow and costly), or they need to become more trusting about including third-party AI tools. It would be interesting to compare the situation in a year’s time to see if these views have changed.

Panel discussions at the Frankfurt Book Fair

UNSILO hosted two panel sessions about the AI in academic publishing survey at this year’s Frankfurt Book Fair. The two sessions were chaired by Toby Green, formerly head of publishing at the OECD. Each panel comprised three participants, and there were also suggestions and comments from the floor. As Toby Green pointed out in his introduction, most publishers were using AI (or planning to use it) to save money or time with their in-house processes. While this undoubtedly provided some clear benefits, it was clear that most publishers weren’t planning to use AI to add new functionality for the end user.

Panel Session One: Implementation and Adoption

Even if there were no case studies of major new functions, nonetheless, there were some impressive descriptions of workflow improvements. Dave Flanagan gave some fascinating examples of how Wiley had been able to improve the quality of the data in some of their chemistry journal articles using AI tools. At the other end of the corporate spectrum, Josh Nicholson, from scite.ai, with only a handful of staff, gave another quick win for AI. Their model is simple: they take thousands of academic articles and identify citations that support or refute a hypothesis. Using these manually annotated examples as a training set, they then apply these to a test set and measure their results. Once the system has been trained, it is able to identify confirmations and refutations in new articles submitted to it, and has now been rolled out to identify some 400 million citations from the published literature. The process, as described by Josh Nicholson in his engaging and straightforward manner, seemed disarmingly simple.

Some themes emerged very clearly from the discussion; one was the need for a human intermediate to manage and curate the AI. Another theme was the contrast between large and small companies. Larger publishers such as Wiley will have their own internal departments developing AI tools, while a startup such as scite.ai concentrates on building just one or two AI-based tools, which publishers then deploy. Small- and medium-sized publishers, such as BMJ, Helen King pointed out, have to make use of external partners to create AI tools. This appears to be an increasingly common arrangement; Smaller publishers choose API-based tools that bolt on to their existing platform and/or workflow.

Panel Session Two: Trust and Bias

The second panel, about trust and bias, also had some examples of small-scale wins: Geoffrey Bilder described some utilities now in use that improve the matching of references within Crossref by some 35%. When it came to trust and bias, there was less talk of quick wins. However, Leslie McIntosh pointed out that plagiarism can in some circumstances be a good thing. In the ‘methods’ section of an academic paper, authors are encouraged to write in a style and layout that is almost identical to the equivalent methods section of thousands of other papers. If such similarities occurred in the abstract or the main body of the article, that might read like

plagiarism. One worrying finding from the survey was that fewer than 10% of publishers using AI tools were checking for bias. Such a figure seems remarkably low compared with the number of times 'bias' is mentioned in an AI discussion today.

When asked for what they wanted to see happening in the next five years, Angela Cochran asked for a single format for references (Endnote currently has around 6,000 output styles, so there is plenty of scope for simplification). At the same time, she pointed out that we are 'abused' by the lack of trust shown by some of the biggest Internet corporations, such as Google and Facebook, which has the consequence of making us distrust AI-based solutions. Finally, Leslie McIntosh offered a more optimistic closing idea: we should build AI tools from those parts of the workflow that we trust, such as the peer review process, and then build on that. It was an encouraging tone on which to close.

Get in touch

UNSILO is a software company based in Denmark, providing machine-learning and AI tools for academic publishers and researchers. Contact us at:

www.unsilo.ai

@unsiloproduct

info@unsilo.ai

UNSILO A/s, Inge Lehmanns Gade 10, 8000 Aarhus, Denmark